Journal of Economic Interaction and Coordination (2025) 20:307–333
https://doi.org/10.1007/s11403-024-00426-y

REGULAR ARTICLE

Mesoscopic structure of the stock market and portfolio optimization

Sebastiano Michele Zema[1,2] · **Giorgio Fagiolo**[2] · **Tiziano Squartini**[3] · **Diego Garlaschelli**[3,4]

Received: 7 June 2023 / Accepted: 26 August 2024 / Published online: 17 September 2024

Abstract

The idiosyncratic and systemic components of market structure have been shown to be responsible for the departure of the optimal mean-variance allocation from the heuristic 'equally weighted' portfolio. In this paper, we exploit clustering techniques derived from Random Matrix Theory to study a third, intermediate (mesoscopic) market structure that turns out to be the most stable over time and provides important practical insights from a portfolio management perspective. First, we illustrate the benefits, in terms of predicted and realized risk profiles, of constructing portfolios by filtering out both random and systemic co-movements from the correlation matrix. Second, we redefine the portfolio optimization problem in terms of stock clusters that emerge after filtering. Finally, we propose a new wealth allocation scheme that attaches equal importance to stocks belonging to the same community and show that it further increases the reliability of the constructed portfolios. Results are robust across different time spans, cross sectional dimensions and set of constraints defining the optimization problem.

Keywords Random matrix theory · Community detection · Mesoscopic structures · Portfolio optimization

JEL Classification C02 · D85 · G11

✉ Sebastiano Michele Zema
sebastiano.zema@sns.it

[1] Scuola Normale Superiore, Pisa, Italy

[2] Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy

[3] Networks, IMT Institute for Advanced Studies, Lucca, Italy

[4] Lorentz Institute for Theoretical Physics, University of Leiden, Leiden, Netherlands

1 Introduction

The pioneering work of Markowitz (1952) laid the foundations of modern portfolio theory through the mean-variance (MV) optimization procedure. According to that model, the portfolio optimizer deals with uncertainty either by minimizing the variance of the investment, given the expected return, or by maximizing the expected return, given a certain level of risk. Despite its simplicity, it is widely recognized that the mean-variance framework delivers poor out-of-sample performances when the historical sample covariance matrix is used in the optimization process (Michaud 1989; Bai et al. 2009). As also shown by Laloux et al. (1999), the smallest eigenvalues of the spectra of the sample covariance matrix, which plays a fundamental role in the estimation of the global minimum variance (GMV) portfolio, are largely affected by noise. As a consequence, a MV optimization procedure which plug-in those estimates could be highly inaccurate, yielding in-sample predictions which seriously depart, in terms of portfolio returns and variances, from the realized ones.

The above mentioned limitations questioned the need for sophisticated optimization procedures for portfolio management purposes, especially considering that the simpler 1/*N* heuristic rule, being less affected by covariance estimation errors, achieves better out-of-sample performances (Duchin and Levy 2009; DeMiguel et al. 2009). To overcome the problems related to the adoption of the MV framework, techniques have been proposed either to ameliorate its theoretical predictions (see Brodie et al. 2009; DeMiguel et al. 2009; Tu and Zhou 2011, among others) or to capture the 'real' essence of the correlation matrix by means of which optimal portfolios are constructed through filtering procedures.[1] Overall, each different estimator and filtering procedure improves upon different portfolio aspects related to the performance, realized risk, reliability and diversification. These improvements also depend on other circumstances as the dimension-to-sample size ratio or the possibility of exploiting short-selling strategies.

In this work we show that once both noise and aggregate systemic fluctuations are considered in the portfolio optimization process (i.e., they are filtered-out from the covariance matrix), the resulting optimal asset allocation closely tracks the 1/N heuristic. As documented by Forbes and Rigobon (2002), the correlation coefficient is indeed conditional on market volatility[2], a direct consequence of which being that variables might appear as strongly correlated only because of temporary turmoil periods. For this reason, focusing on stable interconnections between stocks is fundamental to improve the reliability, in terms of predicted and realized risks, of the portfolio resulting from the optimization process: such a goal can be achieved by identifying which correlations are stable over time, i.e., not resulting from either random co-movements or temporary market effects. Moving from there, we further

[1] A comprehensive empirical study, regarding the possible improvements in the optimal asset allocation through the replacement of the sample correlations estimator with other estimation and filtering techniques, can be found in Pantaleo et al. (2011).

[2] In the rest of the article we will refer to the terms *systemic effects* and *market effects* as synonyms.

show that by redefining the asset allocation problem by giving equal importance to assets belonging to the same communities, i.e., groups of strongly interconnected stocks identified after filtering out noise and common aggregate effects (MacMahon and Garlaschelli 2015; Anagnostou et al. 2021), it is possible to further improve portfolio reliability.

Our contribution directly refers to two different but linked streams of literature. One is focused on network-based approaches that exploit the interconnected and evolving nature of markets. For example, Onnela et al. (2003) and Peralta and Zareei (2016) point out the existence of a relationship between the centrality of each stock in the network of log-return correlations and the weight induced by the MV optimization procedure, suggesting that optimal portfolios should include peripheral stocks to reduce the influence of central assets characterized by higher variance. Other studies heavily rely on hierarchical clustering techniques (Mantegna 1999; Bonanno et al. 2003, 2004; Di Matteo et al. 2004; Onnela et al. 2004; Tumminello et al. 2005): For instance, in Tola et al. (2008) optimal portfolios are constructed by replacing the empirical correlations with ultrametric distances induced by the corresponding hierarchical clustering scheme.

A different stream of literature focuses instead on filtering procedures that rely upon Random Matrix Theory (RMT) (Biely and Thurner 2008; Dimov et al. 2012; Singh and Xu 2016; Zitelli 2020). As shown in MacMahon and Garlaschelli (2015), different components of market structure can be identified by employing RMT-based clustering techniques that returns cohesive groups of stocks on the basis of which the portfolio optimization problem can be reformulated. Such an approach has been recently adopted by Anagnostou et al. (2021), who have focused on Credit Default Swap (CDS) markets, showing that such structures are indeed useful for credit risk modeling, especially because they may encode factors not necessarily related with standard industry/region taxonomies. Taken together, these results point out that filtered correlation matrices are typically more reliable—in terms of predicted and realized risk profiles—than those obtained using the empirical correlations as input.[3]

We organize the paper as follows. In sect. 2 we introduce our filtering procedure and explain how filtered correlations can be exploited to recover the mesoscopic structure of the stock market. We thus apply the proposed methodology on SP500 data showing the results in sect. 3. In sect. 4, we show how that information can be exploited in a portfolio optimization setting. Sect. 5 illustrates the advantages, in terms of predicted and realized risk reliability, of constructing portfolios as above. We thus illustrate the realized risk profiles and Sharpe ratios of the obtained portfolios against the baseline. Sect. 6 finally concludes and discusses possible paths for future research.

[3] This is true especially when the requirement $T \gg N$ cannot be satisfied (Laloux et al. 2000; Plerou et al. 2002).

2 The mesoscopic structure of the stock market

RMT can be employed to filter out the random noise from the correlation matrices of financial returns, by exploiting the *Marčenko–Pastur Law* (Marčenko and Pastur 1967). More formally, let $\{x_{it}\}$, with $i = 1 \dots N$ and $t = 1 \dots T$, be a sample of i.i.d. random variables with zero mean and variance σ^2. Let κ be the ratio T/N, assuming $\kappa \in (1, \infty)$ in the limit $T, N \to \infty$. Then, with probability one, the spectral density function of the sample covariance matrix tends to the Marčenko-Pastur distribution, i.e.,

$$f_\kappa(\lambda) = \frac{\kappa}{2\pi\lambda\sigma^2}\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})} \tag{1}$$

for $\lambda_{\min} \le \lambda \le \lambda_{\max}$, where $\lambda_{\max} = \sigma^2(1 + \sqrt{N/T})^2$ and $\lambda_{\min} = \sigma^2(1 - \sqrt{N/T})^2 > 0$.

The reader interested in the proof is redirected to Bai (1999). The result above implies that the empirical correlation matrix $\mathbf{C}$ can in principle be decomposed as

$$\begin{aligned} \mathbf{C} &= \sum_{i=1}^{N} \lambda_i |v_i\rangle\langle v_i| \\ &= \sum_{i:\lambda_i \in (0,\lambda_{\max}]} \lambda_i |v_i\rangle\langle v_i| + \sum_{i:\lambda_i \in (\lambda_{\max},\lambda_1]} \lambda_i |v_i\rangle\langle v_i| \end{aligned} \tag{2}$$

$$= \mathbf{C}^{(r)} + \mathbf{C}^{(s)}, \tag{3}$$

where $|v_i\rangle$ and $\langle v_i|$ denote the column and row eigenvectors associated with the eigenvalue λ_i, respectively. The above decomposition thus represents the empirical correlation matrix as a sum of matrices, respectively, induced by the *random* component $\mathbf{C}^{(r)}$, whose eigenvalues lie in the Marčenko-Pastur range $[\lambda_{\min}, \lambda_{\max}]$, and the *structural* (non-random) component $\mathbf{C}^{(s)}$.

When dealing with the correlation matrix of financial returns, the largest empirical eigenvalue λ_1 is much larger then the $\lambda_{\max}$ predicted by the MP law, which in turn shifts the lower eigenvalues leftwards below $\lambda_{\min}$ as well. As originally stressed by Laloux et al. (1999), this empirical evidence requires to subtract the contribution of λ_1 from the nominal value $\sigma^2 = 1$ when determining the threshold $\lambda_{\max}$ used to filter out the noise. Following then the prescription of Laloux et al. (1999), we determine $\lambda_{\max}$ in the MP-range by replacing $\sigma^2 = 1$ with $\sigma^2 = 1 - \lambda_1/N$. What remains after removing $\mathbf{C}^{(r)}$ is, then, recognized as signal rather than noise, hence supposed to possess useful economic information.

