INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

Mesoscopic Structure of the Stock Market and Portfolio Optimization

Sebastiano Michele Zema [a]
Giorgio Fagiolo [a]
Tiziano Squartini [b]
Diego Garlaschelli [b,c]

[a] Institute of Economics and EMbeDS, Scuola Superiore Sant'Anna, Pisa, Italy.
[b] Networks, IMT Institute for Advanced Studies, Lucca, Italy.
[c] Lorentz Institute for Theoretical Physics, University of Leiden, Leiden, The Netherlands.

2021/45 **December 2021**
ISSN(ONLINE) 2284-0400

Mesoscopic Structure of the Stock Market and Portfolio Optimization

S.M. Zema[1], G. Fagiolo[1], T. Squartini[2], and D. Garlaschelli[2, 3]

[1]Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, IT
[2]Networks, IMT Institute for Advanced Studies, Lucca, IT
[3]Lorentz Institute for Theoretical Physics, University of Leiden, Leiden, NL

Abstract

The idiosyncratic (microscopic) and systemic (macroscopic) 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 (RMT) 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.

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 a poor out-of-sample performance (Michaud, 1989; Bai et al., 2009). The MV predictions, in fact, seriously depart from empirical observations, thus questioning the need of individuating procedures for investment optimization - especially when 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). Indeed, as Laloux et al. (1999) show, the MV optimization procedure provides a biased estimation of the correlation matrix since the smallest eigenvalues of the latter, which play a fundamental role in the estimation of the global minimum variance (GMV) portfolio, are largely affected by noise.

In order to overcome the limitations of the MV framework, techniques have been proposed either to ameliorate its theoretical predictions 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 and these improvements also depend on other circumstances as the dimension-to-sample size ratio or the possibility of exploiting short-selling strategies.

[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).

One network-based approach exploits the complex, evolving and interconnected 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, an evidence suggesting that optimal portfolios should include peripheral stocks to reduce the influence of central assets characterized by higher levels of 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 the 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 the market structure can be identified by employing an RMT-based clustering technique 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 modelling, 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].

Moving from there, our work makes a step forward and investigates the effects of adjusting the correlation matrix of financial assets by considering not only the amount of correlations induced by noise but also the one induced by systemic co-movements. As documented by Forbes and Rigobon (2002), the correlation coefficient is indeed conditional on market volatility[4], 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 a portfolio reliability for the wealth allocation process: such a goal can be achieved by identifying which market correlations are structural and stable over time, i.e. not resulting from either random co-movements or temporary market effects.

As we will show, optimizing portfolios using the intermediate, mesoscopic level of the spectrum of the correlation matrix yields balanced allocations that improve the reliability of the former ones, in terms of predicted and realized risk, as compared to the standard MV optimization procedure: such an asset allocation closely tracks the heuristic $1/N$ rule but is not sensitive to estimation errors, caused either by random or aggregate systemic fluctuations, affecting correlations; this, in turn, reduces the effective size of portfolios without hampering their performance.

Furthermore, we show that 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), one is able to construct portfolios that are more reliable than those obtained by both the classical MV plug-in estimator[5] and the $1/N$ rule.

The rest of the paper is organized as follows. In section 2 we introduce our filtering procedure and explain how filtered correlations can be exploited to recover the mesoscopic structure of the stock market. In section 3 we show how that information can be exploited in a portfolio optimization setting. Section 4 illustrates the advantages, in terms of predicted and realized risk reliability, of constructing portfolios as above. Section 5 concludes and discusses possible paths for future research.

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

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

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

[5]In what follows, we refer to the historical plug-in estimator for the MV optimization as the 'classical Markowitz' approach.

The reader interested in the proof is redirected to Bai (1999). The result above implies that any empirical correlation matrix $\mathbf{C}$ of financial returns[6], where the largest empirical eigenvalue is denoted as λ_m and is usually (much) larger than $\lambda_{\max}$, can be decomposed as

$$\begin{aligned} \mathbf{C} &= \sum_{i=1}^{N} \lambda_i |v_i\rangle\langle v_i| \\ &= \sum_{i:\lambda_i \in (0,\lambda_{\max}]} \lambda_i |v_i\rangle\langle v_i| + \sum_{i:\lambda_i \in (\lambda_{\max},\lambda_m]} \lambda_i |v_i\rangle\langle v_i| \qquad (2) \\ &= \mathbf{C}^{(r)} + \mathbf{C}^{(s)}, \qquad (3) \end{aligned}$$

where $|v_i\rangle$ and $\langle v_i|$ denote the column and row eigenvectors associated with the eigenvalue λ_i respectively. The above decomposition represents the empirical correlation matrix as a sum of matrices respectively induced by the *random* spectral component $\mathbf{C}^{(r)}$, whose eigenvalues lie in the Marčenko-Pastur range $[\lambda_{\min}, \lambda_{\max}]$ and usually also below $\lambda_{\min}$ (as a result of the fact that, since the trace of a correlation matrix should remain equal to N, the presence of $\lambda_m > \lambda_{\max}$ shifts the lower eigenvalues leftwards) and the *structural* (non-random) component $\mathbf{C}^{(s)}$. The filtering procedure consists in removing $\mathbf{C}^{(r)}$, i.e. the random component of the tensor: what remains is, then, recognized as signal rather than noise, hence supposed to possess useful economic information.

As mentioned above, the spectrum of empirical correlation matrices of financial returns is characterized by a leading eigenvalue λ_m which is much larger than the others. The associated (column) eigenvector $|v_m\rangle$ possesses elements having the same sign and identifies a matrix component $\lambda_m |v_m\rangle\langle v_m|$ affecting all stocks in the same direction and with strong intensity, further inducing the decomposition of $\mathbf{C}^{(s)}$ as follows:

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

i.e. as a sum of a *mesoscopic* spectral component $\mathbf{C}^{(g)}$ and a *systemic* component (or *market mode*) $\mathbf{C}^{(m)}$.

A graphical illustration of this empirical feature is provided in Figure 1 for the S&P500 constituents over the period 2000-2015 which constitute the dataset used in the remainder of the work as well. The systemic component is pervasive and time-varying; hence, 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 100 from the S&P500 constituents, 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 Figure 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 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

[6]Notice that the correlation matrix coincides with the covariance matrix of standardized returns: in this case, $\sigma^2 = 1$ and the range extremes simply read $\lambda_{\min/\max} = (1 \pm \sqrt{N/T})^2$.

