Fine structure of spectral properties for random correlation matrices: An application to financial markets

Giacomo Livan,[1,2,*] Simone Alfarano,[3,†] and Enrico Scalas[4,5,‡]

[1]*Dipartimento di Fisica Nucleare e Teorica, Università degli Studi di Pavia, Via Bassi 6, I-27100 Pavia, Italy*
[2]*Istituto Nazionale di Fisica Nucleare, Sezione di Pavia, Via Bassi 6, I-27100 Pavia, Italy*
[3]*Departament d'Economia, Universitat Jaume I, Campus del Riu Sec, E-12071 Castellón, Spain*
[4]*Dipartimento di Scienze e Tecnologie Avanzate, Laboratorio sui Sistemi Complessi, Università del Piemonte Orientale "Amedeo Avogadro", Viale T. Michel 11, I-15121 Alessandria, Italy*
[5]*BCAM - Basque Center for Applied Mathematics, Bizkaia Technology Park, Building 500, E-48160 Derio, Spain*

(Received 1 March 2011; revised manuscript received 3 May 2011; published 29 July 2011)

We study some properties of eigenvalue spectra of financial correlation matrices. In particular, we investigate the nature of the large eigenvalue bulks which are observed empirically, and which have often been regarded as a consequence of the supposedly large amount of noise contained in financial data. We challenge this common knowledge by acting on the empirical correlation matrices of two data sets with a filtering procedure which highlights some of the cluster structure they contain, and we analyze the consequences of such filtering on eigenvalue spectra. We show that empirically observed eigenvalue bulks emerge as superpositions of smaller structures, which in turn emerge as a consequence of cross correlations between stocks. We interpret and corroborate these findings in terms of factor models, and we compare empirical spectra to those predicted by random matrix theory for such models.

DOI: 10.1103/PhysRevE.84.016113 PACS number(s): 02.50.Ng, 05.10.−a, 07.05.Tp, 89.65.Gh

I. INTRODUCTION

In physics, random matrix theory (RMT) is mainly used to model systems of particles interacting according to unknown laws. This is particularly handy for studying energy levels of complex systems such as heavy nuclei and mesoscopic systems. In such cases, the Hamiltonian operator can be conveniently described by a random matrix featuring some suitable symmetry properties. In particular, two matrix ensembles have been commonly used: the Gaussian orthogonal ensemble of real symmetric random matrices, and the Gaussian unitary ensemble of Hermitian random matrices [1,2]. In both these cases, for proper normalization of matrix elements, the asymptotic statistical properties of the eigenvalues follow the so-called semicircle law:

$$\rho(\lambda) = \frac{1}{2\pi}\sqrt{4-\lambda^2}, \tag{1}$$

where $\rho(\lambda)$ is the marginal probability density function of the eigenvalues. Until recent years physicists often neglected the study of random correlation matrices, even though they find applications in very diverse fields ranging from biology to econometrics. For this reason, applied mathematicians have studied such objects since the 1920s [3]. The asymptotic eigenvalue statistics in this case is given by the Marčenko-Pastur distribution [4], which will be extensively discussed in the following sections. Since the late 1990s, thanks to the growing interest in financial markets as prototypes of complex systems, physicists started working on random correlation matrices [5,6], and this will be the subject of this paper as well.

We consider a set of N stocks whose spot price at time t we denote as $S_i(t), i = 1,\dots,N$. Let $t_1,\dots,t_{T+1}$ be $T+1$ equally spaced time instants, then we introduce the corresponding log returns

$$r_{i,j+1} \overset{\text{def}}{=} \log\frac{S_i(t_{j+1})}{S_i(t_j)}; \tag{2}$$

typically, one can think of the t_i as days. This notation is a little redundant, and we can simply denote time steps as $j = 1,\dots,T+1$. Now, we can assume that the T recorded log-return values are realizations of $N \times T$ random variables $R_j^{(i)}$, so that we globally end up with NT observations r_{ij}, $i = 1,\dots,N$, $j = 1,\dots,T$. Equivalently, the vector

$$\mathbf{r}_i \overset{\text{def}}{=} (r_{i,1},\dots,r_{i,T}) \tag{3}$$

containing all the observations of the ith asset returns can be seen as a realization of a vector random variable $\mathbf{R}^{(i)}$. Such a framework is fully characterized by finite probability distributions [7,8]:

$$\begin{aligned}\mathbb{P}\big(&R_1^{(1)} \in A_1^{(1)},\dots,R_T^{(1)} \in A_T^{(1)};\dots;\\ &R_1^{(N)} \in A_1^{(N)},\dots,R_T^{(N)} \in A_T^{(N)}\big)\\ &= \mathbb{P}(\mathbf{R}^{(1)} \in B^{(1)};\dots;\mathbf{R}^{(N)} \in B^{(N)}),\end{aligned} \tag{4}$$

where $A_j^{(i)} \in \mathbb{R}\ \forall i,j$ and $B^{(i)} \in \mathbb{R}^T\ \forall i$. Depending on the choice of the random variables $\mathbf{R}^{(i)}$, such a picture allows for a huge variety of possible descriptions for the stochastic dynamics of financial data. Most simply, a standard assumption, according to which the log returns are described by uncorrelated Gaussian processes $[(r_{1,i},\dots,r_{N,i}) \sim \mathcal{N}(0,\mathbf{1}_N)$, where $\mathbf{1}_N$ represents the $N \times N$ identity matrix] could be adopted. However, as is well known [9], correlations often play a major role, and a realistic description of financial markets should by no means neglect them. Still, a Gaussian framework can be retained by observing that a set of zero-mean correlated

[*] giacomo.livan@pv.infn.it
[†] alfarano@eco.uji.es
[‡] enrico.scalas@mfn.unipmn.it; URL: www.mfn.unipmn.it/∼scalas

Gaussian numbers generated by a stationary stochastic process is completely characterized by its expectation value vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\mathcal{E}}$:

$$\mathcal{E}_{ij,kl} = \mathbb{E}[r_{ij} r_{kl}]. \tag{5}$$

Following [10] in this paper we shall simplify this structure to the assumption that cross correlations between assets and autocorrelations in time factorize:

$$\mathcal{E}_{ij,kl} = C_{ik} A_{jl}. \tag{6}$$

In the above equation C_{ik} (A_{jl}) represents an element of a $N \times N$ ($T \times T$) positive-definite symmetric matrix $\mathbf{C}$ ($\mathbf{A}$). We shall keep this same kind of notation, that is, denoting matrices by bold letters and the corresponding matrix elements by the same nonbold letters throughout the rest of the paper. We shall assume the $\mathbf{C}$ matrix of cross correlations to be constant over time. Also, most importantly, we shall neglect all possible correlations in time by assuming $\mathbf{A} = \mathbf{1}_T$:

$$\mathcal{E}_{ij,kl} = C_{ik} \delta_{jl}. \tag{7}$$

This last assumption is well motivated both from an empirical viewpoint [9] and a theoretical one, since asset returns can be shown not to display autocorrelations whenever assets are assumed to be described by a submartingale. As a matter of fact, from the submartingale property one can show that

$$\mathbb{E}[r_{i,j} r_{i,k}] \sim \mathbb{E}[\widetilde{r}_{i,j} \widetilde{r}_{i,k}] = 0, \tag{8}$$

where

$$\widetilde{r}_{i,j} = \frac{S_i(t_{j+1}) - S_i(t_j)}{S_i(t_j)} \sim r_{i,j} \tag{9}$$

assuming no dividend payment in the period and $S_i(t_{j+1}) - S_i(t_j) \ll S_i(t_j)$. This relation means that returns are uncorrelated (not necessarily independent) random variables, as can be empirically verified [9]. In the following, we shall always assume the previously mentioned condition $[S_i(t_{j+1}) - S_i(t_j) \ll S_i(t_j)]$ to be fulfilled, thus allowing to identify log returns and returns [as in Eq. (9)].