Interestingly, the (column) eigenvector $|v_1\rangle$ associated to λ_1 possesses elements having the same sign, thus identifying a matrix component $\lambda_1 |v_1\rangle\langle v_1|$ systematically affecting all the stocks in the same direction and with strong intensity. This clearly reflects the presence of the well-known 'market factor', which further induces the decomposition of $\mathbf{C}^{(s)}$ as follows:

$$\mathbf{C}^{(s)} = \sum_{i:\lambda_i \in (\lambda_{\max},\lambda_m)} \lambda_i |v_i\rangle\langle v_i| + \lambda_m |v_m\rangle\langle v_m| = \mathbf{C}^{(g)} + \mathbf{C}^{(m)} \tag{4}$$

i.e., as a sum of a *mesoscopic* spectral component $\mathbf{C}^{(g)}$ and a *systemic* component (or *market mode*) $\mathbf{C}^{(m)}$, where we renamed λ_1 and v_1 as λ_m and v_m, respectively, to stress they refer to the 'market' trend.

A graphical illustration of this empirical feature is provided in Fig. 1 for a sample of stocks constituting the S &P500 index and stably traded over the period 2000–2015. The sample constitutes the dataset used in the remainder of the work as well. Being the systemic component pervasive and time-varying, some stocks might appear interconnected only as a consequence of their common dependence on global market events (see Forbes and Rigobon 2002; Billio et al. 2012, among others). When performing asset allocation strategies based on historical data, it is fundamental to minimize covariance estimation errors induced by any possible time-varying component—which makes historical data not reliable for the future.

To provide an example, let us define the total risk of a system as $\Lambda := \sum_{k=1}^{N} \lambda_k$ and investigate the temporal evolution of the cumulative risk fraction of the different components of the covariance matrix of stock returns by adopting non-overlapping, rolling windows of two years. To this aim, can draw 100 randomized samples of size for each temporal window: The resulting averaged shares of total risk accounted by the random, systemic and mesoscopic component of the spectrum of the covariance matrix are shown in Fig. 2. While the random and systemic cumulative risk fractions vary quite a lot across the considered period, the mesoscopic one is the most stable, a result letting us to conclude that the construction of more reliable portfolios may indeed be based on the stable part of the spectrum.

Let us now use $\mathbf{C}$ to partition the stock market into non-overlapping communities of stocks that are more correlated internally than expected under a suitable null model. Detecting communities in financial markets is not new in the literature: for instance, Fenn et al. (2012) compared different procedures to unfold the community structure of the foreign exchange market and Verma et al. (2019) used clusters to extract relevant factors for volatility modeling. However, the procedure we are now going to illustrate is based on a combination of modularity maximization (Clauset

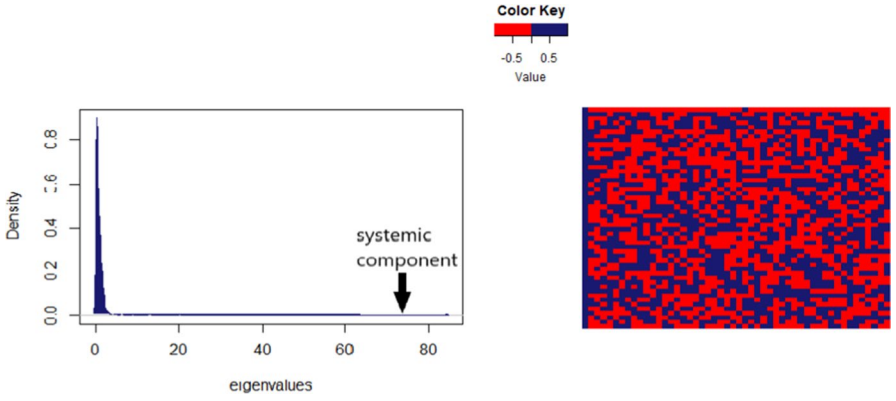

Fig. 1 Eigenvalue density for the 2000–2015 covariance matrix of the S &P500 components (left) and heatmap of the associated eigenvectors (right). The first column of the heatmap represents the eigenvector associated to the systemic component, whose elements all have the same sign

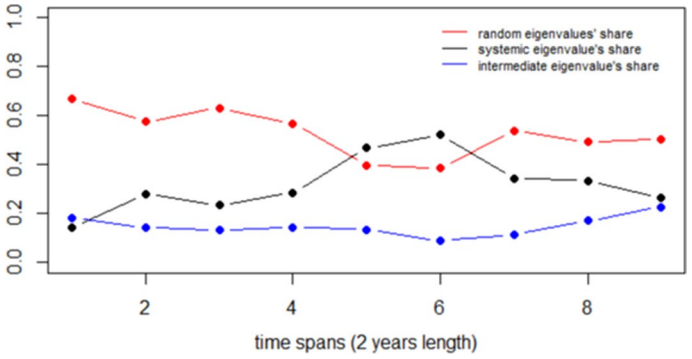

Fig. 2 Cumulative risk fractions associated to the different components of the correlation matrix over different time spans. The random and systemic components vary the most (14% standard deviation for both) while the intermediate, mesoscopic range of the spectrum is more stable (5% standard deviation only)

et al. 2004; Newman 2006) and RMT (MacMahon and Garlaschelli 2015), which was shown to be theoretically superior in the case of correlation matrices.

In the network science literature, the so-called modularity $Q(\boldsymbol{\gamma})$ of a partition $\boldsymbol{\gamma}$ of the N nodes of a network is defined as

$$Q(\boldsymbol{\gamma}) = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{N} \left[w_{ij} - \langle w_{ij}\rangle\right]\delta(\gamma_i, \gamma_j) \tag{5}$$

where w_{ij} the entry of the adjacency matrix of the (possibly weighted) network (i.e., w_{ij} is the weight of the link from node i to node j), $\langle w_{ij}\rangle$ is its expected value under a suitably chosen null model, and the Kronecker delta $\delta(\gamma_i, \gamma_j)$ guarantees that only the nodes belonging to the same community contribute to the modularity. The goal of modularity maximization is to find the partition that maximizes $Q(\boldsymbol{\gamma})$, thus emphasizing the community of nodes whose internal interactions are stronger and maximally unexplained by the (community-free) null model.

For networks, the null model chosen is generally the so-called Weighted Configuration Model (WCM) that randomizes the network topology while preserving the empirical strength $s_i = \sum_{j=1}^{N} w_{ij}$ of each node i. A popular, although generally incorrect (Garlaschelli and Loffredo 2009), expression used to represent this null model is

$$\langle w_{ij}\rangle = \frac{s_i s_j}{2W} \quad \forall\, i,j \tag{6}$$

where $2W = \sum_{i=1}^{N} s_i = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$ is the total edge weight of the network. When considering correlation matrices, the null model above has been shown to be inconsistent (MacMahon and Garlaschelli 2015) as a result of the fact that, unlike (weighted) networks, correlation matrices cannot be directly randomized by considering their entries as independent. Rather, the randomization should occur at

the level of the underlying time series, and the correlation matrix should then be recalculated from the randomized time series. In particular, by reformulating the modularity for correlation matrices as

$$Q(\boldsymbol{\gamma}) = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{N} \left[C_{ij} - \langle C_{ij}\rangle\right]\delta(\gamma_i, \gamma_j), \tag{7}$$

a consistent community-free null model representing random empirical correlations resulting only from noise and possibly global trends comes precisely from RMT and can be expressed as

$$\langle C_{ij}\rangle = C_{ij}^{(r)} + C_{ij}^{(m)} \tag{8}$$

(MacMahon and Garlaschelli 2015). The above null model discounts both the random and the systemic components of correlations. As a consequence,

$$C_{ij} - \langle C_{ij}\rangle = C_{ij}^{(g)}, \tag{9}$$

i.e., the modularity matrix coincides with the mesoscopic component of the original correlation matrix. Therefore, maximizing $Q(\gamma)$ guarantees that the identified communities are necessarily formed by internally positively (after discounting the null model) and mutually negatively (again, after discounting the null model) correlated stocks. In other words, communities are ideally noise-free and mutually anti-correlated with respect to the market.

3 Stock market communities

The dataset used for the present analysis has been downloaded from Yahoo Finance and consists of daily equity data for current S &P500 constituents over the period 2000–2015. The period has been chosen to get a sample that is as homogeneous as possible in terms of index constituents, so to apply the methodology on a sufficiently large sample of stocks that were stably traded on the market over a relatively long period of time. This resulted in a final sample of 450 stocks. After applying RMT to isolate the mesoscopic component of the matrix, we performed modularity maximization by implementing a modified version of the Louvain algorithm (Blondel et al. 2008), taking as input the matrix $\mathbf{C}^{(g)}$. Consistency and stability of this approach have been discussed in MacMahon and Garlaschelli (2015) and Anagnostou et al. (2021) to which the interested reader is referred for additional technical clarifications.