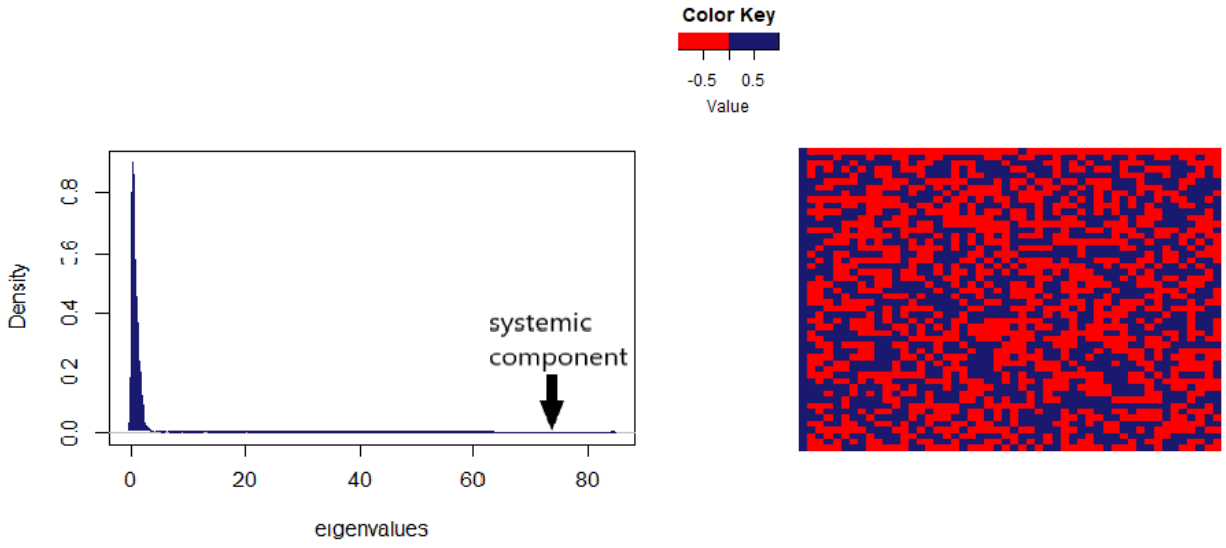

Figure 1: Eigenvalue density for the 2000-2015 covariance matrix of the S&P500 components (left) and heatmap of the associated eigenvectors (right). The first column of the heatmap represents the eigenvector associated to the systemic component, whose elements all have the same sign.

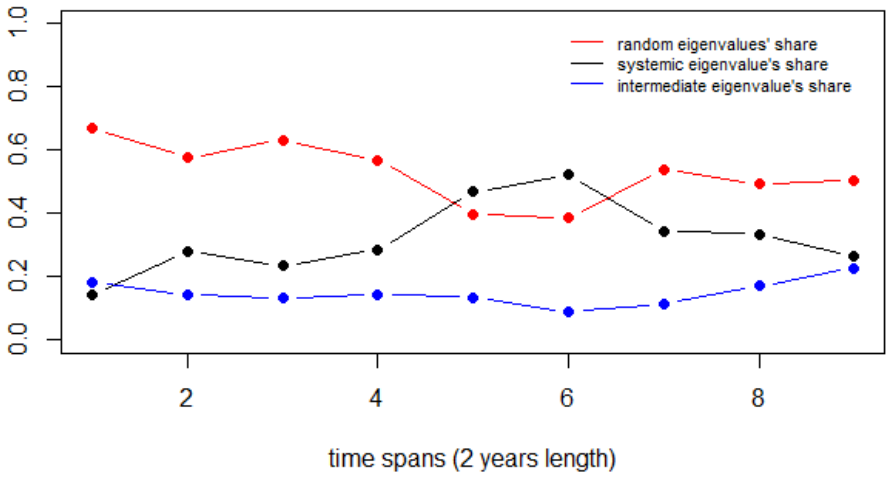

Figure 2: Cumulative risk fractions associated to the different components of the correlation matrix over different time spans. The random and systemic components vary the most (14% standard deviation for both) while the intermediate, mesoscopic range of the spectrum is more stable (5% standard deviation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

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

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 finding 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 in general 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 the modularity $Q(\gamma)$ 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, the 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 equity data for the 450 most capitalized companies in the US, stably traded over the last 20 years, all constituting the S&P500. After applying RMT to isolate the mesoscopic component of the matrix, we performed the modularity maximization procedure by implementing a modified version of the Louvain algorithm (Blondel et al., 2008), taking as input the matrix $\mathbf{C}^{(g)}$. The 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 conducted the above analysis across the 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 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

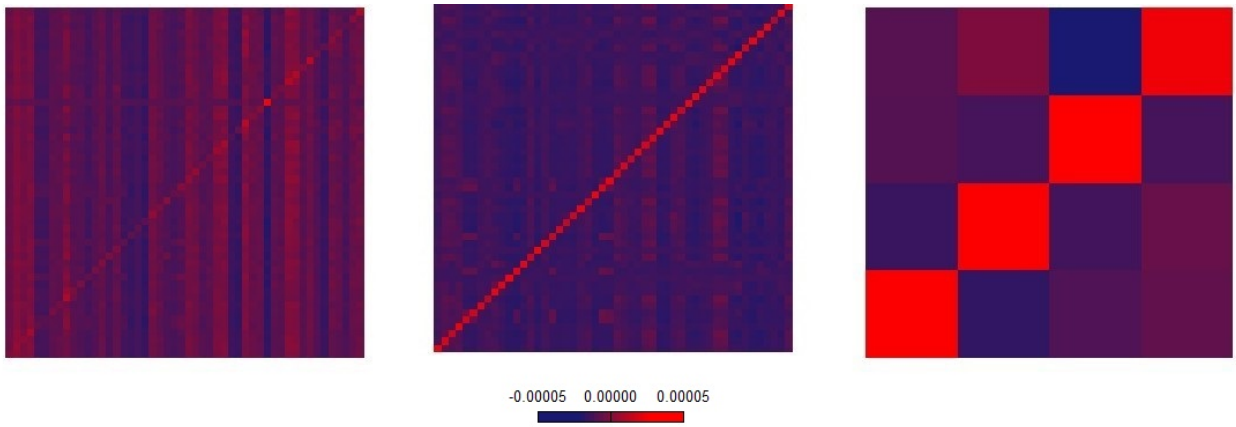

Figure 3: Sequence of transformations applied to the empirical correlation matrix (left), leading first to the 'noise and systemic free' correlation matrix (middle) and then to the internally positively and mutually negatively correlated clusters.

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 the 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 the 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, in turn stabilizing the optimal partition. This can be easily seen by comparing the evolution of the density of the unfiltered, empirical correlations of our sample of stocks with that of the filtered, mesoscopic correlations employed to perform the clustering procedure. As Figure 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 the coefficients calculated over the entire time span are not representative of the underlying sub-periods. By contrast, 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. In this case, the overall distribution is representative of the distributions for the sub-periods.