The Gaussian probability measure leading to the correlation structure (7) can be shown to be

$$P(\mathbf{R})\mathrm{D}\mathbf{R} = \frac{1}{(2\pi)^{NT/2}(\det\mathbf{C})^{T/2}} \exp\left(-\frac{1}{2}\mathrm{Tr}\mathbf{R}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{R}\right)\mathrm{D}\mathbf{R}, \tag{10}$$

where $\mathbf{R}$ is a rectangular $N \times T$ matrix containing all of the returns observations ($R_{ij} = r_{ij}$), while $\mathrm{D}\mathbf{R} \overset{\text{def}}{=} \prod_{i=1}^{N} \prod_{j=1}^{T} \mathrm{d}R_{ij}$ is the flat integration measure over matrix elements.

Being symmetric, the $\mathbf{C}$ matrix in (10) is made of $N(N+1)/2$ independent entries. Now the typical challenge to be faced in many multivariate analysis problems is to estimate these numbers from N time series of T observations, that is, NT data points. When such data are collected in a $N \times T$ matrix $\mathbf{R}$ as in (10), then a standard estimator for $\mathbf{C}$ is given by the matrix $\mathbf{c} \overset{\text{def}}{=} \mathbf{R}\mathbf{R}^{\mathrm{T}}/T$. In other words, an estimator for the matrix element C_{ij} in $\mathbf{C}$ is given by

$$c_{ij} = \frac{1}{T} \sum_{t=1}^{T} R_{it} R_{jt}, \tag{11}$$

which is the well-known Pearson estimator for large T [for small values of T the $1/T$ factor would need to be replaced by $1/(T-1)$]. Of course, the c_{ij} are a noise-dressed representation of the C_{ij}. As a matter of fact, even though the random variables in $\mathbf{R}$ were exactly described by the probability distribution in (10) (i.e., by the correlation matrix $\mathbf{C}$), the finiteness of the data sample under study would anyway cause the c_{ij} to deviate, on average, from their "true" counterparts C_{ij}. As is intuitively clear, the two will become closer as more observations are collected, that is, as $T \to \infty$, or equivalently as $q \to 0$, where q is the so called "rectangularity ratio"

$$q \overset{\text{def}}{=} \frac{N}{T}. \tag{12}$$

However, realistic situations in financial practice typically involve large numbers of variables and similarly large numbers of observations. Ideally, this is not far from a "thermodynamic limit" situation in which

$$N, T \to \infty\,, \quad \text{with} \quad \frac{N}{T} = q = \text{constant}. \tag{13}$$

Remarkably this is precisely the regime under which some powerful RMT results are valid [4]. In particular, this is the limit under which it is possible to make analytical statements about the relation between the eigenvalue spectra of the theoretical covariance matrix $\mathbf{C}$ and its estimator (11) [11–13]. We shall exploit those results in the following sections.

A very general class of models fulfilling (7) is the one of the so called factor models [14–18]. Such models aim at describing the time evolution of each asset in terms of a few "driving forces," or factors, which typically describe the impact that a market sector or the whole market itself have on a given asset. In a K factors model, the time evolution of asset returns is given by

$$r_{it} = r_i(t) = \sum_{j=1}^{K} g_i^{(j)} m_j(t) + g_i^{(0)} \epsilon_i(t), \tag{14}$$

where $g_i^{(j)}$ and $g_i^{(0)}$ are constant parameters, whereas m_j and ϵ_i are independent and identically distributed normal random variables. We shall assume these latter to be normalized as follows:

$$\begin{gathered} \mathbb{E}[m_i(t)] = \mathbb{E}[\epsilon_i(t)] = 0, \\ \mathbb{E}[m_i(t) m_j(t')] = \mathbb{E}[\epsilon_i(t)\epsilon_j(t')] = \delta_{ij}\delta_{tt'}, \\ \mathbb{E}[m_i(t)\epsilon_j(t')] = 0. \end{gathered} \tag{15}$$

In the next section we shall specialize the model in (14) to a particular case. However, in a very general fashion, factor models have proven to be able to reproduce, at least qualitatively, some relevant features of empirical covariance matrix eigenvalue spectra.

The general appearance of the return covariance matrix eigenvalue spectrum of a given number of assets (for zero mean and unit standard deviation data) is the one depicted in Fig. 1 for the log returns of the daily prices for the assets composing the S&P500 and FTSE350 Indices (data downloaded from Yahoo Finance). Three main features are clearly visible: a large bulk close to zero, a number of larger eigenvalues "leaking out" of such bulk, and a much larger and isolated

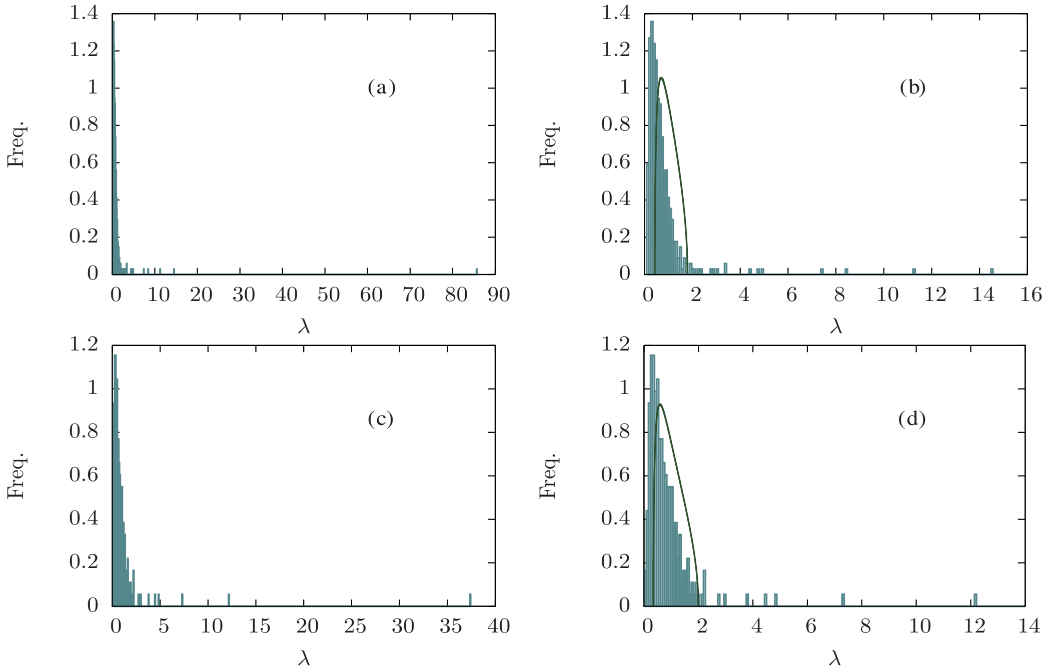

FIG. 1. (Color online) (a) Empirical eigenvalue density of the covariance matrix for $T = 3400$ daily returns of $N = 396$ assets belonging the S&P500 Index over the years 1996–2009. (b) The same density as in (a) without the largest eigenvalue. (c) Eigenvalue density for $T = 1423$ daily returns of $N = 243$ assets belonging to the FTSE350 Index over the years 2005–2010. (d) The same density as in (c) without the largest eigenvalue. All figures were produced with standardized data. In (b) and (d) the Marčenko-Pastur distributions for the corresponding values of $q = N/T$ are also plotted.

eigenvalue. Since the pioneering works [5,6], RMT has become a standard tool to analyze these macroscopic features. More specifically, the aforementioned eigenvalue bulk has mostly been identified with the Marčenko-Pastur distribution [4], that is, the limiting eigenvalue marginal probability density for the (already introduced) matrix $\mathbf{c} = \mathbf{R}\mathbf{R}^{\mathrm{T}}/T$ when all the entries R_{ij} are drawn from a normal distribution $\mathcal{N}(0,\sigma)$. Quite importantly, this result is rigorously derived only in the thermodynamic limit (13) of infinite matrix sizes growing to infinity at a fixed rate. In this limit, the Marčenko-Pastur distribution reads

$$\rho_{\mathbf{c}}(\Lambda) = \frac{1}{2\pi q\sigma^2}\frac{\sqrt{(\Lambda_+ - \Lambda)(\Lambda - \Lambda_-)}}{\Lambda}, \qquad \Lambda_\pm \overset{\text{def}}{=} \sigma^2(1 \pm \sqrt{q})^2, \tag{16}$$

where q is the rectangularity ratio defined in (12). However, as can be seen in Figs. 1(b)–1(d), the Marčenko-Pastur distribution actually provides a very poor fit of empirical distributions when q and σ are assumed to be equal to N/T and 1 (for standardized data), respectively. The aforementioned eigenvalue bulks are reasonably well fitted by a Marčenko-Pastur distribution only when q and σ are assumed to be *free* parameters, whose values are to be determined via fitting. In particular, this typically causes q to deviate from the ratio N/T, thus introducing the concept of effective system size.