We performed community detection on the entire time span 2000–2015 and identified an optimal partition of the 450 stocks into 4 communities. Figure 3 shows the heatmaps depicting the sequence of transformations leading from the original stocks to such a set of mutually, negatively correlated communities. Figure 4 shows the detected communities, together their relative compositions, according to the industrial classification. The number of detected communities is lower than the number

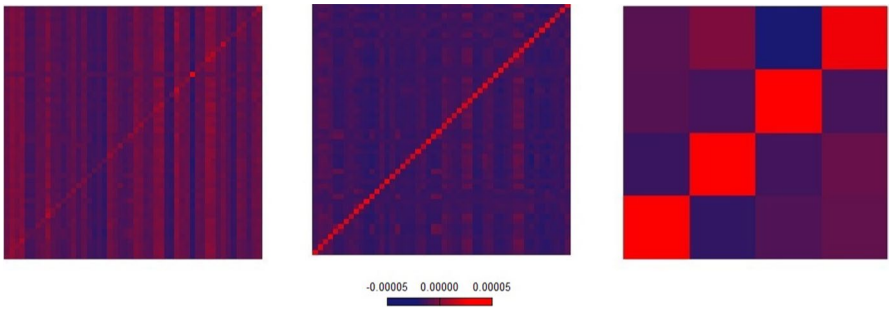

Fig. 3 Sequence of transformations applied to the empirical correlation matrix (left), leading first to the 'noise and systemic free' correlation matrix (middle) and then to the internally positively and mutually negatively correlated clusters

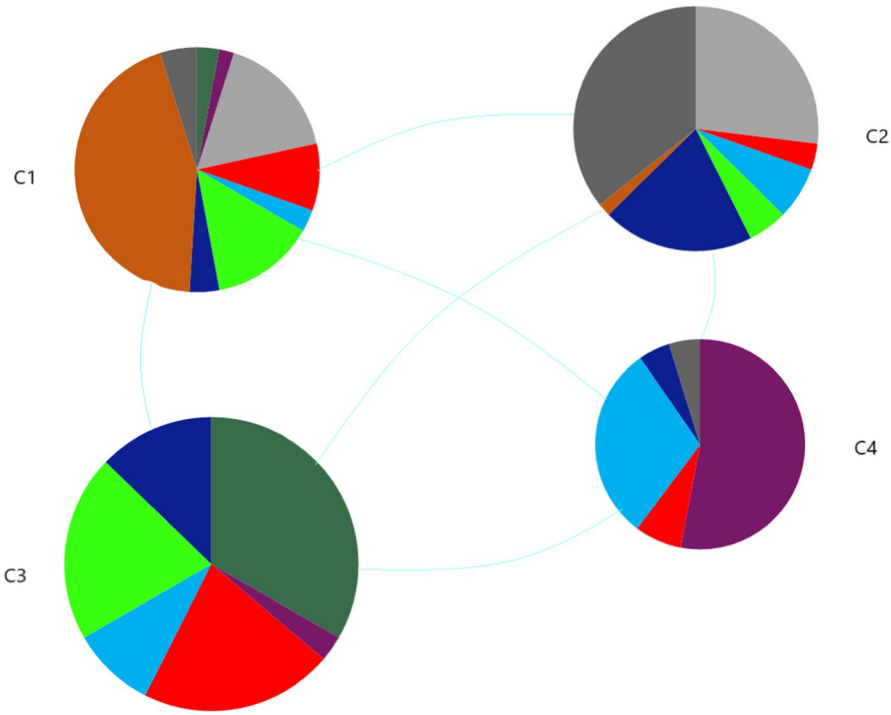

Fig. 4 Community structure of the selected 450 stocks during the period January 2000-December 2015: Finance, Energy, Healthcare, Industrials, Materials, Discretionary, Staples, Technology, Utilities following the GICS classification

of considered sectors, showing the tendency of stocks to be strongly interconnected across different sectors as well. Still, it can be noticed how stocks belonging to specific sectors tend to cluster more than others—a behavior detected also in Borghesi et al. (2007) by employing hierarchical clustering techniques. In particular, almost

all stocks in the financial sector are clustered together in C3 while stocks in the energetic and technological sectors are, respectively, placed in C4 and C1; utilities are quite clustered in C2. The remaining sectors (namely industrials, materials, consumer discretionary, consumer staples and healthcare), instead, are more dispersed across different communities. This result confirms that the data-driven cluster identification leads to communities that are unpredictable from the nominal sectoral classification of stocks, as also observed in MacMahon and Garlaschelli (2015) and Anagnostou et al. (2021).

Performing community detection over the whole available time span, i.e., from 2000 to 2015, might seem unreasonable since structural changes have arguably occurred in occasion of the 2008 financial crisis and possibly other events. This is only partially true: It turns out that while the original, unfiltered empirical correlations do change a lot over time (especially during market turmoils), mesoscopic correlations remain remarkably stable, thus stabilizing the optimal partition. This can be easily seen by comparing the evolution of the density of unfiltered, empirical correlations of our sample of stocks with that of filtered, mesoscopic correlations employed to perform the clustering procedure. As Fig. 5 shows, the distribution of the empirical correlation coefficients clearly shifts toward higher values in the second half of the considered time span (which contains a period of higher turmoil), so that coefficients calculated over the entire time span are not representative of the underlying sub-periods; by contrast,

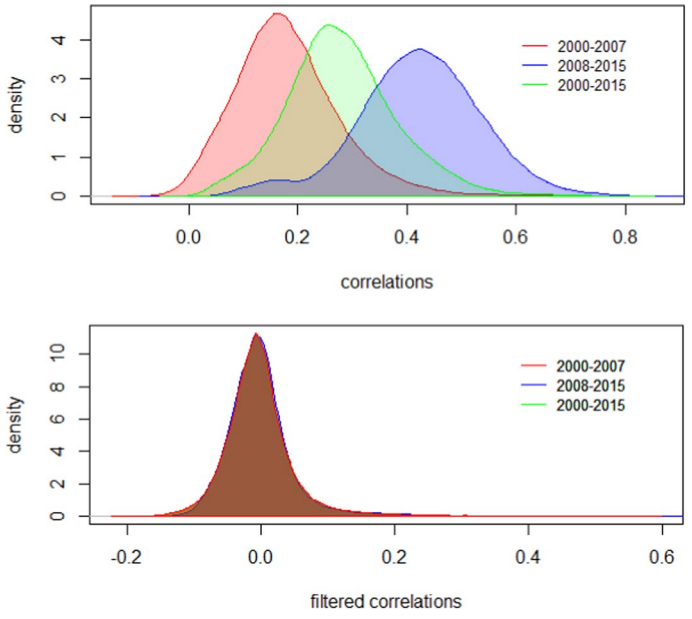

Fig. 5 Densities of the unfiltered empirical (top) and filtered mesoscopic (bottom) correlation coefficients over different periods. Notice that, while a clear shift occurs between the first and the second half of the overall 2000–2015 period for the unfiltered coefficients, no shift occurs for the filtered mesoscopic ones. As a consequence, the distribution of the unfiltered matrix entries calculated over the entire period is not representative of the distributions for the individual sub-periods, while that of the filtered matrix entries is

when considering only the mesoscopic component of the correlation matrix over different periods, we find that the distribution of the entries of such component almost perfectly overlap with each other over time. Thus, in this case, the overall distribution is representative of the distributions for the sub-periods.

Estimates of the optimal weights for asset allocation strategies typically take historical data as input. Including historical information, however, does not necessarily bring advantages, as what happened in the past does not necessarily repeat itself with the same regularity—especially during periods of high market volatility and structural breaks. This consideration will shape our optimal asset allocation strategies, which will be performed by taking into account only the stable part of the covariance matrices of the assets, namely the mesoscopic ('intermediate') one. Noticeably, optimal portfolios built by considering only 'stable' information will be shown to closely track the equal weight strategy.

It is worth stressing that our objective is not that of suggesting that the market component should be eliminated from the definition of a realistic data generating process for stock returns; rather, here we aim at showing how heavily its structural and pervasive nature affects portfolios as well as investigating its impact, in terms of reliability, on the construction of the optimal ones.

4 Back to basic portfolio optimization

Let us now address the implications of the market mesoscopic structure from a portfolio management perspective.

In order to do so, let us briefly review the classical Markowitz portfolio optimization scheme. Consider N risky assets with covariance matrix $\mathbf{\Sigma}$ and vector of expected returns $\boldsymbol{\mu}$. Given the wealth allocation vector $\boldsymbol{\omega} = [\omega_1 \dots \omega_N]$, such that $\sum_i \omega_i = 1$, the portfolio expected return reads

$$\mu_p = \sum_{i=1}^{N} \omega_i \mu_i \tag{10}$$

with associated variance reading

$$\sigma_p^2 = \sum_{i=1}^{N} \omega_i^2 \sigma_i^2 + \sum_{i>j} 2\omega_i \omega_j C_{ij} \sigma_i \sigma_j. \tag{11}$$

The well-known MV approach consists in finding the allocation vector $\boldsymbol{\omega}$ which minimizes σ_p^2 subject to a given value of μ_p or, equivalently, the one that maximize the return subject to a given level of variance. The optimization problem to be solved, expressed in matrix form, reads

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} \\ &\text{s.t.}\mu_p = \boldsymbol{\omega}'\boldsymbol{\mu} \\ &\sum_i^N \omega_i = 1 \end{aligned} \tag{12}$$

and has solution

$$\boldsymbol{\omega}^* = \boldsymbol{b}\boldsymbol{\Sigma}^{-1}\mathbf{1} + \boldsymbol{c}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{13}$$

with

$$\begin{aligned} \boldsymbol{b} &= \frac{A - \boldsymbol{\mu_p} B}{\Delta} \qquad \boldsymbol{c} = \frac{\boldsymbol{\mu_p} C - B}{\Delta} \\ A &= \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \qquad B = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ C &= \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1} \qquad \Delta = CA - B^2. \end{aligned}$$

In the case of a completely risk-adverse investor that is only interested in minimizing the risk with no constraint on the expected returns, the solution simply becomes

$$w_{\text{gmv}} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^t\boldsymbol{\Sigma}^{-1}\mathbf{1}} \tag{14}$$

where ω_{gmv} denotes the investment plan associated with the global minimum variance (GMV) portfolio. We, then, focus on *reliability* of optimal portfolios by comparing their predicted risk, σ_p, obtained via the correlation matrices estimated using historical data, with (ex-post) realized risk, σ_p^r. As in Tola et al. (2008), we deem a portfolio as reliable if

$$\mathcal{R} = \frac{|\sigma_p^r - \sigma_p|}{\sigma_p} \tag{15}$$

is 'small' - the main difference of our approach being that we never assume perfect knowledge of future volatilities for the investor, letting uncertainty affect the whole covariance matrix. Being interested in understanding whether the adoption of the stable (i.e., mesoscopic) part of the correlation matrix over time reduces the discrepancies between predicted and realized variances, the reliability index is our main object of interest. However, for the sake of completeness, the main analysis will be complemented by checking other standard and important metrics as the Sharpe Ratios and the realized risk for the GMV portfolio.