Since the stability of a distribution does not necessarily imply the stability of the individual entries of the matrix components, as a more stringent test we look at the evolution of the norm of the matrix containing the relative changes of each coefficient. Given the generic entry $C_{ij}^{(f)}$, where f indicates which matrix component we are considering, we define the matrix $\mathbf{\Delta C}^{(f)}$ with entries

$$\Delta C_{ij}^{(f)} = \frac{C_{ij}^{(f)}(t) - C_{ij}^{(f)}(t-1)}{C_{ij}^{(f)}(t-1)}. \tag{10}$$

The above quantity is the relative variation of the correlation coefficient across two consecutive time spans. We focus on relative variations since the components of the correlation matrix are characterized by magnitudes that are so different that a comparison of absolute changes would be meaningless. The stability of each component of the correlation matrix can be inspected by employing the p-norm $||\mathbf{\Delta C}^{(f)}||_p$. As we can see from Table 1, the mesoscopic component $\mathbf{C}^{(g)}$ turns out to be the most stable over time, given its smallest temporal variations.

To further assess the stability of the mesoscopic part of the correlation matrix, we repeat the computation just described by arranging the terms of $\mathbf{C}^{(g)}$ according to the community a given node belongs to, so to have a matrix

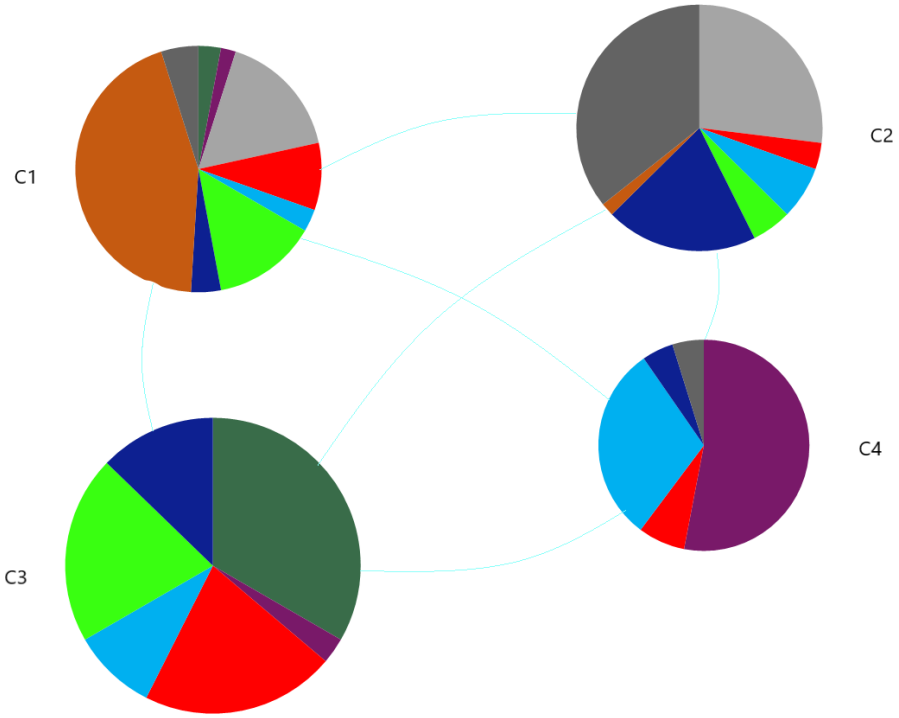

Figure 4: Community structure of the 450 most capitalized stocks of the US stock market during the period January 2000-December 2015: ■ Finance, ■ Energy, ■ Healthcare, ■ Industrials, ■ Materials, ■ Discretionary, ■ Staples, ■ Technology, ■ Utilities.

arranged in blocks, as displayed in Figure 3. Afterwards, to attach equal importance to nodes according to the community they belong to, we replace the values in each given block with the average computed over the block itself and denote this new matrix as $\mathbf{C}^{(\overline{g})}$. Notice that no clear difference arises with respect to the norms computed for $\mathbf{C}^{(g)}$, a result confirming the stability of the detected structures.

Taken together, all the above checks of the stability of the filtered, mesoscopic component of empirical correlations lay the ground for our subsequent analyses in the rest of the paper.

	$\|\|\Delta\mathbf{C}^{(r)}\|\|_1$	$\|\|\Delta\mathbf{C}^{(m)}\|\|_1$	$\|\|\Delta\mathbf{C}^{(g)}\|\|_1$	$\|\|\Delta\mathbf{C}^{(\overline{g})}\|\|_1$
ΔT_1	10.02	2.69	0.59	0.64
ΔT_2	15.51	8.93	1.87	1.83
ΔT_3	8.58	0.81	0.67	0.68

Table 1: Norms of the entry-by-entry relative changes of the correlation coefficients associated to the different components of the matrix. T_1, T_2 and T_3 denote time spans of equal length covering the period 2000-2015. The 1-norm has been chosen for simplicity.

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

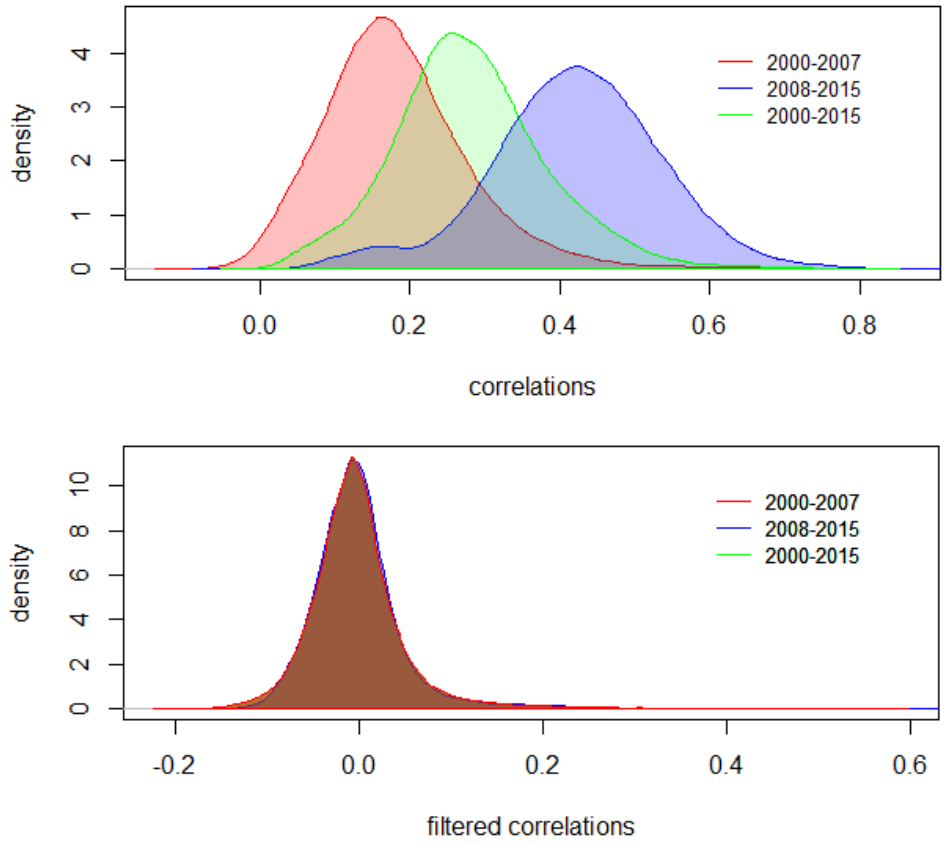

Figure 5: Densities of the unfiltered empirical (top) and filtered mesoscopic (bottom) correlation coefficients for S&P constituents over different periods. Notice that, while a clear shift occurs between the first and the second half of the overall 2000-2015 period for the unfiltered coefficients, no 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.