Since the Marčenko-Pastur distribution emerges as the limiting density for the covariance matrix of N *uncorrelated* time series made of T observations, identifying eigenvalue bulks such as the ones in Fig. 1 with it basically amounts to state that most of the information contained in empirical covariance matrix spectra is actually no information at all, being equivalent to the spectrum one would obtain in the presence of pure noise [12,19]. On the other hand, this viewpoint allows one to give a specific meaning to the "large" eigenvalues out of the bulk. As would also be possible to verify with principal component analysis (PCA) [20], such eigenvalues correspond to groups of correlated assets, most typically belonging to the same market sector. Analogously, the largest eigenvalue of the distribution is usually identified with the "market mode": such an eigenvalue appears as a consequence of those fluctuations that involve the market as a whole, and as a matter of fact the PCA can easily show it to account for a large part of the return variance.

As already anticipated, factor models (14) represent good candidates to reproduce most of the empirical features shown in Fig. 1. In the following sections we shall make use of such models to challenge the previously mentioned common knowledge, according to which the eigenvalue bulks in empirical covariance matrix spectra essentially correspond to noise. Such a common knowledge has already been revised critically in a number of works (see, for example, [13,21–24]), and in this paper we wish to present an additional amount of evidence in this direction.

The paper is organized as follows. In Sec. II the "direct" problem of analytically estimating eigenvalue densities is addressed. In particular, some specific versions of the factor model in Eq. (14) will be introduced and the eigenvalue spectra for the correlation matrix $\mathbf{C}$ of such model will be derived (sometimes performing approximations). Then the

RMT results provided in [12,13] will be applied in order to derive exact results for the noise-dressed version **c** of the correlation matrix. Eventually, a subsection will be devoted to discuss the results obtained via Monte Carlo simulations in order to validate the analytical formulas. In the light of such numerical results, we shall also briefly discuss again the applicability limits of the Marčenko-Pastur distribution. In Sec. III the "inverse" problem of inferring eigenvalue densities from the empirically observed ones will be discussed. More specifically, a filtering procedure will be devised in order to highlight some of the cluster structure in empirical correlation matrices. Such a procedure will be performed on two data sets (relative to the S&P500 and the FTSE350 Indices): the results we obtain confirm, at least on a qualitative level, the ability of factor models to reproduce relevant stylized facts observed in real stock market returns, and we believe this to be one of the main points in our paper. Eventually, in Sec. IV some conclusions and possible future perspectives of this work will be outlined.

II. THEORY: THE DIRECT PROBLEM

A. Cluster models: Heuristic analysis

Let us now specialize the factor model (14). In particular, let us start from the situation where all asset returns obey the following equation:

$$r_i(t) = \gamma_N m_N(t) + (1-\gamma_N)\epsilon_i(t), \tag{17}$$

where $m_N(t),\epsilon_i(t) \sim \mathcal{N}(0,1)\ \forall t$, and $\gamma_N \in [0,1]$. In the previous equation m_N represents a common mode driving all assets with the same "intensity" γ_N. We shall now build K clusters of correlated assets from Eq. (17). Thus, let there be K groups of N_k variables ($k = 1,\ldots,K$) with $\bar{N} \overset{\text{def}}{=} \sum_{k=1}^{K} N_k \leqslant N$, and let us order the assets so that r_i belongs to the kth assets for $i = 1 + \sum_{l=1}^{k-1} N_l,\ldots,\sum_{l=1}^{k} N_l$. We can also denote the generic element in the kth cluster as $r_i^{(k)}$. We shall define it as

$$\begin{aligned} r_i^{(k)}(t) &= \gamma_k m_k(t) + (1-\gamma_k) r_i(t)\ , \\ i &= 1 + \sum_{l=1}^{k-1} N_l,\ldots,\sum_{l=1}^{k} N_l, \end{aligned} \tag{18}$$

where $\gamma_k \in [0,1]$, $m_k(t) \sim \mathcal{N}(0,1)$ is a cluster mode and r_i is as in Eq. (17). Thus, we can rewrite the previous relation as

$$\begin{aligned} r_i^{(k)}(t) &= \gamma_k m_k(t) + (1-\gamma_k)\gamma_N m_N(t) \\ &\quad + (1-\gamma_k)(1-\gamma_N)\epsilon_i(t)\ , \\ i &= 1 + \sum_{l=1}^{k-1} N_l,\ldots,\sum_{l=1}^{k} N_l. \end{aligned} \tag{19}$$

We still simply call r_i ($i = 1 + \bar{N},\ldots,N$) those elements which do not belong to any cluster, and we assume them to evolve according to (17). We always have $\mathbb{E}[r_i(t)] = \mathbb{E}[r_j^{(k)}(t)] = 0\ \forall,i,j,k,t$. Recalling the relations in (15), which can be generalized to include m_N in a straightforward way, we can calculate all possible covariance matrix elements between assets described by (17) and (19). Four separate cases can be distinguished:

$$\begin{aligned} \mathbb{E}[r_i(t)r_j(t)] &= (1-\gamma_N)^2\delta_{ij} + \gamma_N^2, \\ \mathbb{E}\big[r_i(t)r_j^{(k)}(t)\big] &= (1-\gamma_k)\gamma_N^2, \\ \mathbb{E}\big[r_i^{(k)}(t)r_j^{(k)}(t)\big] &= (1-\gamma_k)^2(1-\gamma_N)^2\delta_{ij} + (1-\gamma_k)^2\gamma_N^2 + \gamma_k^2, \\ \mathbb{E}\big[r_i^{(k)}(t)r_j^{(l)}(t)\big] &= (1-\gamma_k)(1-\gamma_l)(1-\gamma_N)^2\delta_{ij} \\ &\quad + (1-\gamma_k)(1-\gamma_l)\gamma_N^2. \end{aligned} \tag{20}$$

We are now in position to compute the correlation matrix **C** of the model, whose matrix elements read

$$C_{ij} = \frac{\mathbb{E}[r_i(t)r_j(t)]}{\sqrt{\mathrm{Var}[r_i(t)]\mathrm{Var}[r_j(t)]}} \tag{21}$$

with straightforward generalizations to those involving elements belonging to clusters.