4.1 Noise-free and systemic free optimization

As shown before, the systemic component affects all stocks in the same direction, inducing a positive amount of covariance between the variables, i.e., $\sigma_{ij}^{(m)} > 0$; it is, then, straightforward to show that, for a risk minimizer investor, the adoption of the

Fig. 6 Asset composition comparison for the periods 2000–2003 (top), 2004–2007 (middle) and 2008–2011 (bottom) between the $1/N$ rule (horizontal black line), sample covariance GMV (red) and the portfolio optimization based on mesoscopic correlations (blue). For each stock on the x-axis, the relative weight on the y-axis is shown, the mesoscopic-based optimization closely follow the heuristic $1/N$ rule. When short-selling is not allowed we have $\omega_i \geq 0, \forall\, i$ ▶

mesoscopic variance $\sigma_{ij}^{(g)} = C_{ij}^{(g)}\sigma_i\sigma_j$, in place of $C_{ij}^{(m)}$, rebalances the portfolio: Hence, total wealth will be no longer concentrated over few assets characterized by the lowest past variances. In other words, while in presence of co-movements (e.g., because of market turmoil), a risk minimizer investor would lower the portfolio risk by concentrating the wealth over the less risky assets, an investor who is aware of the temporarily nature of aggregate shocks causing crashes in the market, would filter out the systemic effects from past data and trust only the stable part of the correlation matrix.

To provide empirical evidence for such a statement, we performed MV optimization on the S &P500 constituents over multiple periods, comparing the wealth allocation vectors obtained using the empirical and the mesoscopic correlation matrices and keeping the equally weighted portfolio as a benchmark[4]: Fig. 6 shows the results, considering both cases in which short selling is either possible or not. Noticeably, the MV optimization procedure implemented by removing both noise and systemic effects from the correlation matrix closely follow the $1/N$ rule, yielding as an optimal solution a portfolio which closely tracks the equally weighted one. That is, when we implement portfolio optimization focusing on the intermediate and most stable part of the correlation matrix only, the heuristic 1/N strategy turns out to be optimal. In what follows, we will refer to such an approach as the 'mesoscopic-based' optimization procedure.

We are not the firsts to show that the heuristic 1/N strategy can be seen as the optimal solution of a portfolio optimization problem. For instance, Pflug et al. (2012) have shown the 1/N to be optimal when there is a 'very-high' level of uncertainty about future asset returns, which is, however, an argument similar to that invoked by DeMiguel et al. (2009) as well. Thus, in this work we make a further step by linking the 'uncertainty' to which these authors refer to the different spectral component of the empirical correlation matrix of financial returns. That is, we show how the systematic and randomic components of the correlation matrix are those responsible for the departure of the GMV portfolio from the 1/N strategy.

On the contrary, the MV framework which plugs-in the sample covariance matrix, returns a much more heterogeneous composition being more sensitive to estimation errors. This confirms that the optimization procedure based on the mesoscopic structure of the correlation matrix is less sensitive to both noisy and aggregated fluctuations, thus yielding more balanced portfolios. As an additional test, in Fig. 7 we compare the mesoscopic and $1/N$ weights with the ones we would obtain by cleaning the correlation matrix only from noise through the

[4] A short analytical description of the rebalancing effect for the $N = 2$ assets case is provided in the appendix.

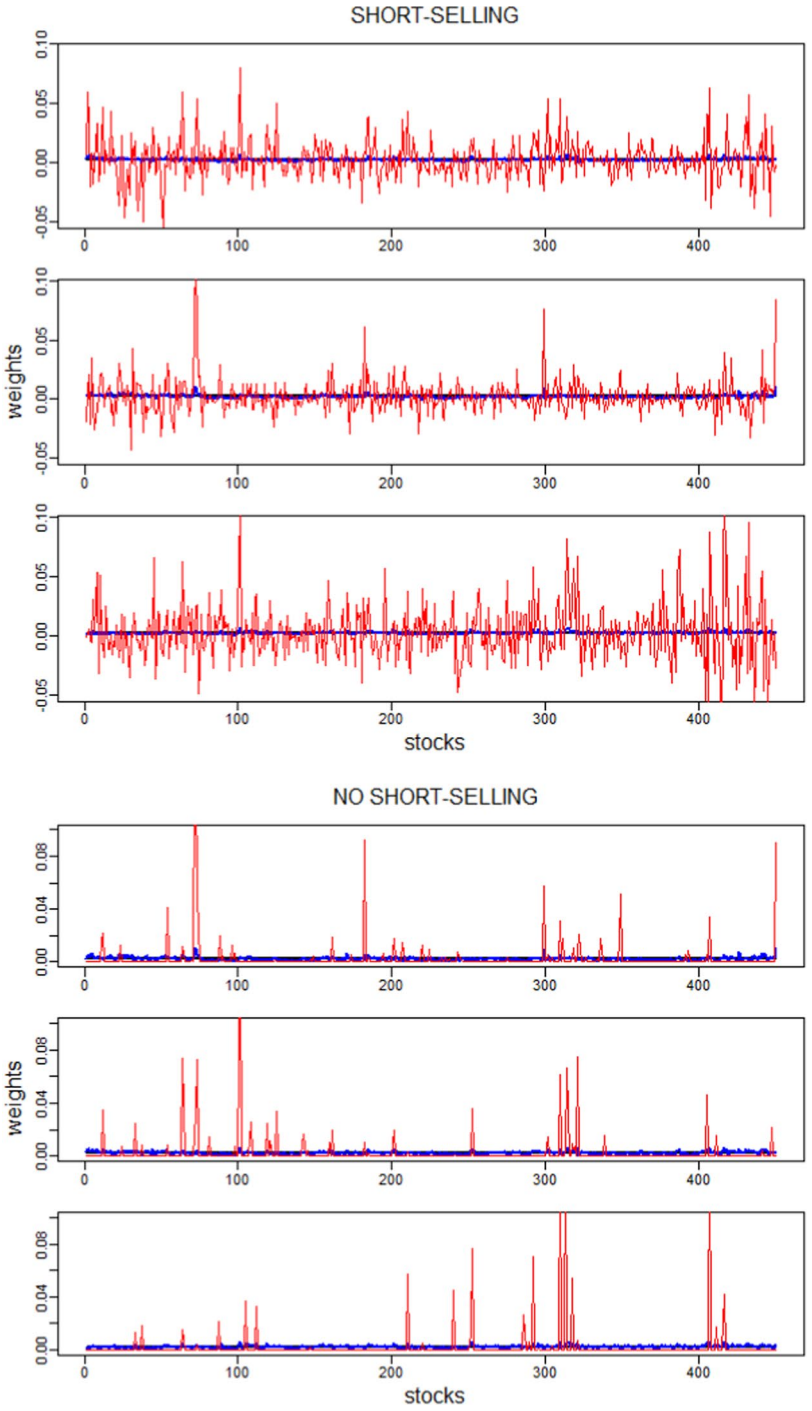
SHORT-SELLING
weights
stocks
NO SHORT-SELLING
weights
stocks

Fig. 7 Asset composition comparison for the periods 2000–2003 (top), 2004–2007 (middle) and 2008–2011 (bottom) between the $1/N$ rule (horizontal black line), RMT approach (green) and the portfolio optimization based on mesoscopic correlations (blue). For each stock on the x-axis, the relative weight on the y-axis is shown. Cleaning from the noise is not sufficient to closely track the heuristic rule as it is when adjusting from the market component as well

standard RMT-based approach: in order to closely track the balanced $1/N$ allocation it is necessary to filter out both the noisy and the systemic components.

A measure of similarity to the equally weighted portfolio is provided by the number of stocks with a 'significant' amount of money invested into. Following Bouchaud and Potters (2003), this quantity can be defined as

$$\mathcal{N} = \frac{1}{\sum_{i=1}^{N} \omega_i^2}. \tag{16}$$

Indeed, when the wealth is equally divided among the N assets, $\mathcal{N}$ is equal to N; on the other hand, it is equal to 1 when the wealth is invested only in one asset. As stressed in Tola et al. (2008), the quantity $\mathcal{N}$ simply provides a rough estimate of the number of stocks which could be effectively used to build a portfolio that is smaller than the original one, while preserving most of the risk-return properties of the latter.

In Fig. 8, the effective size of portfolios obtained with different allocation rules are displayed and compared: The compared allocation strategies are, again, the sample covariance GMV, the $1/N$ rule, the mesoscopic-based GMV portfolios and, finally, the noise-free GMV portfolios. It can be noticed that, independently from factors such as the time span, the subsample and the subsample size considered, the mesoscopic-based GMV portfolios are always much closer to the $1/N$ rule compared to the sample covariance GMV and to the RMT portfolios where only the random component is filtered-out. As it will be shown afterward, this result brings non-trivial practical implications in terms of reliability.

4.2 Mesoscopic community-based optimization scheme

The adoption of mesoscopic correlations leads to balanced portfolios closely tracking the equally weighted investment plan. Let us now show how the portfolio optimization problem can be simply reformulated by considering the detected clusters of stocks, instead of the single ones, to further reduce the uncertainty characterizing each specific asset.