$$\mu_p = \sum_{i=1}^{N} \omega_i \mu_i \tag{11}$$

with associated variance reading

$$\sigma_p^2 = \sum_{i=1}^{N} \omega_i^2 \sigma_i^2 + \sum_{i>j} 2\omega_i \omega_j C_{ij} \sigma_i \sigma_j. \tag{12}$$

The well-known Markowitz 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

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & \boldsymbol{\omega'\Sigma\omega} \\ \text{s.t.} \quad & \mu_p = \boldsymbol{\omega'\mu} \\ & \sum_{i}^{N} \omega_i = 1 \end{aligned} \tag{13}$$

and has solution

$$\boldsymbol{\omega^*} = \boldsymbol{b\Sigma^{-1}1} + \boldsymbol{c\Sigma^{-1}\mu} \tag{14}$$

with

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

In this work, we focus on the variance and consider a completely risk-adverse investor that is only interested in minimizing the risk with no constraints on the expected return. In that case, the solution simply becomes

$$w_{gmv} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{\Sigma}^{-1}\mathbf{1}} \tag{15}$$

where ω_{gmv} denotes the investment plan associated with the global minimum variance (GMV) portfolio.

We then focus on the *reliability* of an optimal portfolio by comparing its predicted risk, σ_p, obtained via the correlation matrices estimated using historical data, with the (ex-post) realized risk, σ_p^r. As in Tola et al. (2008), we deem a portfolio as reliable if

$$\mathcal{R} = \frac{|\sigma_p^r - \sigma_p|}{\sigma_p} \tag{16}$$

is 'small' - the main difference of our approach being that we will never assume perfect knowledge of future volatilities for the investor, letting uncertainty affect the whole covariance matrix.

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 straightforward to show that, for a risk-minimizer investor, the adoption of the mesoscopic variance $\sigma_{ij}^{(g)} = C_{ij}^{(g)} \sigma_i \sigma_j$, in place of $C_{ij}^{(m)}$ holds) rebalances the portfolio: hence, the total wealth will not be concentrated anymore over few assets characterized by the lowest past variances. In other words, while in presence of co-movements (e.g. because of market turmoils), an 'ingenuous' investor would (try to) 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, for the future.

To provide empirical evidence for such a statement, we performed the MV optimization procedure 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[7]: Figure 6 shows the results, considering both cases in which short selling is either possible or not. Noticeably, the MV optimization based on mesoscopic correlations closely follow the $1/N$ rule, yielding as an optimal solution a portfolio which is very similar to the equally-weighted one; on the contrary, the standard Markowitz optimization framework outputs a much more heterogeneous composition being more sensitive to the estimated sample covariances. This confirms what expected, i.e. 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 Figure 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 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

$$\mathcal{N} = \frac{1}{\sum_{i=1}^{N} \omega_i^2}; \tag{17}$$

indeed, when the wealth is equally divided among the N assets, the quantity $\mathcal{N}$ is equal to N; on the other had, 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 but preserves most of the risk-return properties of the latter.

In Figure 8, the effective size of the portfolios obtained with different allocation rules are displayed and compared: the compared allocation strategies are, again, the Markowitz GMV portfolios, 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 than those output by the classical Markowitz approach and the one based on RMT but filtering out only the random component. As it will be shown afterwards, this result brings non-trivial practical implications in terms of reliability.

4.2 Mesoscopic community-based optimization scheme

The adoption of the 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 taking into account 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 $c \in \{1, 2 \dots n\}$). 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

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & \boldsymbol{\omega}' \boldsymbol{\Sigma}^{(g)} \boldsymbol{\omega} \\ \text{s.t.} \quad & \mu_p = \boldsymbol{\omega}' \boldsymbol{\mu} \\ & \sum_{c}^{n} W_c = 1 \\ & \omega_i = \omega_j \quad \forall \ i, j \in c. \end{aligned} \tag{18}$$

and is the same as the problem in 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