We shall focus for now on the limiting case in which correlations between cluster elements are very strong, that is, when $\gamma_k \to 1$ in each cluster. One can see from (20) that under this assumption the model's correlation matrix has a simple block-diagonal structure:

$$\mathbf{C} = \begin{pmatrix} \mathbf{E}^{(N_1)} & 0 & \ldots & 0 & 0 \\ 0 & \mathbf{E}^{(N_2)} & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{E}^{(N_K)} & 0 \\ 0 & 0 & \ldots & 0 & \mathbf{F}^{(N-\bar{N})} \end{pmatrix}, \tag{22}$$

where $\mathbf{E}^M$ is the $M \times M$ matrix whose entries are all equal to unity ($E_{ij}^M = 1\ \forall i,j$), while $\mathbf{F}^{(N-\bar{N})}$ is a $(N-\bar{N}) \times (N-\bar{N})$ matrix with a slightly more complicated structure:

$$\mathbf{F}^{(N-\bar{N})} = \begin{pmatrix} 1 & \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} & \ldots & \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} \\ \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} & 1 & \ldots & \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} & \frac{\gamma_N^2}{(1-\gamma_N)^2+\gamma_N^2} & \ldots & 1 \end{pmatrix}. \tag{23}$$

The block structure in (22) allows for the computation of the eigenvalue spectrum. In fact, since we have

$$\begin{aligned} \det(\mathbf{E}^{(M)} - \Lambda \mathbf{I}_M) &= \Lambda^{M-1}(M-\Lambda), \\ \det(\mathbf{F}^{(N-\bar{N})} - \Lambda \mathbf{I}_{N-\bar{N}}) &= \left[\frac{(N-\bar{N})\gamma_N^2 + (1-\gamma_N)^2}{(1-\gamma_N)^2+\gamma_N^2} - \Lambda\right] \\ &\quad \times \left[\frac{(1-\gamma_N)^2}{(1-\gamma_N)^2+\gamma_N^2} - \Lambda\right]^{N-\bar{N}-1}, \end{aligned} \tag{24}$$

the characteristic equation for the **C** matrix reads

$$\Lambda^{\bar{N}-K}\left[\Lambda - \frac{(N-\bar{N})\gamma_N^2 + (1-\gamma_N)^2}{(1-\gamma_N)^2+\gamma_N^2}\right] \times \left[\Lambda - \frac{(1-\gamma_N)^2}{(1-\gamma_N)^2+\gamma_N^2}\right]^{N-\bar{N}-1} \prod_{k=1}^{K}(\Lambda - N_k) = 0. \tag{25}$$

This eigenvalue spectrum is able to reproduce, at least on a heuristic level, some of the features of empirical spectra (see Fig. 1): each cluster gives rise to a large eigenvalue equal to the cluster size N_k, and the common mode produces one large eigenvalue $\sim(N-\bar{N})\gamma_N^2/[(1-\gamma_N)^2+\gamma_N^2]$ too. It is worth mentioning that this latter eigenvalue might not necessarily be the largest one: as a matter of fact, large enough N_k and a small γ_N can lead to situations in which the largest eigenvalue is given by $\max_k N_k$. Even though this seems not to be the case in most financial applications, it is still worth stressing that the largest eigenvalue in empirical spectra should not be labeled as the "market eigenvalue" right away, but only after some further checks.

Going back to Eq. (25), a $(N-\bar{N}-1)$-fold degenerate eigenvalue $\{$equal to $(1-\gamma_N)^2/[(1-\gamma_N)^2+\gamma_N^2]\}$ can be recognized. Also, Eq. (25) indicates that each cluster gives rise to N_k-1 zero modes, altogether forming a group of $\bar{N}-K$ zero modes. In a noise-marred situation, as it can be verified by means of Monte Carlo simulations, the degeneracies in (25) are broken and give rise to two bulks. In the highly correlated cluster assumption ($\gamma_k\to 1$) yielding (25) the two bulks typically remain well separated. However, when such assumption is relaxed, allowing for small values of the γ_k, the two bulks get closer, and for properly chosen values of the parameters they eventually "collide" and merge into one single structure (see Sec. II C and the figures in it). This latter might be identified with the typical eigenvalue bulks appearing in empirical spectra (see Fig. 1). It is important to stress, already at this heuristic level, that the emergence of such a bulk in this factor model stems from the presence of (weak) correlations between the assets, oppositely to the Marčenko-Pastur distribution (16), which in turn, as already discussed, originates from pure noise. Nevertheless, quite subtly the Marčenko-Pastur distribution can still provide good fits to such bulks in a number of situations, as we shall illustrate later.

The previous factor model, yielding Eq. (25) for the eigenvalue spectrum of its correlation matrix, can be further simplified to the case where no common factor is driving the asset returns. This can be directly achieved on the eigenvalue spectrum by setting $\gamma_N=0$ in Eq. (25). This gives

$$\Lambda^{\bar{N}-K}(\Lambda-1)^{N-\bar{N}}\prod_{k=1}^{K}(\Lambda-N_k)=0. \tag{26}$$

This spectrum still yields one large eigenvalue for each cluster and two degenerate eigenvalues equal to zero and one, respectively. Just like in the previous discussion, let us now relax the assumption of strong correlations ($\gamma_k\to 1$) within clusters. For the sake of simplicity, let us assume that all assets in each cluster are mutually correlated with the same correlation coefficient $\rho_k\in[0,1]$ [which can be explicitly computed from (20)]. So, the correlation matrix of the model would read

$$\mathbf{C}=\begin{pmatrix}\tilde{\mathbf{E}}^{(N_1)} & 0 & \dots & 0 & 0\\ 0 & \tilde{\mathbf{E}}^{(N_2)} & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & \tilde{\mathbf{E}}^{(N_K)} & 0\\ 0 & 0 & \dots & 0 & \mathbf{I}_{N-\bar{N}}\end{pmatrix}, \tag{27}$$

where each $\tilde{\mathbf{E}}^{(N_k)}$ is a $N_k\times N_k$ matrix such that $\tilde{E}_{ij}^{(N_k)}=\rho_k$ for $i\neq j$ and $\tilde{E}_{ii}^{(N_k)}=1$, while the last block is now given by the identity matrix [as can be seen from the first relation in (20) for $\gamma_N=0$]. One can verify that

$$\det(\tilde{\mathbf{E}}^{(N_k)}-\Lambda\mathbf{I}_{N_k})=[\Lambda-(1-\rho_k)]^{N_k-1}\times\{\Lambda-[N_k\rho_k+(1-\rho_k)]\}. \tag{28}$$

In order to further simplify things, let us consider the case where we have just one cluster of $\bar{N}$ assets with mutual correlation ρ. Equation (25) in this case would need to be modified to read

$$[\Lambda-(1-\rho)]^{\bar{N}-1}(\Lambda-1)^{N-\bar{N}}\{\Lambda-[\bar{N}\rho+(1-\rho)]\}=0, \tag{29}$$

thus giving one large eigenvalue and two degenerate eigenvalues equal to $(1-\rho)$ and one. The latter emerges as a consequence of the $N-\bar{N}$ mutually uncorrelated assets, that is, as a consequence of pure noise, while the former is due to the presence of a cluster. Just like in the case discussed previously, a noise-dressed version of (29) would lead to two eigenvalue bulks, and suitably chosen values of ρ would make the two bulks merge into one (see Sec. II C). Thus, in this case too, the emergence of a main bulk would not be a consequence of pure noise alone.

B. Cluster models: Exact results

The proper mathematical framework to deal with covariance matrices featuring degenerate spectra [as the ones in Eqs. (25), (26), and (29)] is the one provided in [12,13], and we shall exploit it extensively in the following. So, first let us introduce some basic notions and notations of RMT. Just like we did so far, we shall denote the eigenvalues of the correlation matrix $\mathbf{C}$ of a given model as Λ_i $(i=1,\dots,N)$, while the eigenvalues of the corresponding estimator (11) will be denoted as λ_i. Quite straightforwardly, one can define the eigenvalue density for the theoretical correlation matrix as

$$\rho_{\mathbf{C}}(\Lambda)=\frac{1}{N}\sum_{i=1}^{N}\delta(\Lambda-\Lambda_i), \tag{30}$$

and this is related to the matrix moments $M_{\mathbf{C}}^{(k)}$:

$$M_{\mathbf{C}}^{(k)}\overset{\text{def}}{=}\frac{1}{N}\mathrm{Tr}\,\mathbf{C}^k=\frac{1}{N}\sum_{i=1}^{N}\Lambda_i^k=\int d\Lambda\rho_{\mathbf{C}}(\Lambda)\Lambda^k. \tag{31}$$

In analogy to (30), one can define an expected spectral density for the estimator $\mathbf{c}$ in equation (11):

$$\rho_{\mathbf{c}}(\lambda)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\delta(\lambda-\lambda_i)], \tag{32}$$

where the expectation is to be meant with respect to the probability measure (10). Generalizing (31), we can then define the expected matrix moments as

$$m_{\mathbf{c}}^{(k)}\overset{\text{def}}{=}\frac{1}{N}\mathbb{E}[\mathrm{Tr}\mathbf{c}^k]=\int d\lambda\rho_{\mathbf{c}}(\lambda)\lambda^k. \tag{33}$$