Let $N = N_1 + N_2 + \cdots + N_n$ be the total number of asset, n the number of detected communities, N_c being the number of assets in a given community (denoted by the subscript $c \in \{1, 2 \ldots n\}$). The problem, now, is that of finding the share of wealth W_c which has to be invested into a given community, with $\omega_c = W_c/N_c$ being the share of wealth invested into the generic asset i belonging to that community. The problem can be, thus, reformulated as follows

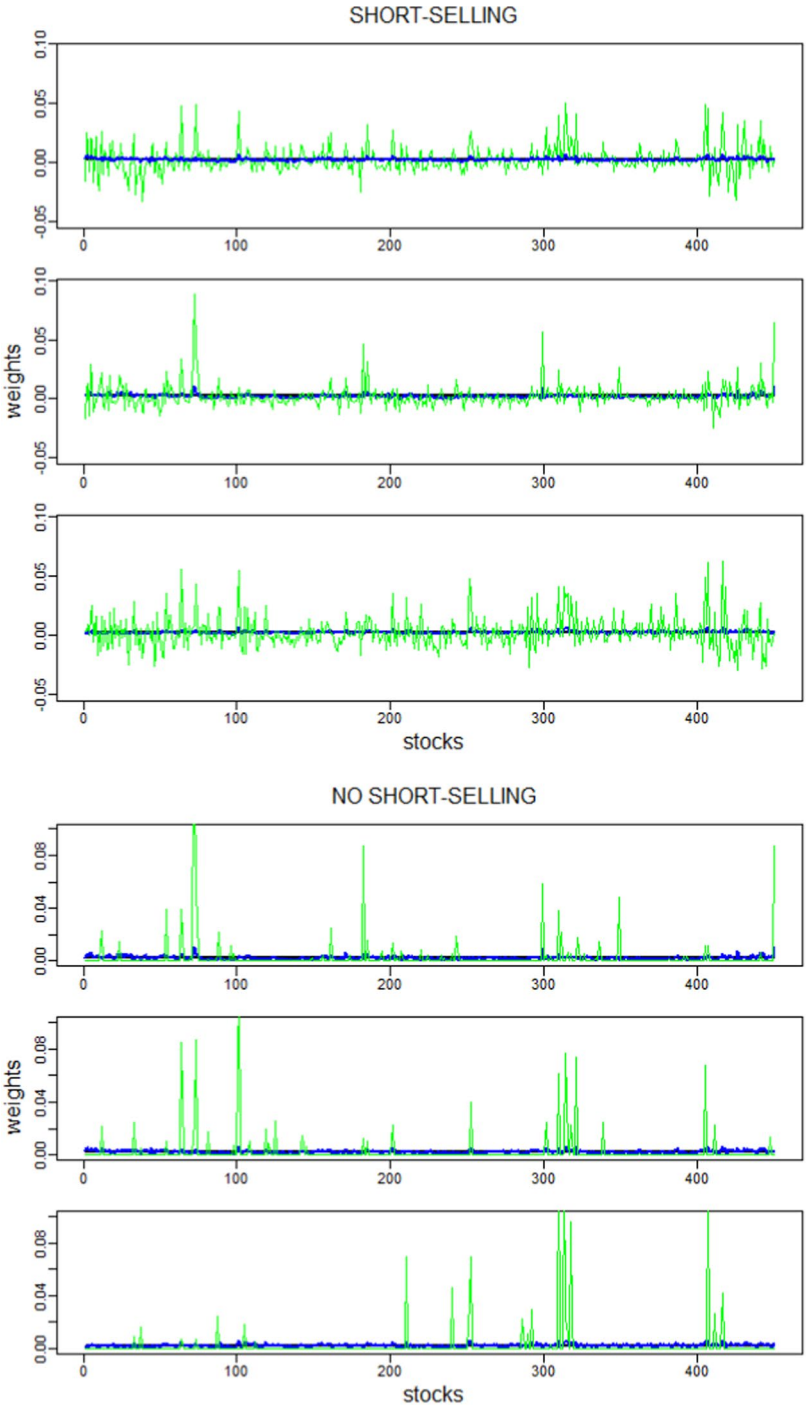
SHORT-SELLING
weights
stocks
NO SHORT-SELLING
weights
stocks

Fig. 8 Effective sizes $\mathcal{N}$ for each of the 100 random subsamples. In panel **a**, the size of the subsamples is 100, while in panel **b** is 50. The subsamples are of length $T = 3$ years with the plots covering together, inside each panel, the 2000–2012. Mesoscopic GMV portfolios in blue, RMT filtered in green, sample covariance GMV in red

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \boldsymbol{\omega}' \boldsymbol{\Sigma}^{(g)} \boldsymbol{\omega} \\ &\text{s.t.} \mu_p = \boldsymbol{\omega}' \boldsymbol{\mu} \\ &\sum_{c}^{n} W_c = 1 \\ &\omega_i = \omega_j \quad \forall \ \ i,j \in c. \end{aligned} \tag{17}$$

and is the same as the problem in Eq. 3 with the difference that weights are constrained to be equal for all the stocks belonging to the same community c; naturally, the total wealth share, being the sum of the wealth shares invested into each community, must still sum up to one. Reformulating the problem as in Eq. 17 leads to the minimization of the following objective function, where no constraint on the expected return is imposed:

$$\sigma_p^2 = \sum_{c=1}^{n} \omega_c^2 \left[N_c \overline{\sigma}_c^2 + N_c (N_c - 1) \overline{\sigma}_{ijc}^{(g)} \right] + \sum_{c=1}^{n-1} \sum_{k=c+1}^{n} 2\omega_c \omega_k \left[N_c N_k \overline{\sigma}_{ck}^{(g)} \right]; \tag{18}$$

notice that $\overline{\sigma}_c^2$ and $\overline{\sigma}_{ck}$, respectively, denote the average of the variances inside a given community and the average of the mesoscopic covariances between assets belonging to different communities. The reliability analysis of the proposed approach will be assessed in the next section.

5 Reliability and performance analysis

Let us now compare the reliability $\mathcal{R}$ of the portfolios obtained by implementing the sample covariance GMV portfolios, the mesoscopic-based portfolios (i.e., both noise and systemic free), the mesoscopic community-based portfolios and, finally, the heuristic equally weighted strategy.

Both cases with and without short-selling will be analyzed, focusing on the GMV portfolios computed over different periods and for different sample sizes. For the sake of completeness, we repeat the comparison also imposing constraints on the expected portfolio return μ_p. The different portfolios constituting the efficient frontier are constructed, obtaining for each of them the index $\mathcal{R}$ given the out-of-sample realized portfolios variances. The analyses are carried out over different time spans and considering different sample sizes: The time spans analyzed, i.e., $T_1 = 2000 - 2007$, $T_2 = 2004 - 2011$ and $T_3 = 2008 - 2015$ are divided in two additional sub-periods of equal length by fixing t_0. Upon doing so, we create portfolios given the data collected over the period $t_0 - \Delta t$ and quantify their out-of-sample performance over the period $t_0 + \Delta t$. For what concerns the

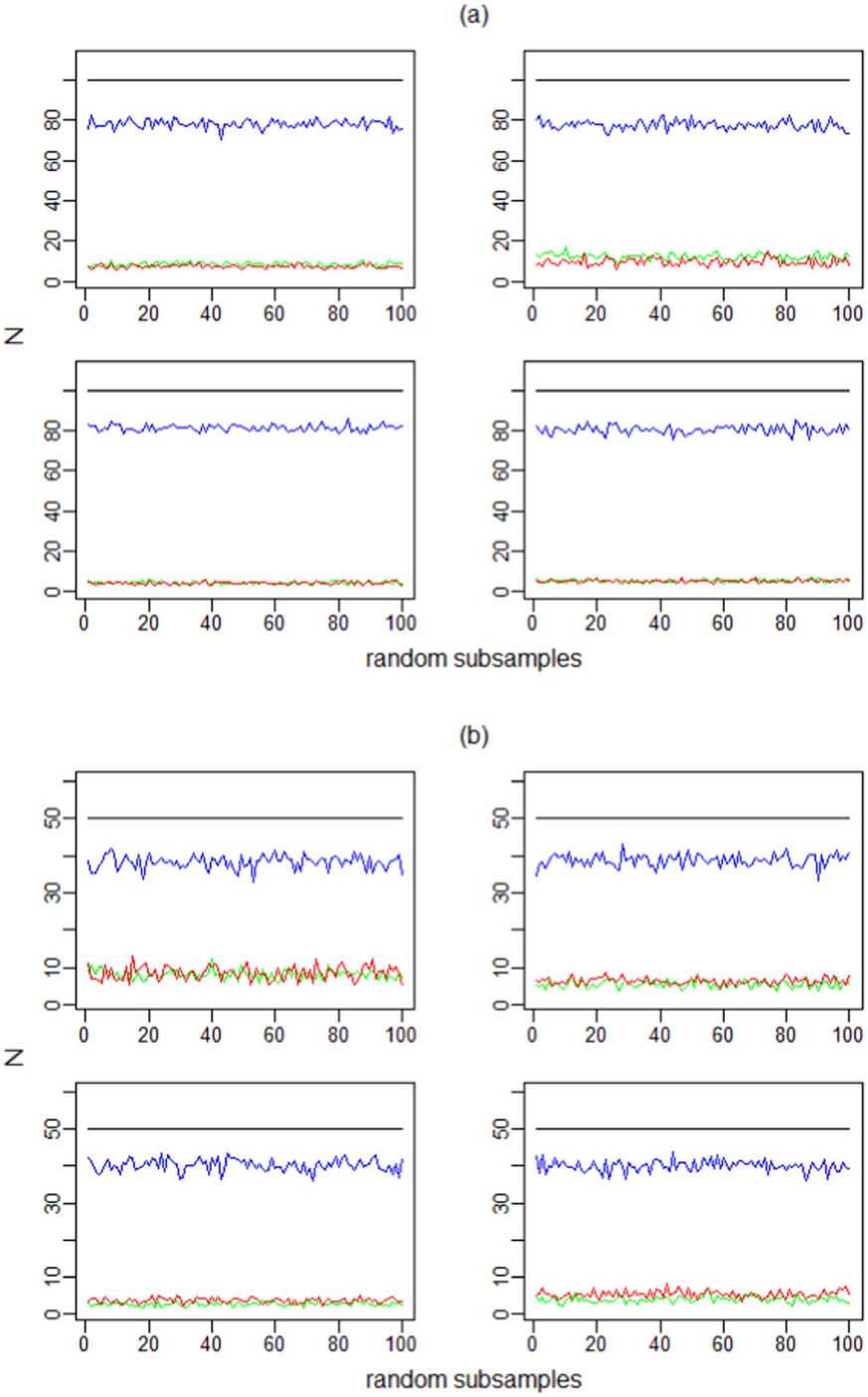
(a)
N
random subsamples
(b)
N
random subsamples