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

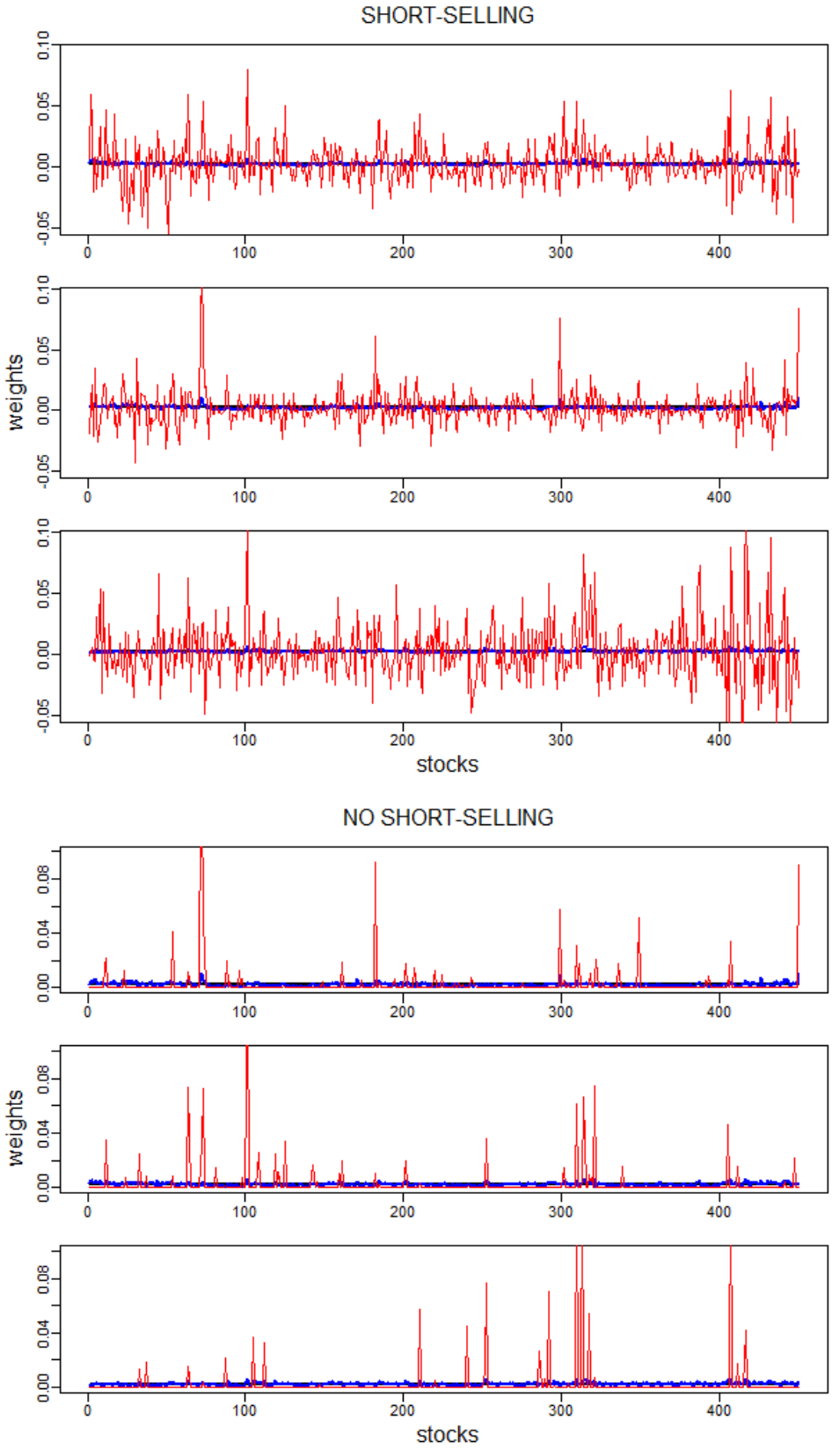

Figure 6: Asset composition comparison for the periods 2000-2003 (top), 2004-2007 (middle) and 2008-2011 (bottom) between the $1/N$ rule (horizontal black line), classical Markowitz (red) and the portfolio optimization based on mesoscopic correlations (blue). For each stock on the x-axis, the relative 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 $\omega_i \geq 0$, $\forall\, i$.

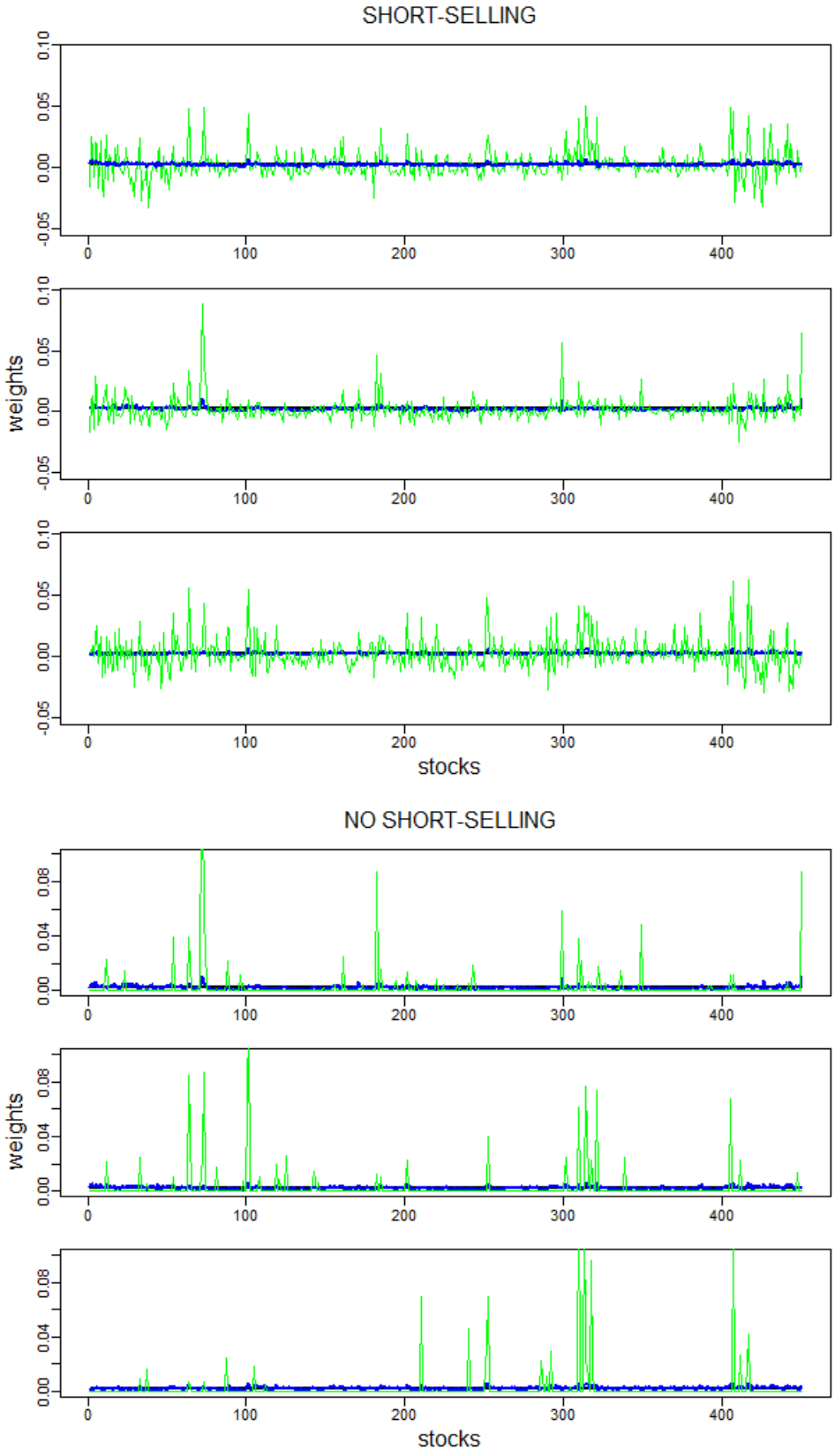

Figure 7: Asset composition comparison for the periods 2000-2003 (top), 2004-2007 (middle) and 2008-2011 (bottom) between the $1/N$ rule (horizontal black line), RMT approach (green) and the portfolio optimization based on mesoscopic correlations (blue). For each stock on the x-axis, the relative weight 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.