The two corresponding resolvents, or Green's functions, are given by

$$\begin{aligned}\mathbf{G}_{\mathbf{C}}(Z) &= (Z\mathbf{I}_N - \mathbf{C})^{-1},\\ \mathbf{g}_{\mathbf{c}}(z) &= \mathbb{E}[(z\mathbf{I}_N - \mathbf{c})^{-1}],\end{aligned} \tag{34}$$

where $Z,z \in \mathbb{C}$. Then one can introduce the moment generating functions, and it is possible to show that they are closely related to the Green's functions in the following way:

$$\begin{aligned}M_{\mathbf{C}}(Z) &\overset{\text{def}}{=} \sum_{k=1}^{\infty} \frac{M_{\mathbf{C}}^{(k)}}{Z^k} = Z G_{\mathbf{C}}(Z) - 1,\\ m_{\mathbf{c}}(z) &\overset{\text{def}}{=} \sum_{k=1}^{\infty} \frac{m_{\mathbf{c}}^{(k)}}{z^k} = z g_{\mathbf{c}}(z) - 1,\end{aligned} \tag{35}$$

where we have

$$\begin{aligned}G_{\mathbf{C}}(Z) &\overset{\text{def}}{=} \frac{\mathrm{Tr}[\mathbf{G}_{\mathbf{C}}(Z)]}{N},\\ g_{\mathbf{c}}(z) &\overset{\text{def}}{=} \frac{\mathrm{Tr}[\mathbf{g}_{\mathbf{c}}(z)]}{N}.\end{aligned} \tag{36}$$

Moreover, from the well known relation $\lim_{\epsilon\to 0^+}(\lambda + \mathrm{i}\epsilon)^{-1} = \mathcal{P}(\lambda^{-1}) - \mathrm{i}\pi\delta(\lambda)$ (where $\mathcal{P}$ denotes the principal value), one can show that the eigenvalue densities (30) and (32) can be directly derived from the corresponding Green's functions (34):

$$\begin{aligned}\rho_{\mathbf{C}}(\Lambda) &= -\frac{1}{\pi} \lim_{\epsilon\to 0^+} \mathrm{Im}\, G_{\mathbf{C}}(\Lambda + \mathrm{i}\epsilon),\\ \rho_{\mathbf{c}}(\lambda) &= -\frac{1}{\pi} \lim_{\epsilon\to 0^+} \mathrm{Im}\, g_{\mathbf{c}}(\lambda + \mathrm{i}\epsilon).\end{aligned} \tag{37}$$

So, basically, the Green's function contains the same information as the whole eigenvalue density, and, through (35), the same is also true for the moment generating function. In particular, for $\Lambda,\lambda > 0$, the previous relations can be converted into

$$\begin{aligned}\rho_{\mathbf{C}}(\Lambda) &= -\frac{1}{\pi\Lambda} \lim_{\epsilon\to 0^+} \mathrm{Im}\, M_{\mathbf{C}}(\Lambda + \mathrm{i}\epsilon),\\ \rho_{\mathbf{c}}(\lambda) &= -\frac{1}{\pi\lambda} \lim_{\epsilon\to 0^+} \mathrm{Im}\, m_{\mathbf{c}}(\lambda + \mathrm{i}\epsilon).\end{aligned} \tag{38}$$

A fundamental relation between between the moment generating functions of a "true" correlation matrix and its estimator in the infinite matrix size limit (13) can be derived [12,13] either in the framework of free random variables [10,25,26] or using planar diagrammatic methods [12,27,28]. The starting point is the following simple relation between moment generating functions:

$$m_{\mathbf{c}}(z) = M_{\mathbf{C}}(Z), \tag{39}$$

where the two complex arguments are related by the following transformation:

$$Z = \frac{z}{1 + q m_{\mathbf{c}}(z)}. \tag{40}$$

Once $M_{\mathbf{C}}(Z)$ is known, $m_{\mathbf{c}}(z)$ can be derived in principle from the following functional equation:

$$m_{\mathbf{c}}(z) = M_{\mathbf{C}}\left[\frac{z}{1 + q m_{\mathbf{c}}(z)}\right]. \tag{41}$$

Bearing in mind the previous discussion on factor models, we shall focus on correlation matrices whose spectra display degenerate eigenvalues. Let us then assume the correlation matrix $\mathbf{C}$ to have L distinct eigenvalues Λ_i $(i = 1,\ldots,L)$ with degeneracies n_i. The moment generating function for such a matrix is given by

$$M_{\mathbf{C}}(Z) = \frac{1}{N}\sum_{i=1}^{L} \frac{n_i\Lambda_i}{Z - \Lambda_i} = \sum_{i=1}^{L} \frac{w_i\Lambda_i}{Z - \Lambda_i}, \tag{42}$$

where the weights $w_i = n_i/N$ have been introduced. Thus, from (41) we get

$$m_{\mathbf{c}}(z) = \sum_{i=1}^{L} \frac{w_i\Lambda_i[1 + q m_{\mathbf{c}}(z)]}{z - \Lambda_i[1 + q m_{\mathbf{c}}(z)]}. \tag{43}$$

For each fixed z this becomes a polynomial equation of degree $L + 1$ in $m_{\mathbf{c}}(z)$, yielding as many solutions. The problem arises of choosing the right one: as extensively discussed and detailed in [29], the right branch of the map in Eq. (40) to pick up is the one giving $Z \to z$ for $z \to \infty$. In the simplest case one has $\mathbf{C} = \mathbf{I}_N$ and, of course, the correlation matrix has just one N-fold degenerate eigenvalue equal to one: it can be shown that, in this case, Eq. (43) leads precisely to the Marčenko-Pastur distribution (16), as one would expect. On the other hand, already when considering two distinct eigenvalues, quite different scenarios are possible, including the previously discussed cases of well separated or merging bulks (see the next subsection and [28]). Let us also remark that Eq. (43) cannot be applied to the large nondegenerate eigenvalues typically displayed by factor models. This is because, as already stated, we shall always work in the thermodynamic limit (13), where the weight $(1/N)$ of such eigenvalues vanishes as $N \to \infty$. As a matter of fact, this kind of eigenvalues need to be investigated *per se*, and actually extensive areas of the RMT literature are devoted to the study of statistical properties of single eigenvalues as well as order statistics [30]. In particular, it has been shown in [31] that large nondegenerate sample eigenvalues of correlation matrices follow a normal distribution (see the next subsection for a numerical confirmation).

C. Monte Carlo simulations

In this subsection we present and detail the Monte Carlo simulations we performed in order to test and validate the analytical results described so far. In all cases, we generated T realizations of N stochastic processes described by the factor model introduced in Eqs. (18), (19), and (20) (from a numerical viewpoint, this just boils down to the generation of standard Gaussian random numbers). By choosing different parameter values, we implemented the different versions of the model which were discussed in the previous subsection, corresponding to different theoretical correlation matrices Eqs. (22) and (27)]. The eigenvalues of the corresponding estimators (11) were obtained via numerical diagonalization (by means of the diagonalization algorithm provided by Matlab®).