Table 1 Reliability $\mathcal{R}$ for each strategy adopted and for each sample size, under different time spans, ranging from the $N = 50$ case to the whole sample case ($N = 450$). The mesoscopic approach closely tracks the reliability of the 1/N heuristics, which is in turn higher than the classical GMV portfolio based on the sample covariance matrix plugged-in as input

	Time span	$\mathcal{R}_{\text{equally}}$	Short-selling		No short-selling	
			$\mathcal{R}_{\text{mesoscopic}}$	$\mathcal{R}_{\text{GMV}}$	$\mathcal{R}_{\text{mesoscopic}}$	$\mathcal{R}_{\text{GMV}}$
$N = 50$	T_1	0.53	0.47	0.79	0.48	0.68
	T_2	4.06	4.05	5.67	4.05	4.78
	T_3	0.8	0.79	0.3	0.8	0.58
$N = 100$	T_1	0.17	0.15	1.12	0.23	0.53
	T_2	1.38	2.07	3.74	1.3	2.72
	T_3	0.27	0.27	0.31	0.4	0.3
$N = 200$	T_1	0.13	.11	1.26	0.16	0.46
	T_2	1.03	1.05	3.72	1.38	2
	T_3	0.2	0.2	0.43	0.26	0.2
Whole sample	T_1	0.47	0.45	6.25	0.45	0.61
	T_2	4.75	4.18	17.11	4.88	6.65
	T_3	0.8	0.78	2.32	0.79	0.54

Table 2 Comparison of the community-based portfolios with the mesoscopic (i.e., noise and systemic free but without community constraints) and the classical GMV one

	Time span	$\mathcal{R}_{\text{equally}}$	$\mathcal{R}_{\text{mesoscopic}}$	$\mathcal{R}_{\text{community}}$	$\mathcal{R}_{\text{GMV}}$
$\mathcal{R}$	T_1	0.47	0.45	0.41	0.61
	T_2	4.75	4.88	3.45	6.65
	T_3	0.8	0.78	0.74	0.54

samples size, we randomly extract 100 subsamples out of the S &P500 components, for each considered size. Average values computed for the $\mathcal{R}$ indices are reported in Tables 1, 2.

We can summarize the results as follows. When short-selling is allowed and no constraint on the expected portfolio return is present, the sample covariance GMV is always underperforming—the only exception being represented by the lowest-dimensionality case ($N = 50$)—when compared with the $1/N$ and the mesoscopic-based optimization rule, irrespectively from the sample size and the time span considered. In addition, our methodology performs slightly better than the equally weighted portfolio, thus revealing itself to be the most reliable investment plan considered: The difference, however, is almost negligible, confirming how close the mesoscopic-based optimization procedure is to the equally weighted strategy.

The poor performance of the sample covariance GMV portfolios is not a novel result, especially when no constraint about the possibility of exploiting short-selling strategies is imposed (see Frost and Savarino 1988; Eichhorn et al. 1998; Britten-Jones 1999; Jagannathan and Ma 2003)—Note that, from a purely mathematical perspective, imposing constraints is equivalent to letting

a shrinkage operator to act on the covariance matrix of the assets, which helps when the number of parameters to estimate is too large and, as a consequence, estimation errors are large as well. Our empirical analysis confirms these results: Notice the huge improvement of the MV optimization based on sample covariances compared to the situation in which short-selling was not allowed, although the mesoscopic-based optimization provides better reliability indices in all cases except in period T_3.

Let us now check whether the clusters detected on $\mathbf{C}^{(g)}$ can be used as a further source of information. In particular, let us attach homogeneous optimal weights to stocks belonging to the same clusters, denoting with $\mathcal{R}_{\text{community}}$ the corresponding reliability index. To make the comparison as clear as possible, we compare the reliability of the community-based portfolios only to the best performing competing approaches, in each time span. Results in Table 3 noticeably confirm the informativeness of the detected communities. Portfolios in which optimal weights are recovered by constraining stocks in the same community to have the same weight further improve their reliability indices, outperforming both the equally weighted and the mesoscopic-based strategies for all considered time spans. Still, the simple MV based on sample covariances with the no short-selling constraint is the more reliable one in T_3.

Let us now consider all approaches, i.e., the sample covariance one, the mesoscopic-based and the mesoscopic community-based ones and compute the reliability

Table 3 Summary statistics of the reliability $\mathcal{R}$ indices for the noise plus systemic free, standard mean-variance approach with the sample covariance matrix plugged-in, and community-based efficient frontiers over different periods, with and without short-selling strategies. For each efficient frontier, specific quantiles are taken and the $\mathcal{R}$ index of the portfolio located on that point of the frontier is computed

		Min.	1st quartile	Median	Mean	3rd quartile
Short-selling						
T_1	$\mathcal{R}_{\text{community}}$	0.34	0.39	0.43	0.5	0.52
	$\mathcal{R}_{\text{mesoscopic}}$	0.63	1.05	1.81	2.4	3.15
	$\mathcal{R}_{\text{standard MV}}$	1.13	1.46	2.1	2.5	3.31
T_2	$\mathcal{R}_{\text{community}}$	3.56	3.96	4.09	4.02	4.19
	$\mathcal{R}_{\text{mesoscopic}}$	7.4	8.13	9.1	9.4	10.4
	$\mathcal{R}_{\text{standard MV}}$	4.05	7.07	7.67	7.85	8.5
T_3	$\mathcal{R}_{\text{community}}$	0.79	0.81	0.81	0.81	0.83
	$\mathcal{R}_{\text{mesoscopic}}$	0.49	0.84	1.36	1.7	2.33
	$\mathcal{R}_{\text{standard MV}}$	0.88	1.07	1.4	1.51	1.85
No short-selling						
T_1	$\mathcal{R}_{\text{community}}$	0.006	0.14	0.20	0.26	0.46
	$\mathcal{R}_{\text{mesoscopic}}$	0.03	0.28	0.64	0.72	0.94
	$\mathcal{R}_{\text{standard MV}}$	0.02	0.22	0.48	0.48	0.75
T_2	$\mathcal{R}_{\text{community}}$	3.23	3.33	3.39	3.45	3.74
	$\mathcal{R}_{\text{mesoscopic}}$	0.043	0.46	2.70	3.71	4.75
	$\mathcal{R}_{\text{standard MV}}$	0.02	0.42	2.19	1.52	2.37
T_3	$\mathcal{R}_{\text{community}}$	0.74	0.76	0.77	0.76	0.78
	$\mathcal{R}_{\text{mesoscopic}}$	0.03	0.31	0.56	1.34	2.64
	$\mathcal{R}_{\text{standard MV}}$	0.52	0.53	0.54	0.56	0.59

Table 4 Out of sample annualized Sharpe ratios $\mathcal{S}$ for each strategy adopted under different time spans. Annualized portfolios' return and risk provided as well

	Time span	Mesoscopic	Community	GMV
$\mathcal{S}$	T_1	1.103	1.201	1.745
	T_2	0.076	0.124	0.132
	T_3	1.116	1.129	1.076
R_p	T_1	0.149	0.181	0.325
	T_2	0.022	0.036	0.025
	T_3	0.144	0.148	0.113
σ_p^r	T_1	0.135	0.151	0.186
	T_2	0.289	0.290	0.188
	T_3	0.129	0.131	0.105

index for each portfolio of each efficient frontier. Results are summarized in Table 4. For each portfolio on the predicted frontier we track its out-of-sample dynamics and compute the $\mathcal{R}$ index for each of such portfolios. We thus provide the summary statistics of the reliability indices in each time span, for each strategy, and for each type of constraint imposed[5]: when adding constraints on expected returns, sample covariances outperform our methodology, both in time span T_2 (when also the constraint on short-selling is imposed) and in time span T_3, after the first quartile. In time span T_1, instead, the community-based approach outperforms the others. Overall, our approach is confirmed to perform better in all periods when we impose constraints only on expected returns but not on the weights. In particular, optimizing by taking into account the detected clusters stabilize the results, hence providing the best reliability.

Providing a deep explanation for such a result is hard given the higher degree of uncertainty introduced by the constraints on the expected returns. What is clear, however, is that cleaning the correlation matrices from both noise and systemic effects helps to ameliorate the reliability of the portfolios and exploiting stocks communities identified through the mesoscopic correlation further improves the results. The same holds true when constraints on expected returns are imposed but allowing for short-selling strategies. When both constraints on returns and weights are in place; however, the MV approach based on sample covariances is found to be hardly beatable.

Despite having a 'reliable' portfolio is important for an investor in terms of discrepancies between realized and predicted risk, it has to be mentioned that reliability is not the reference metric in portfolio management. For this reason, we complement the analysis by providing the risk-adjusted performances in terms of Sharpe ratios as well. Moreover we also show the ex-post realized risk σ_p^r with respect to the GMV portfolios. As a matter of fact, even if an investor would be happy to invest in a more reliable portfolio, she could reconsider her asset allocation if the price to pay for that reliability is a much higher level of risk in absolute

[5] That is, the table is a summary of the 'realized frontiers' out of sample in terms of $\mathcal{R}$ indices.

terms. In the following we show that this is not the case, with the mesoscopic and community-based optimization schemes yielding absolute volatility levels quite in line with GMV portfolio.

The results are shown in Table 4. The mesoscopic optimization has been shown to track the 1/N strategy; while, the community-based optimization is an 'intra-community 1/N' one. It is well-known in the literature that the 'Talmudic allocation' can be considered as a way to achieve maximum portfolio's diversification. As a consequence, being the portfolio highly diversified, it brings to the investor the benefits also in terms of realized out-sample volatility, with the advantage of being also more reliable.