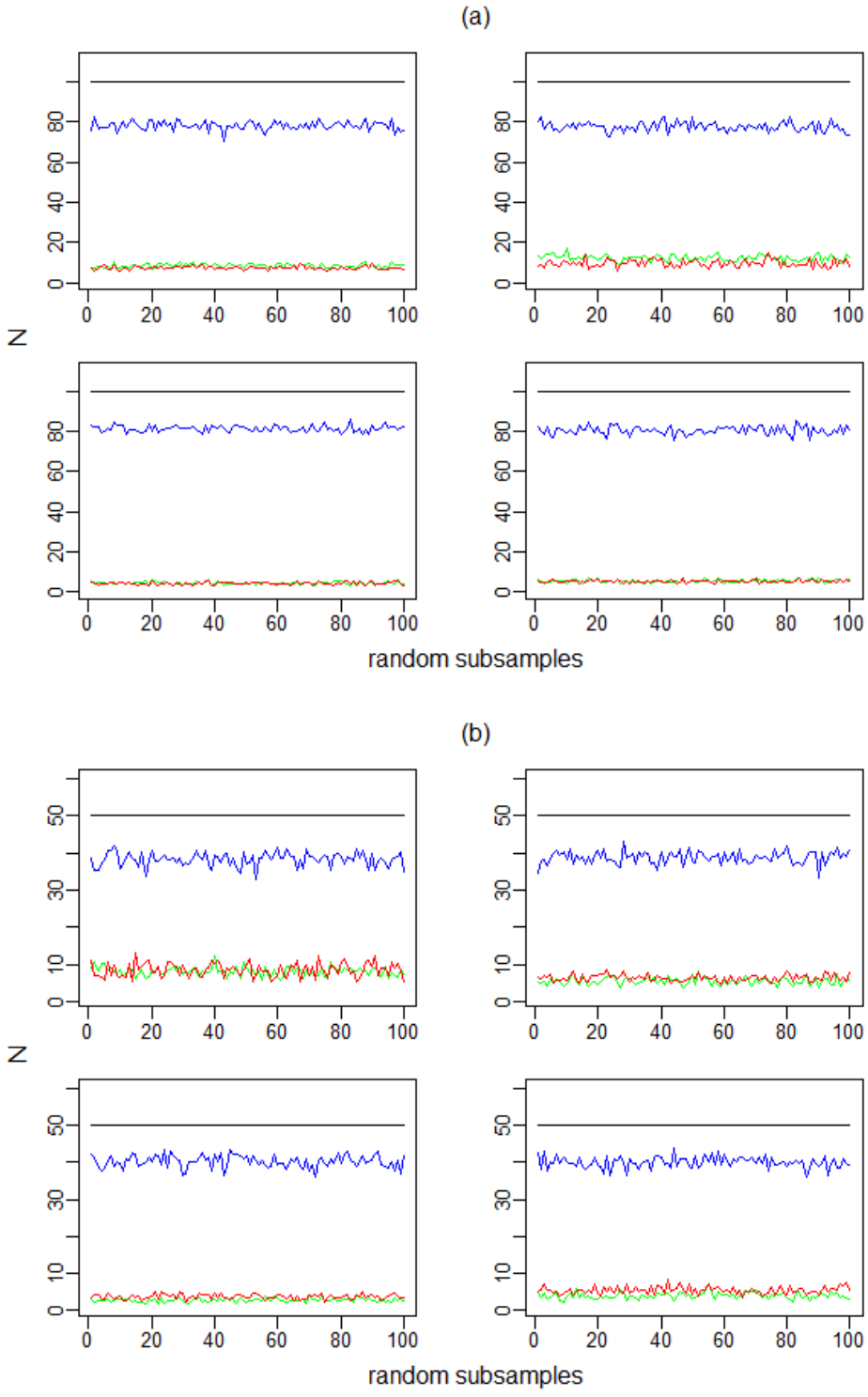

Figure 8: Effective sizes $\mathcal{N}$ for each of the 100 random subsamples. In panel (a), the size of the subsamples is 100, while in panel (b) is 50. The subsamples are of length $T = 3$ years with the plots covering together, inside each panel, the 2000-2012. Mesoscopic GMV portfolios in blue, RMT filtered in green, classical plug-in Markowitz in red.

			Short-selling		No short-selling	
	Time span	$\mathcal{R}_{\text{equally}}$	$\mathcal{R}_{\text{mesoscopic}}$	$\mathcal{R}_{\text{Markowitz}}$	$\mathcal{R}_{\text{mesoscopic}}$	$\mathcal{R}_{\text{Markowitz}}$
$N = 50$	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	**0.11**	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: 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 ($N = 450$). No randomization occured when the whole sample is taken. The measures $\mathcal{R}_{\text{mesoscopic}}$ and $\mathcal{R}_{\text{Markowitx}}$ refer to the GMV portfolio cases with and without short-selling strategies. Entries in bold refer to the strategies with the best performances, both with and without short-selling constraints. Whenever the performance is worse than the equally weighted portfolio, the relative entry is not highlighted.

into each community, must still sum up to one. Reformulating the problem as in 16 leads to the minimization of the following objective function, where no constraint on the expected return is placed:

$$\sigma_p^2 = \sum_{c=1}^{n} \omega_c^2 \left[N_c \overline{\sigma}_c^2 + N_c(N_c - 1)\overline{\sigma}_{ijc}^{(g)} \right] + \sum_{c=1}^{n-1} \sum_{k=c+1}^{n} 2\omega_c \omega_k \left[N_c N_k \overline{\sigma}_{ck}^{(g)} \right]; \tag{19}$$

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 analysis

Let us now compare the reliability $\mathcal{R}$ of the portfolios obtained by implementing the classical Markowitz portfolio optimization approach, the mesoscopic-based optimization approach, the mesoscopic plus community-based optimization approach 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. Then, for the sake of completeness, we repeat the comparison considering 30 values of expected portfolio return μ_p to insert, as additional constraints, on the minimization problem. The different portfolios constituting the efficient frontier are, then, computed, and for each of them the index $\mathcal{R}$ is obtained, 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 subperiods of equal length by fixing t_0. Upon doing so, we create porfolios 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 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 Table 2. When short-selling is allowed and no constraint on the expected portfolio return is present, the Markowitz approach 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,

	Time span	$\mathcal{R}_{\text{equally}}$	$\mathcal{R}^*_{\text{mesoscopic}}$	$\mathcal{R}_{\text{community}}$	$\mathcal{R}^*_{\text{Markowitz}}$
$\mathcal{R}$	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**

Table 3: Comparison between the community-based portfolios and the other methodologies. With $\mathcal{R}^*_{\text{mesoscopic}}$ and $\mathcal{R}^*_{\text{Markowitz}}$ we indicate that we chose the best performance between the short-selling and no short-selling cases

our methodology performs slightly better than the equally-weighted portfolio, thus revealing it to be the most reliable investment plan considered: the difference, however, is almost negligible, a result confirming the closeness of the mesoscopic-based optimization procedure and the equally-weighted strategy. From a purely mathematical perspective, imposing constraints is equivalent at letting a shrinkage operator act on the covariance matrix of the assets, an operation helping when the number of parameters to estimate is too large - and, as a consequence, estimation errors are large as well.