In Fig. 2 a first example of eigenvalue spectra deriving from factor models is presented. In this first example a common mode (introduced via a nonzero γ_N coefficient, see the figure caption for all the details on parameter values) as

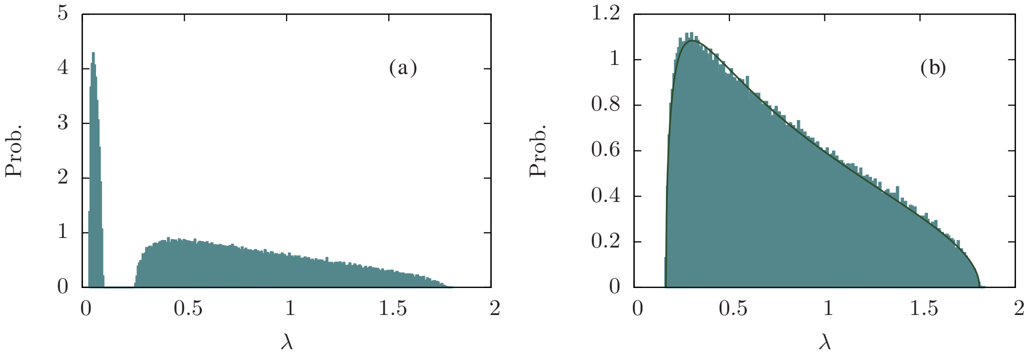

FIG. 2. (Color online) (a) Eigenvalue density for 100 simulations of the factor model described in (18), (19), and (20) with $N = 500$, $T = 2000$ ($q = 0.25$). One cluster made of $N_1 = \bar{N} = 100$ variables, correlated via a coefficient $\gamma_1 = 0.7$, is present. A common mode is introduced via a coefficient $\gamma_N = 0.3$. As expected from Eq. (25), two eigenvalue bulks are clearly visible. The mean value in the bulk on the left is 0.04, while Eq. (25) would predict zero as a consequence of the strong correlation limit ($\gamma_k \to 1$) approximation. On the other hand, the mean value in the bulk on the right is 0.85, in remarkable agreement with the predicted value $(1-\gamma_N)^2/[(1-\gamma_N)^2+\gamma_N^2] = 0.84$. For the sake of readability, the "large" eigenvalues are not shown. (b) By setting $\gamma_1 = 0.4$, the two bulks in (a) merge into a single one. Such a structure, despite emerging as a consequence of (weak) correlations, is very well fitted by a Marčenko-Pastur distribution [see Eq. (16)] with $q = 0.29$ and $\sigma = 0.88$ (solid line).

well as a correlated cluster of variables are present. As already discussed in the previous subsection, the degeneracies in Eq. (25) are broken, and, in the limit of strong correlations in the cluster ($\gamma_k \to 1$), two well separated eigenvalue bulks emerge [Fig. 2(a)]. On the other hand, when such correlations get weaker, the two eigenvalue bulks get closer, eventually melting into one single structure (Fig. 2). Remarkably such a structure is quite well fitted by a Marčenko-Pastur distribution, which is however characterized by values of the q and σ parameters that differ from the ones which would be obtained for standardized uncorrelated data ($q = N/T$ and $\sigma = 1$).

Such features are further illustrated in Fig. 3, which refers to the case of a factor model with no common mode ($\gamma_N = 0$). Again the progressive fusion (induced by weaker correlations) between separated eigenvalue bulks is shown. Also, we compare the numerically obtained spectra to the eigenvalue densities obtained from the solution of Eq. (43), obtaining a very good agreement between the two [Figs. 3(a) and 3(b)]. Just like in the previous case, the Marčenko-Pastur distribution seems to provide quite a good fit of the "limiting" eigenvalue bulk obtained for small correlations [Fig. 3(c)]. However, we performed a Kolmogorov-Smirnov [32] test under the null hypothesis of data distributed according to a Marčenko-Pastur distribution, and we found such hypothesis to be rejected for all the significance levels we considered (see the caption of Fig. 3 for further details). On the other hand, the same test prevented us from rejecting the hypothesis of data distributed according to the eigenvalue density obtained from the solution of Eq. (43), its degenerate eigenvalues being given by Eq. (29). This is quite surprising, given the great similarity between the two densities [see Fig. 3(d)], which would be almost undistinguishable if plotted on the scale of the whole distribution [as in Fig. 3(c)]. In the following section, we shall apply these ideas to financial data.

We also believe these findings to provide some interesting evidence against the use of the Marčenko-Pastur distribution whenever nonnegligible correlations are present between random variables. Despite being close, in a number of situations, to the eigenvalue densities deriving from the solution of Eq. (43), the Marčenko-Pastur distribution always needs to be fitted on the data under study, even when they are completely under control (as in the case of Monte Carlo simulations). Then, as already pointed out, the presence of correlations causes the parameters q and σ to deviate from the corresponding values which would be obtained in a pure noise situation. In particular, given the definition in Eq. (12), this leads to the introduction of the artificial, and possibly misleading, concept of effective system size. On the other hand, in the conclusions section we provide suggestions on how to use cluster models to fit empirical spectra just by means of filtering algorithms and Monte Carlo simulations.

Eventually, concluding this subsection on Monte Carlo simulations, in Fig. 4 we show a numerically obtained distribution of the largest sample eigenvalue for a factor model yielding the eigenvalue spectrum in (29). As can be seen by direct inspection, the corresponding histogram is well fitted by a Gaussian distribution, as already anticipated in the previous subsection. Moreover, three statistical tests (whose details are provided in the caption) were performed under the null hypothesis of normally distributed data, and all the results we obtained prevent from rejecting such hypothesis. It is known that in a number of situations largest sample eigenvalues of correlation matrices are distributed according to Tracy-Widom (TW) distributions [30]. Now, since in some cases TW distributions can look quite close to normal distributions, we also performed statistical tests in order to rule out the former for the distribution of the largest sample eigenvalue of factor models.

III. EMPIRICAL DATA: THE INVERSE PROBLEM

The goal of this section is to show that some of the features displayed in correlation matrix spectra of factor models are actually present in empirical spectra of financial correlation matrices too. In particular, our goal is to show that the Gaussian factor models outlined in the previous sections and