We then provide a comprehensive analysis in terms of Sharpe ratios, in the same spirit of the analysis done in terms of reliability. In the same spirit of table 3, in table 5 it is possible to appreciate the Sharpe ratios of the compared strategies, again over three different time horizons and considering the possibility to do short-selling or not. As we can see, there is no clear evidence in favor of any given strategy, with results in terms of risk-adjusted performances which depend both on the time horizon and constraint put in the optimization process. Interestingly, we can notice that the community-based asset allocation is particularly stable, in terms of Sharpe ratios, over different constraints (both in terms of returns and short-selling), never yielding a negative return in all the time horizons considered. On the other hand, a simple plug-in approach in a classical MV optimization framework can yield Sharpe ratios which are sensibly higher than the proposed approaches.

Table 5 Summary statistics of the Sharpe ratios $\mathcal{S}$ for the noise plus systemic free, sample covariance MV approach (i.e., standard MV), and community-based efficient frontiers over different periods, with and without short-selling strategies. For each efficient frontier, specific quantiles are taken and the Sharpe ratios of the portfolio located on that point of the frontier are computed

		Min	1st quartile	Median	Mean	3rd quartile
Short-selling						
T_1	$\mathcal{S}_{\text{community}}$	0.98	1.03	1.08	1.08	1.13
	$\mathcal{S}_{\text{mesoscopic}}$	0.7	1.11	1.31	1.23	1.39
	$\mathcal{S}_{\text{standard MV}}$	1.04	1.35	1.57	1.51	1.69
T_2	$\mathcal{S}_{\text{community}}$	0.17	0.22	0.26	0.25	0.29
	$\mathcal{S}_{\text{mesoscopic}}$	−0.07	−0.03	0.02	0.02	0.08
	$\mathcal{S}_{\text{standard MV}}$	−0.23	−0.11	0.12	0.15	0.41
T_3	$\mathcal{S}_{\text{community}}$	0.73	0.84	0.96	0.97	1.1
	$\mathcal{S}_{\text{mesoscopic}}$	−0.01	0.43	0.83	0.76	1.11
	$\mathcal{S}_{\text{standard MV}}$	−0.79	−0.55	−0.20	−0.16	0.21
No short-selling						
T_1	$\mathcal{S}_{\text{community}}$	0.94	1.06	1.08	1.08	1.11
	$\mathcal{S}_{\text{mesoscopic}}$	0.13	0.81	1.13	0.99	1.34
	$\mathcal{S}_{\text{standard MV}}$	0.12	0.83	1.50	1.23	1.76
T_2	$\mathcal{S}_{\text{community}}$	0.22	0.24	0.26	0.26	0.28
	$\mathcal{S}_{\text{mesoscopic}}$	0.07	0.08	0.17	0.24	0.4
	$\mathcal{S}_{\text{standard MV}}$	0.08	0.13	0.19	0.26	0.4
T_3	$\mathcal{S}_{\text{community}}$	0.93	0.94	0.99	1.02	1.06
	$\mathcal{S}_{\text{mesoscopic}}$	0.81	1.11	1.13	1.11	1.16
	$\mathcal{S}_{\text{standard MV}}$	0.63	0.93	1.06	0.99	1.09

This leads us to the final discussions and conclusions.

6 Discussion and conclusions

In this work we investigated the mesoscopic structure of the stock market correlations that emerge after filtering out both microscopic (stock-specific noise) and macroscopic (market-wide trends) components. We showed that such mesoscopic correlations are the most stable over time, thereby encoding important information in the context of portfolio optimization. Indeed, we found that the noisy and the systemic components of the stock market are unstable, leading to biased and poor out-of-sample performances and being surprisingly responsible for the departure of the sample covariance GMV investment prescription from the heuristic, equally weighted strategy. Upon filtering out these unstable components, the market can be partitioned into internally positively and mutually negatively correlated communities of stocks. We proposed to use these stable mesoscopic communities to construct portfolios characterized by higher levels of reliability in terms of predicted and realized risk.

Results can be summarized as follows. The adoption of 'noise- and systemic free' correlations leads to an asset allocation which closely tracks the reliability of the heuristic equally weighted portfolio. That is, once the investor takes into account random co-movement as well as that induced by the presence of common aggregate fluctuations, the heuristic strategy turns out to be empirically optimal at the end. In addition, both the equally weighted portfolios and the ones induced by the proposed optimization scheme have been found to be more reliable than the sample covariance plug-in portfolios. Importantly, portfolio reliability can be further improved by performing a mesoscopic optimization while simultaneously accounting for the community to which a given stock belongs: This is especially true when short selling is allowed, that is, we can go long on some communities while going short on others. Only when constraints on both weights and expected returns are imposed, the homogeneous community-based portfolios do not bring improvements compared to classical approach—with the exception of the period $T_1 = 2000 - 2007$ and for few specific levels of targeted expected returns. Most important, we have shown that the portfolios created following our methodology do not systematically reduce the performances in terms of realized risk out-of-sample or Sharpe ratios. This aspect could foster new research on clustering based trading strategies characterized by higher levels of reliability, since higher reliability does not necessarily imply higher levels of risk.

To conclude, the proposed methodology works well when focusing on the minimum variance portfolio or when short-selling can be performed, suggesting the adoption of network clustering techniques for risk management applications. The uncovered mesoscale structure might bring insights about additional, and complementary, ways of creating stock market indices to monitor market trends - something which might be the object of further studies aimed at understanding co-movements between industries and sectors in the stock market.

Appendix A The 2-asset case

Consider an investor who splits her wealth between $N = 2$ assets and want to minimize the variance of her investment. The problem to solve simply is

$$\min_{\omega_1} \quad \omega_1^2\sigma_1^2 + (1-\omega_1)^2\sigma_2^2 + 2\omega_1(1-\omega_1)C_{12}\sigma_1\sigma_2, \tag{A1}$$

whose first order condition is

$$2\omega_1\sigma_1^2 - 2(1-\omega_1)\sigma_2^2 + 2(1-2\omega_1)C_{12}\sigma_1\sigma_2 = 0 \tag{A2}$$

implying the following optimal wealth allocation with respect to asset 1

$$\omega_1^* = \frac{\sigma_2^2 - C_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2C_{12}\sigma_1\sigma_2}. \tag{A3}$$

Given the decomposition in (8), we know that the noise-free correlation coefficients and covariances are $C_{ij} = C_{ij}^{(g)} + C_{ij}^{(m)}$ and $\sigma_{ij} = \sigma_{ij}^{(g)} + \sigma_{ij}^{(m)}$, we thus write

$$\omega_1^* = \frac{\sigma_2^2 - \left(\sigma_{12}^{(m)} + \sigma_{12}^{(g)}\right)}{\sigma_1^2 + \sigma_2^2 - 2\left(\sigma_{12}^{(m)} + \sigma_{12}^{(g)}\right)}. \tag{A4}$$

For a risk minimizer investor who filters out the systemic induced covariances being aware of its temporarily nature, or equivalently in absence of significant systemic co-movements, the optimal adjusted weight is the one obtained using the mesoscopic covariances

$$\omega_1^{adj} = \frac{\sigma_2^2 - C_{12}^{(g)}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2C_{12}^{(g)}\sigma_1\sigma_2}. \tag{A5}$$

This difference can be easily quantified taking $\Delta\omega_1^* = \omega_1^* - \omega_1^{adj}$, which after some manipulation and terms rearranging yields

$$\Delta\omega_1^* = \frac{2\sigma_{12}^{(m)}\sigma_2^2 \; - \; \sigma_{12}^{(m)}\left(\sigma_1^2 + \sigma_2^2\right)}{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_{12}^{(g)}\left(\sigma_1^2 + \sigma_2^2 + \sigma_{12}^{(m)} + \sigma_{12}^{(g)}\right)}. \tag{A6}$$

If $\sigma_{12}^{(m)} = 0 \rightarrow \Delta\omega_1^* = 0$ and no difference in the wealth allocation occurs.

Otherwise, $\sigma_{12}^{(m)} > 0 \rightarrow \Delta\omega_1^* > 0$ if $\sigma_2^2 > \sigma_1^2$, which clarify the rebalancing of the portfolio stated in the paper and empirically displayed.

Appendix B GMV decomposition

Consider the solution of the GMV portfolio

$$w_{gmv} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{\Sigma}^{-1}\mathbf{1}}, \tag{B7}$$

and the spectral decomposition of the covariance matrix

$$\mathbf{\Sigma}^{-1} = \boldsymbol{PD}^{-1}\boldsymbol{P}^{-1}. \tag{B8}$$

where $\boldsymbol{D}$ is the diagonal matrix from which we are able to identify the eigenvalues associated to random covariances exploiting the *MP-Law*, and the biggest one associated to the systemic component. Thus, $\boldsymbol{D}$ can be splitted as

$$\boldsymbol{D} = \boldsymbol{D}^{(r)} + \boldsymbol{D}^{(g)} + \boldsymbol{D}^{(m)} \tag{B9}$$

and its inverse can be obtained by simply replacing each non zero element in the main diagonal (i.e., eigenvalues) with its reciprocal, having

$$\boldsymbol{D}^{-1} = \boldsymbol{D}_{(r)}^{-1} + \boldsymbol{D}_{(g)}^{-1} + \boldsymbol{D}_{(m)}^{-1}. \tag{B10}$$