The poor performance of the GMV Markowitz 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). Our empirical analysis confirms these results: notice the huge improvement of the Markowitz approach compared to the situation in which short-selling was not allowed - although the mesoscopic-based optimization provieds 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 with the best performing competing approach, in each time span. Results in Table 3 noticeably confirm the informativeness of the detected communities from a risk-management perspective. Portfolios in which optimal weights are recovered by constraining stocks in the same community to weigh the same further improve their reliability indices, outperforming both the equally-weighted and the mesoscopic-based strategies for all considered time spans. Still, Markowitz with the no short-selling constraint is the more reliable in T_3.

Let us now consider all approaches, i.e. the classical Markowitz one, the mesoscopic-based one and the mesoscopic plus community-based one and compute the reliability index for each portfolio of each efficient frontier. Results are summarized in Table 4, where the $\mathcal{R}$ indices are ordered and compared between the different quartiles of the expected return distribution [8]: when adding constraints on expected returns, Markowitz outperforms our methodology, in time span T_2, when also the constraint on short-selling is imposed and in time span T_3 after the first quartile of the distributions? In time span T_1, instead, the community-based approach outperform 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 minimum variance 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, Markowitz approach is found to be hardly beatable.

[8] For each time span we take the historical expected return distribution of our assets (i.e. in-sample averages) and use the quartiles of the latter as input for the expected return constraints in the optimization problem.

Short-selling						
		min.	1st quartile	median	mean	3rd quartile
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
	$\mathcal{R}_{\text{Markowitz}}$	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
	$\mathcal{R}_{\text{Markowitz}}$	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
	$\mathcal{R}_{\text{Markowitz}}$	0.88	1.07	1.4	1.51	1.85
No short-selling						
		min.	1st quartile	median	mean	3rd quartile
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
	$\mathcal{R}_{\text{Markowitz}}$	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
	$\mathcal{R}_{\text{Markowitz}}$	**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
	$\mathcal{R}_{\text{Markowitz}}$	0.52	0.53	**0.54**	**0.56**	**0.59**

Table 4: Summary statistics of the reliability $\mathcal{R}$ indexes for the noise plus systemic free, classical Markowitz, and community-based efficient frontiers over different periods, with and without short selling strategies.

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 responsible for the surprising departure of the classical Markowitz 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, and slightly outperforms, the reliability of the heuristic equally weighted portfolio, while at the same time requiring a smaller number of assets over which the wealth need be effectively invested. 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 Markowitz plug-in estimator. Importantly, the reliability of portfolios can be further improved by performing the mesoscopic optimization while simultaneously accounting for the community to which a given stock belongs: remarkably, also when constraints on short-selling are imposed, this new methodology performs noticeably better than classical Markowitz.

Only when constraints on both weights and expected returns are imposed, the homogeneous community-based portfolios do not bring improvements compared to classical Markowitz - with the exception of the period $T_1 = 2000 - 2007$ and for few specific levels of targeted expected returns. Thus, 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. In particular, 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.

References

Anagnostou, I., T. Squartini, D. Kandhai, and D. Garlaschelli (2021). Uncovering the mesoscale structure of the credit default swap market to improve portfolio risk modelling. *Quantitative Finance*, 1–18.

Bai, Z., H. Liu, and W.-K. Wong (2009). Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics 19*(4), 639–667.

Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica 9*(3), 611–662.

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

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics 104*(3), 535–559.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*(10), P10008.

Bonanno, G., G. Caldarelli, F. Lillo, and R. N. Mantegna (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E 68*(4), 046130.

Bonanno, G., G. Caldarelli, F. Lillo, S. Micciche, N. Vandewalle, and R. N. Mantegna (2004). Networks of equities in financial markets. *The European Physical Journal B 38*(2), 363–371.

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

Bouchaud, J.-P. and M. Potters (2003). *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge university press.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance 54*(2), 655–671.

Clauset, A., M. E. Newman, and C. Moore (2004). Finding community structure in very large networks. *Physical review E 70*(6), 066111.

Di Matteo, T., T. Aste, and R. Mantegna (2004). An interest rates cluster analysis. *Physica A: Statistical Mechanics and its Applications 339*(1-2), 181–188.

Dimov, I. I., P. N. Kolm, L. Maclin, and D. Y. Shiber (2012). Hidden noise structure and random matrix models of stock correlations. *Quantitative Finance 12*(4), 567–572.

Duchin, R. and H. Levy (2009). Markowitz versus the talmudic portfolio diversification strategies. *The Journal of Portfolio Management 35*(2), 71–74.

Eichhorn, D., F. Gupta, and E. Stubbs (1998). Using constraints to improve the robustness of asset allocation. *Journal of Portfolio Management 24*(3), 41.

Fenn, D. J., M. A. Porter, P. J. Mucha, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones (2012). Dynamical clustering of exchange rates. *Quantitative Finance 12*(10), 1493–1520.

Forbes, K. J. and R. Rigobon (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance 57*(5), 2223–2261.

Frost, P. A. and J. E. Savarino (1988). For better performance. *The Journal of Portfolio Management 15*(1), 29–34.

Garlaschelli, D. and M. I. Loffredo (2009). Generalized bose-fermi statistics and structural correlations in weighted networks. *Physical review letters 102*(3), 038701.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance 58*(4), 1651–1683.

Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters (1999). Noise dressing of financial correlation matrices. *Physical review letters 83*(7), 1467.

Laloux, L., P. Cizeau, M. Potters, and J.-P. Bouchaud (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance 3*(03), 391–397.

MacMahon, M. and D. Garlaschelli (2015, Apr). Community detection for correlation matrices. *Phys. Rev. X 5*, 021006.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems 11*(1), 193–197.

Marčenko, V. A. and L. A. Pastur (1967, apr). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik 1*(4), 457–483.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial Analysts Journal 45*(1), 31–42.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*(23), 8577–8582.

Onnela, J.-P., A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto (2003, Nov). Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E 68*, 056110.

Onnela, J.-P., K. Kaski, and J. Kertész (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B 38*(2), 353–362.

Pantaleo, E., M. Tumminello, F. Lillo, and R. N. Mantegna (2011). When do improved covariance matrix estimators enhance portfolio optimization? an empirical comparative study of nine estimators. *Quantitative Finance 11*(7), 1067–1080.

Peralta, G. and A. Zareei (2016). A network approach to portfolio selection. *Journal of Empirical Finance 38*, 157–180.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley (2002). Random matrix approach to cross correlations in financial data. *Physical Review E 65*(6), 066126.

Singh, A. and D. Xu (2016). Random matrix application to correlations amongst the volatility of assets. *Quantitative Finance 16*(1), 69–83.