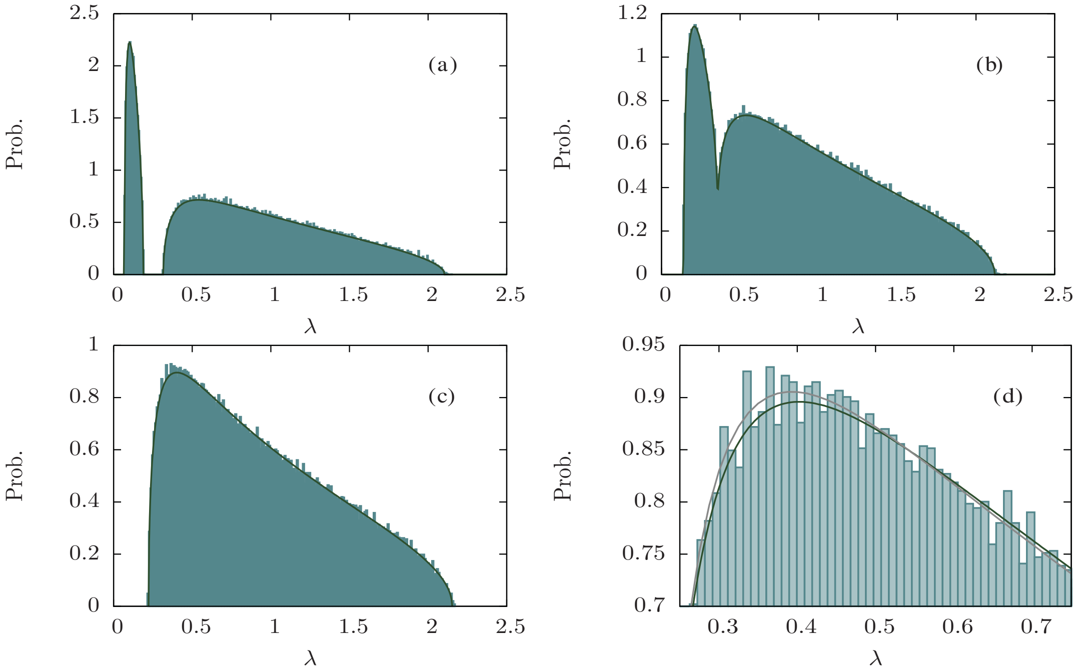

FIG. 3. (Color online) (a) Eigenvalue spectrum of the correlation matrix for the factor model yielding the spectrum in (29) for $N = 500$, $T = 2000$, $\bar{N} = 100$, and $\rho = 0.84$. This model yields two degenerate eigenvalues: $\Lambda_1 = 1 - \rho = 0.16$ and $\Lambda_2 = 1$ (see also Fig. 2). The histogram is the result of 100 Monte Carlo simulations of such model, while the solid line represents the density obtained from the solution of Eq. (43) (solved as in [29]). (b) Eigenvalue spectrum for the same model with $\rho = 0.65$, that is, for $\Lambda_1 = 0.35$. It can be clearly seen that the two separated bulks shown in (a) start to merge as a consequence of the smaller correlations (smaller value of ρ). Again, the solid line represents the density obtained from Eq. (43). (c) Posing $\rho = 0.30$ the two eigenvalue bulks merge completely into one single structure. In analogy to Fig. 2(b), such a structure is apparently well fitted by a Marčenko-Pastur distribution with $q = 0.26$ and $\sigma = 0.97$, plotted as a solid line. On this scale, the Marčenko distributions would be barely distinguishable from the density obtained from Eq. (43) with $\Lambda_1 = 1 - \rho = 0.7$ and $\Lambda_2 = 1$. (d) Comparison between two such densities (the dark line represents the Marčenko-Pastur distribution, while the grey line is the solution of Eq. (43)) in correspondence of their peak, where they differ the most. Despite the quite small deviation between the two, a Kolmogorov-Smirnov (KS) performed on the data gave the following results. The critical values for different significance levels α, are given by $\mathrm{V}^*_{\mathrm{KS}}(\alpha = 0.10) = 5.5 \times 10^{-3}$, $\mathrm{V}^*_{\mathrm{KS}}(\alpha = 0.05) = 6.1 \times 10^{-3}$ and $\mathrm{V}^*_{\mathrm{KS}}(\alpha = 0.01) = 7.3 \times 10^{-3}$. Under a null hypothesis of data distributed according to the Marčenko-Pastur distribution, the value of the KS statistic was $\mathrm{S}_{\mathrm{KS}} = 7.9 \times 10^{-3}$, allowing for the rejection of the null hypothesis for all the significance levels considered. On the other hand, under the null assumption of data distributed according to the density obtained from Eq. (43), we obtained $\mathrm{S}_{\mathrm{KS}} = 2.3 \times 10^{-3}$, thus preventing from rejecting the null hypothesis. Clearly the large statistics in this example plays a relevant role in helping the KS test to "distinguish" the two densities. Smaller data samples would prevent the Marčenko-Pastur distribution from being rejected. Indeed, when performing a KS test on a smaller version ($N = 250$, $T = 1000$, $\bar{N} = 50$, $\rho = 0.30$) of the system in (d), we obtained $\mathrm{S}_{\mathrm{KS}} = 3.0 \times 10^{-3}$ with the density given by Eq. (43) and $\mathrm{S}_{\mathrm{KS}} = 5.0 \times 10^{-3}$ for the Marčenko-Pastur distribution, the critical value being $\mathrm{V}^*_{\mathrm{KS}}(\alpha = 0.05) = 8.6 \times 10^{-3}$. Thus this example shows how rescaled and smaller system dimensions might prevent from discriminating the two densities.

the eigenvalue spectra they yield for correlation matrices do actually make contact with financial data on a qualitative level, thus confirming that the empirically observed eigenvalue bulks, as the ones shown in Fig. 1, cannot be regarded as a consequence of pure noise. In the light of the discussions in Sec. II, we consider a peak separation [similar to those shown in Figs. 2(a) and 3(a)] the most important phenomenological evidence one should look for. For this reason we shall try to empirically recreate, as best as possible, the conditions under which such a peak separation might be achieved. It is important to stress that nothing guarantees *a priori* that such conditions should be fulfilled by financial data. Thus, in the following we shall devise a filtering procedure able to detect those correlation structures (i.e., strongly correlated clusters plus additional uncorrelated assets) which might replicate the block-diagonal correlation matrices generated by the factor models introduced in the previous sections closely enough. We shall restrict our attention only to a relatively small number of assets in our data sets (396 assets from the S&P500 Index and 243 assets from the FTSE350 Index), namely only those ones which ideally replicate the block-diagonal structure of the correlation matrix in (27). When a single cluster is considered, such matrix yields the eigenvalue spectrum of Eq. (29), whose noise-dressed version is expected to produce well separated eigenvalue bulks. Let us rewrite the matrix in Eq. (27) for this specific case:

$$\mathbf{C} = \begin{pmatrix} \tilde{\mathbf{E}}^{(\bar{N})} & 0 \\ 0 & \mathbf{I}_{N-\bar{N}} \end{pmatrix}, \tag{44}$$

recalling that $\tilde{E}^{(\bar{N})}_{ij} = \rho$ for $i \neq j$ and $\tilde{E}^{(\bar{N})}_{ii} = 1$, $\bar{N}$ being the number of elements in the correlated cluster.

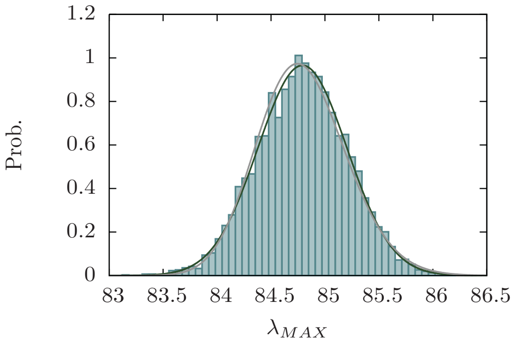

FIG. 4. (Color online) Distribution of the largest eigenvalue $\lambda_{\rm MAX}$ from 5000 Monte Carlo simulations of the factor model yielding the spectrum in (29) with $\rho = 0.85$, $N = 500$, and $\bar{N} = 100$. The distribution is well fitted by a normal distribution (dark line) with expected value $m = 84.79$, very close to the theoretical value predicted by Eq. (29): $\bar{N}\rho + (1-\rho) = 85.15$. Three different statistical tests (Jarque-Bera, Lilliefors, and Kolmogorov-Smirnov) [32] were performed, assuming a null hypothesis of normally distributed data. In the following we report the different critical values ($\rm V^*$) obtained for the different tests and for different significance levels α. Also, we report the statistic values (S), which, if smaller than the critical values, prevent the null hypothesis from being rejected. Jarque-Bera test: $\rm S_{JB} = 3.114$, $\rm V^*_{JB}(\alpha = 0.10) = 4.605$, $\rm V^*_{JB}(\alpha = 0.05) = 5.992$, $\rm V^*_{JB}(\alpha = 0.01) = 9.210$. Lilliefors test: $\rm S_L = 0.78 \times 10^{-2}$, $\rm V^*_L(\alpha = 0.10) = 1.14 \times 10^{-2}$, $\rm V^*_L(\alpha = 0.05) = 1.25 \times 10^{-2}$, and $\rm V^*_L(\alpha = 0.01) = 1.56 \times 10^{-2}$. Kolmogorov-Smirnov test: $\rm S_{KS} = 0.78 \times 10^{-2}$, $\rm V^*_{KS}(\alpha = 0.10) = 1.73 \times 10^{-2}$, $\rm V^*_{KS}(\alpha = 0.05) = 1.92 \times 10^{-2}$, $\rm V^*_{KS}(\alpha = 0.01) = 2.30 \times 10^{-2}$. All of the previous results prevent from rejecting the hypothesis of normally distributed data. We also generated a TW distribution with the algorithm described in [33], fitted it (grey line) to the data, and performed a KS test on it (the KS being the only non-Gaussian test among the ones mentioned previously) obtaining $\rm S_{KS} = 2.16 \times 10^{-2}$, which allows to reject the null hypothesis of data distributed according to a TW distribution for $\alpha = 0.05, 0.10$.

In order to reproduce the structure in (44), we start from the empirical covariance matrices (let us denote them as $\mathbf{c}$, according to the previously adopted notation) of our data sets and apply the following procedure.