Combining the above equations we get

$$\mathbf{\Sigma}^{-1} = \boldsymbol{PD}_r^{-1}\boldsymbol{P}^{-1} + \boldsymbol{PD}_g^{-1}\boldsymbol{P}^{-1} + \boldsymbol{PD}_m^{-1}\boldsymbol{P}^{-1}$$

$$= \mathbf{\Sigma}_r^{-1} + \mathbf{\Sigma}_g^{-1} + \mathbf{\Sigma}_m^{-1}$$

which allows to split the GMV solution as

$$w_{gmv} = \frac{\Sigma_r^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}} + \frac{\Sigma_g^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}} + \frac{\Sigma_m^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}}$$

$$= w_{gmv}^{(r)} + w_{gmv}^{(g)} + w_{gmv}^{(m)}.$$

Appendix C Community-based optimization procedure

Consider the variance of the portfolio

$$\sigma_p^2 = \sum_{i=1}^{N} \omega_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{i\neq j} \omega_i\omega_j\sigma_{ij}. \tag{C11}$$

Remember that $N = N_1 + N_2 + \cdots + N_n$ is the total number of asset, n the number of detected communities, and N_c the number of assets in a given community denoted by the subscript $c \in \{1, 2 \ldots n\}$. We also drop the superscript (g) taking for granted that we always refer to the covariance between assets already filtered from both noise and systemic effects. Maximizing with respect to the n detected communities, so to

have homogeneous weights inside a given community, can be achieved by splitting the variance of the portfolio as follows

$$\begin{aligned}\sigma_p^2 = &\sum_{i=1}^{N_1}\omega_1^2\sigma_{i1}^2 + \sum_{i=1}^{N_2}\omega_2^2\sigma_{i2}^2 + \cdots + \sum_{i=1}^{N_n}\omega_n^2\sigma_{in}^2 \\ &+ \sum_{i=1}^{N_1}\sum_{i\neq j}\omega_1^2\sigma_{ij1} + \sum_{i=1}^{N_2}\sum_{i\neq j}\omega_2^2\sigma_{ij2} + \cdots + \sum_{i=1}^{N_n}\sum_{i\neq j}\omega_n^2\sigma_{ijn} \\ &+ \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\omega_1\omega_2\sigma_{ij12} + \sum_{i=1}^{N_1}\sum_{j=1}^{N_3}\omega_1\omega_3\sigma_{ij13} + \cdots + \sum_{i=1}^{N_1}\sum_{j=1}^{N_n}\omega_1\omega_n\sigma_{ij1n} \\ &+ \sum_{i=1}^{N_2}\sum_{j=1}^{N_3}\omega_2\omega_3\sigma_{ij23} + \cdots + \sum_{i=1}^{N_2}\sum_{j=1}^{N_n}\omega_2\omega_n\sigma_{ij2n} \\ &\vdots \\ &+ \sum_{i=1}^{N_{n-1}}\sum_{j=1}^{N_n}\omega_{n-1}\omega_n\sigma_{ij(n-1)n}\end{aligned} \tag{C12}$$

which is equivalent to

$$\sigma_p^2 = \sum_{c=1}^{n}\omega_c^2 N_c\bar{\sigma}_c^2 + \sum_{c=1}^{n}\omega_c^2 N_c(N_c-1)\bar{\sigma}_{ijc} + \sum_{c=1}^{n-1}\sum_{k=c+1}^{n}2\omega_c\omega_k N_c N_k\bar{\sigma}_{ck} \tag{C13}$$

Thus the objective function to minimize with respect to the community weights become

$$\sigma_p^2 = \sum_{c=1}^{n}\omega_c^2\left[N_c\bar{\sigma}_c^2 + N_c(N_c-1)\bar{\sigma}_{ijc}\right] + \sum_{c=1}^{n-1}\sum_{k=c+1}^{n}2\omega_c\omega_k\left[N_c N_k\bar{\sigma}_{ck}\right] \tag{C14}$$

with $W_c = \omega_c N_c$ being the total share of wealth invested in community c.

Acknowledgements The authors would like to thank Giulio Bottazzi, Daniele Giachini, Manuel Luci and the participants to the CCS 2021 and CEF 2023 conferences for the insightful comments. We are also very grateful to the anonymous referee and the Editor whose comments helped us to achieve a better version of the paper. This work has been supported by the European Union—NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR), project 'SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics'—Grant IR0000013 (n. 3264, 28/12/2021). This work has been also supported by the project NetRes—'Network analysis of economic and financial resilience', Italian DM n. 289, 25-03-2021 (PRO3 Scuole), CUP D67G22000130001 (https://netres.imtlucca.it).

Funding Open access funding provided by Scuola Normale Superiore within the CRUI-CARE Agreement.

Data availability The datasets and codes are available from the corresponding author on reasonable request.

Declarations

Conflict of interest None.

References

Anagnostou I, Squartini T, Kandhai D, Garlaschelli D (2021) Uncovering the mesoscale structure of the credit default swap market to improve portfolio risk modelling. Quant Financ 21:1–18

Bai Z, Liu H, Wong WK (2009) Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. Math Financ: Int J Math, Stat Financ Econ 19(4):639–667

Bai ZD (1999) Methodologies in spectral analysis of large dimensional random matrices, a review. Stat Sin 9(3):611–662

Biely C, Thurner S (2008) Random matrix ensembles of time-lagged correlation matrices: derivation of eigenvalue spectra and analysis of financial time-series. Quant Financ 8(7):705–722

Billio M, Getmansky M, Lo AW, Pelizzon L (2012) Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J Financ Econ 104(3):535–559

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):P10008

Bonanno G, Caldarelli G, Lillo F, Mantegna RN (2003) Topology of correlation-based minimal spanning trees in real and model markets. Phys Rev E 68(4):046130

Bonanno G, Caldarelli G, Lillo F, Micciche S, Vandewalle N, Mantegna RN (2004) Networks of equities in financial markets. Eur Phys J B 38(2):363–371

Borghesi C, Marsili M, Miccichè S (2007) Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. Phys Rev E 76:026104

Bouchaud JP, Potters M (2003) Theory of financial risk and derivative pricing: from statistical physics to risk management. Cambridge University Press, Cambridge

Britten-Jones M (1999) The sampling error in estimates of mean-variance efficient portfolio weights. J Financ 54(2):655–671

Brodie J, Daubechies I, De Mol C, Giannone D, Loris I (2009) Sparse and stable markowitz portfolios. Proc Natl Acad Sci 106(30):12267–12272

Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111

DeMiguel V, Garlappi L, Nogales FJ, Uppal R (2009) A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. Manage Sci 55(5):798–812

DeMiguel V, Garlappi L, Uppal R (2009) Optimal versus naive diversification: how inefficient is the 1/n portfolio strategy? Rev Financ Stud 22(5):1915–1953

Di Matteo T, Aste T, Mantegna R (2004) An interest rates cluster analysis. Physica A 339(1–2):181–188

Dimov II, Kolm PN, Maclin L, Shiber DY (2012) Hidden noise structure and random matrix models of stock correlations. Quant Financ 12(4):567–572

Duchin R, Levy H (2009) Markowitz versus the talmudic portfolio diversification strategies. J Portf Manage 35(2):71–74

Eichhorn D, Gupta F, Stubbs E (1998) Using constraints to improve the robustness of asset allocation. J Portf Manag 24(3):41

Fenn DJ, Porter MA, Mucha PJ, McDonald M, Williams S, Johnson NF, Jones NS (2012) Dynamical clustering of exchange rates. Quant Financ 12(10):1493–1520
Forbes KJ, Rigobon R (2002) No contagion, only interdependence: measuring stock market comovements. J Financ 57(5):2223–2261
Frost PA, Savarino JE (1988) For better performance. J Portf Manage 15(1):29–34
Garlaschelli D, Loffredo MI (2009) Generalized bose-fermi statistics and structural correlations in weighted networks. Phys Rev Lett 102(3):038701
Jagannathan R, Ma T (2003) Risk reduction in large portfolios: why imposing the wrong constraints helps. J Financ 58(4):1651–1683
Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. Phys Rev Lett 83(7):1467
Laloux L, Cizeau P, Potters M, Bouchaud JP (2000) Random matrix theory and financial correlations. Int J Theor Appl Financ 3(03):391–397
MacMahon M, Garlaschelli D (2015) Community detection for correlation matrices. Phys Rev X 5:021006
Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B-Condens Matter Complex Syst 11(1):193–197
Marčenko VA, Pastur LA (1967) DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. Math USSR-Sbornik 1(4):457–483
Markowitz H (1952) Portfolio selection. J Financ 7(1):77–91
Michaud RO (1989) The markowitz optimization enigma: is 'optimized'optimal? Financ Anal J 45(1):31–42
Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582
Onnela JP, Chakraborti A, Kaski K, Kertész J, Kanto A (2003) Dynamics of market correlations: taxonomy and portfolio analysis. Phys Rev E 68:056110
Onnela JP, Kaski K, Kertész J (2004) Clustering and information in correlation based financial networks. Eur Phys J B 38(2):353–362
Pantaleo E, Tumminello M, Lillo F, Mantegna RN (2011) When do improved covariance matrix estimators enhance portfolio optimization? an empirical comparative study of nine estimators. Quant Financ 11(7):1067–1080
Peralta G, Zareei A (2016) A network approach to portfolio selection. J Empir Financ 38:157–180
Pflug GC, Pichler A, Wozabal D (2012) The 1/n investment strategy is optimal under high model ambiguity. J Bank Financ 36(2):410–417
Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Guhr T, Stanley HE (2002) Random matrix approach to cross correlations in financial data. Phys Rev E 65(6):066126
Singh A, Xu D (2016) Random matrix application to correlations amongst the volatility of assets. Quant Financ 16(1):69–83
Tola V, Lillo F, Gallegati M, Mantegna RN (2008) Cluster analysis for portfolio optimization. J Econ Dyn Control 32(1):235–258
Tu J, Zhou G (2011) Markowitz meets talmud: a combination of sophisticated and naive diversification strategies. J Financ Econ 99(1):204–215
Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci 102(30):10421–10426
Verma A, Buonocore RJ, Di Matteo T (2019) A cluster driven log-volatility factor model: a deepening on the source of the volatility clustering. Quant Financ 19(6):981–996
Zitelli G (2020) Random matrix models for datasets with fixed time horizons. Quant Financ 20(5):769–781