Tola, V., F. Lillo, M. Gallegati, and R. N. Mantegna (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control 32*(1), 235 – 258. Applications of statistical physics in economics and finance.

Tumminello, M., T. Aste, T. Di Matteo, and R. N. Mantegna (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences 102*(30), 10421–10426.

Verma, A., R. J. Buonocore, and T. Di Matteo (2019). A cluster driven log-volatility factor model: a deepening on the source of the volatility clustering. *Quantitative Finance 19*(6), 981–996.

Zitelli, G. (2020). Random matrix models for datasets with fixed time horizons. *Quantitative Finance 20*(5), 769–781.

A Brief analytical clarifications with 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

$$\min_{\boldsymbol{\omega_1}} \ \omega_1^2\sigma_1^2 + (1-\omega_1)^2\sigma_2^2 + 2\omega_1(1-\omega_1)C_{12}\sigma_1\sigma_2, \tag{20}$$

whose first order condition is

$$2\omega_1\sigma_1^2 - 2(1-\omega_1)\sigma_2^2 + 2(1-2\omega_1)C_{12}\sigma_1\sigma_2 = 0 \tag{21}$$

implying the following optimal wealth allocation with respect to asset 1

$$\omega_1^* = \frac{\sigma_2^2 - C_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2C_{12}\sigma_1\sigma_2}. \tag{22}$$

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

$$\omega_1^* = \frac{\sigma_2^2 - (\sigma_{12}^{(m)} + \sigma_{12}^{(g)})}{\sigma_1^2 + \sigma_2^2 - 2(\sigma_{12}^{(m)} + \sigma_{12}^{(g)})}. \tag{23}$$

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 comovements, the optimal adjusted weight is the one obtained using the mesoscopic covariances

$$\omega_1^{adj} = \frac{\sigma_2^2 - C_{12}^{(g)}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2C_{12}^{(g)}\sigma_1\sigma_2}. \tag{24}$$

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

$$\Delta\omega_1^* = \frac{2\sigma_{12}^{(m)}\sigma_2^2 \; - \; \sigma_{12}^{(m)}(\sigma_1^2 + \sigma_2^2)}{(\sigma_1^2 + \sigma_2^2)^2 - 4\sigma_{12}^{(g)}(\sigma_1^2 + \sigma_2^2 + \sigma_{12}^{(m)} + \sigma_{12}^{(g)})}. \tag{25}$$

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.

B Technical steps GMV decomposition

Consider the solution of the GMV portfolio

$$w_{gmv} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^t\mathbf{\Sigma}^{-1}\mathbf{1}}, \tag{26}$$

and the spectral decomposition of the covariance matrix

$$\mathbf{\Sigma}^{-1} = \boldsymbol{PD^{-1}P^{-1}}. \tag{27}$$

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

$$\boldsymbol{D} = \boldsymbol{D^{(r)}} + \boldsymbol{D^{(g)}} + \boldsymbol{D^{(m)}} \tag{28}$$

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

$$\boldsymbol{D^{-1}} = \boldsymbol{D_{(r)}^{-1}} + \boldsymbol{D_{(g)}^{-1}} + \boldsymbol{D_{(m)}^{-1}}. \tag{29}$$

Combining the above equations we get

$$\begin{aligned}\mathbf{\Sigma^{-1}} &= \boldsymbol{PD_r^{-1}P^{-1}} + \boldsymbol{PD_g^{-1}P^{-1}} + \boldsymbol{PD_m^{-1}P^{-1}} \\ &= \mathbf{\Sigma_r^{-1}} + \mathbf{\Sigma_g^{-1}} + \mathbf{\Sigma_m^{-1}}\end{aligned}$$

which allows to split the GMV solution as

$$\begin{aligned}w_{gmv} &= \frac{\Sigma_r^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}} + \frac{\Sigma_g^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}} + \frac{\Sigma_m^{-1}\mathbf{1}}{\mathbf{1}^t\Sigma^{-1}\mathbf{1}} \\ &= w_{gmv}^{(r)} + w_{gmv}^{(g)} + w_{gmv}^{(m)}.\end{aligned}$$

C Details on the implementation of the community-based optimization procedure

Consider the variance of the portfolio

$$\sigma_p^2 = \sum_{i=1}^{N}\omega_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{i\neq j}\omega_i\omega_j\sigma_{ij}. \tag{30}$$

Remember that $N = N_1 + N_2 + ... + N_n$ 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 $c \in \{1, 2, ...n\}$. 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

$$\begin{aligned}
\sigma_p^2 \quad &= \sum_{i=1}^{N_1} \omega_1^2 \sigma_{i1}^2 + \sum_{i=1}^{N_2} \omega_2^2 \sigma_{i2}^2 + \cdots + \sum_{i=1}^{N_n} \omega_n^2 \sigma_{in}^2 \\
&+ \sum_{i=1}^{N_1} \sum_{i \neq j} \omega_1^2 \sigma_{ij1} + \sum_{i=1}^{N_2} \sum_{i \neq j} \omega_2^2 \sigma_{ij2} + \cdots + \sum_{i=1}^{N_n} \sum_{i \neq j} \omega_n^2 \sigma_{ijn} \\
&+ \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \omega_1 \omega_2 \sigma_{ij12} + \sum_{i=1}^{N_1} \sum_{j=1}^{N_3} \omega_1 \omega_3 \sigma_{ij13} + \cdots + \sum_{i=1}^{N_1} \sum_{j=1}^{N_n} \omega_1 \omega_n \sigma_{ij1n} \\
&+ \sum_{i=1}^{N_2} \sum_{j=1}^{N_3} \omega_2 \omega_3 \sigma_{ij23} + \cdots + \sum_{i=1}^{N_2} \sum_{j=1}^{N_n} \omega_2 \omega_n \sigma_{ij2n} \\
&\vdots \\
&+ \sum_{i=1}^{N_{n-1}} \sum_{j=1}^{N_n} \omega_{n-1} \omega_n \sigma_{ij(n-1)n}
\end{aligned} \tag{31}$$

which is equivalent to

$$\sigma_p^2 = \sum_{c=1}^{n} \omega_c^2 N_c \bar{\sigma}_c^2 + \sum_{c=1}^{n} \omega_c^2 N_c (N_c - 1) \bar{\sigma}_{ijc} + \sum_{c=1}^{n-1} \sum_{k=c+1}^{n} 2 \omega_c \omega_k N_c N_k \bar{\sigma}_{ck} \tag{32}$$

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

$$\sigma_p^2 = \sum_{c=1}^{n} \omega_c^2 \left[N_c \bar{\sigma}_c^2 + N_c (N_c - 1) \bar{\sigma}_{ijc} \right] + \sum_{c=1}^{n-1} \sum_{k=c+1}^{n} 2 \omega_c \omega_k \left[N_c N_k \bar{\sigma}_{ck} \right] \tag{33}$$

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