(1) We first identify a small cluster of $\bar{N}$ strongly mutually correlated assets. If we denote the corresponding set of indices as I_U, then we have

$$c_{ij} \geqslant \rho_U, \quad i,j \in I_U \tag{45}$$

for some threshold value $\rho_U > 0$, which should be fixed to be quite high (a reasonable choice for financial data is $\rho_U \approx 0.6$) in order to actually ensure strong correlations between all the stocks in the cluster. This selection procedure is meant to reproduce the $\tilde{\mathbf{E}}^{(\bar{N})}$ diagonal block in Eq. (44).

(2) Then, all assets which are weakly correlated to the elements in the cluster are pointed out. Among those, only the ones with small mutual correlations are retained. By grouping their indices in another set I_D, we can write

$$\begin{aligned} |c_{ij}| &\leqslant \rho'_D, \quad i \in I_U,\ j \in I_D, \\ |c_{kl}| &\leqslant \rho''_D, \quad k,l \in I_D,\ k \neq l \end{aligned} \tag{46}$$

for some threshold values $\rho'_D, \rho''_D \in (0, \rho_U)$. ρ'_D and ρ''_D should be chosen to be quite small (reasonable choices are $\rho'_D, \rho''_D \approx 0.1$), so that the first condition in (46) will ensure very weak correlations between elements in I_D and elements in the cluster I_U, mimicking the zero off-diagonal blocks in (44). Similarly, for small values of ρ''_D the second condition in (44) will reproduce the identity matrix in the right-lower block of (44).

(3) If we now redefine N to be the total number of stocks in I_U and I_D, so that I_D contains $N - \bar{N}$ elements, and we properly sort them, then the approximation to (44) is given as follows

$$\mathbf{c} = \begin{pmatrix} \mathbf{c}_{I_U} & \mathbf{c}_{I_U,I_D} \\ \mathbf{c}_{I_D,I_U} & \mathbf{c}_{I_D} \end{pmatrix}, \tag{47}$$

where $\mathbf{c}_{I_U}$ and $\mathbf{c}_{I_D}$ are square matrices (of dimensions $\bar{N}$ and $N - \bar{N}$, respectively) containing the correlation matrix elements pertaining to the two sets I_U and I_D. On the other hand, the $\mathbf{c}_{I_U,I_D}$ matrix ($\mathbf{c}_{I_D,I_U}$ being its transpose) contains the "interaction" terms between the two sets.

The goal of such a construction is to empirically make contact with the spectrum in Eq. (29). As a matter of fact, for suitably chosen threshold values ρ_U, ρ'_D, and ρ''_D, we expect the eigenvalue spectrum of the $\mathbf{c}$ matrix in (47) to be the noise-dressed version of the one in (29). In particular, small values of ρ'_D and ρ''_D should guarantee the $\mathbf{c}_{I_D}$ block to yield $N - \bar{N}$ eigenvalues close to one. On the other hand, the block $\tilde{\mathbf{E}}^{(\bar{N})}$ in (44) yields $\bar{N} - 1$ small eigenvalues equal to $1 - \rho$ and a large one equal to $\bar{N}\rho + (1 - \rho)$. Now it is reasonable to assume ρ to be equal to the average mutual correlation between the assets (i.e., the average of off-diagonal correlation matrix elements) in I_U,

$$\rho = \frac{1}{\bar{N}(\bar{N}-1)} \sum_{i \neq j} [c_{I_U}]_{ij}, \tag{48}$$

and to suppose that $\mathbf{c}_{I_U}$ will produce $\bar{N} - 1$ eigenvalues close to this value. Lastly, before moving on let us mention that applying the previously outlined filtering procedure to uncorrelated or weakly correlated data would not produce any relevant result, even for less restrictive threshold values than the ones we use.

In the following we present and discuss the results we obtained applying this procedure to our data sets. In the case of the S&P500 Index, we identified a cluster made of $\bar{N}_{\rm S\&P} = 7$ strongly mutually correlated assets [$\rho_{\rm S\&P} = 0.712$, with $\rho_{\rm S\&P}$ computed as in (48)], all of which happen to belong to the energy sector. We then identified a group of 33 stocks, belonging to various sectors, which satisfy the previously described requirements: a small mutual correlation (mean value $= 0.099$) and a small correlation with the $\bar{N}$ elements in the cluster (mean value $= 0.096$). So, all in all we have $N_{\rm S\&P} = 40$. Analogously, also in the FTSE350 Index case we were able to identify a cluster made of $\bar{N}_{\rm FTSE} = 7$ highly mutually correlated stocks ($\rho_{\rm FTSE} = 0.707$), all corresponding to investment trusts. In this case, however, we only found 21 more stocks (so that $N_{\rm FTSE} = 28$) satisfying the aforementioned requirements (mean value of mutual correlation $= 0.015$, mean value of correlations with elements in the cluster $= 0.014$). In Fig. 5 graphical representations of the empirical correlation matrices were obtained, and a

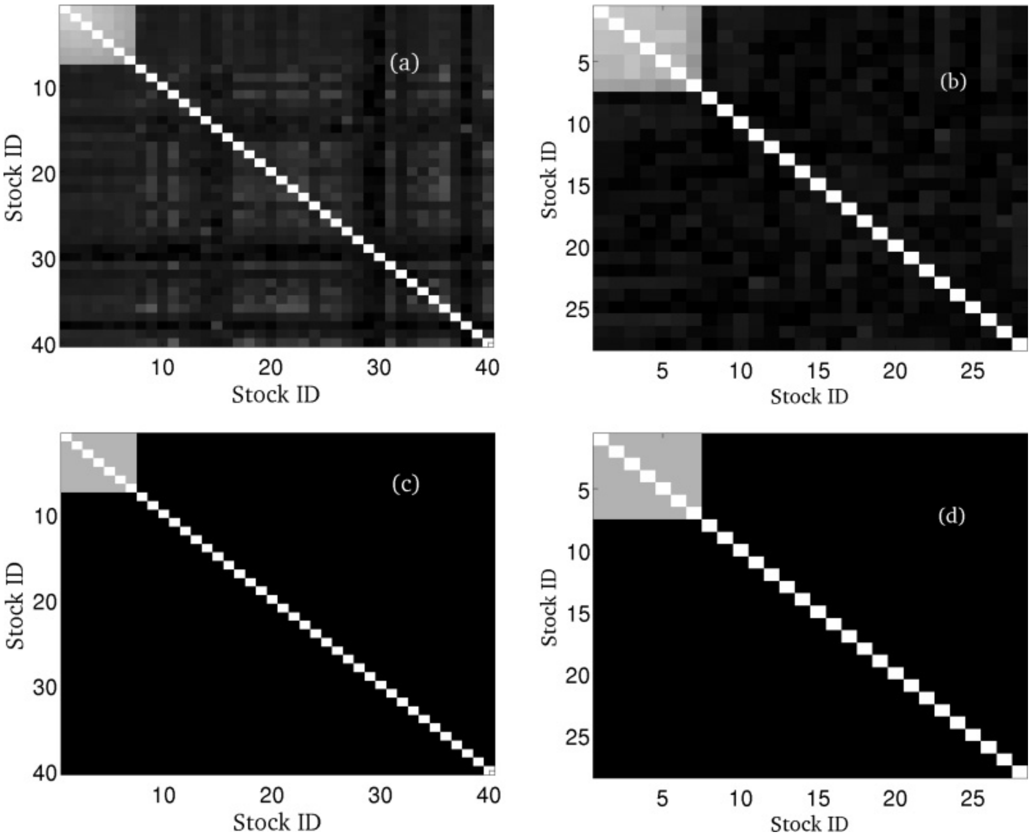

FIG. 5. Graphical representation of correlation matrices. (a) 40 × 40 correlation matrix for the selected returns belonging to the S&P500 Index. (b) 28 × 28 correlation matrix for the selected returns belonging to the FTSE350 Index. In (a) and (b) the stocks have been sorted in order to highlight the cluster structure. (c) and (d) Model correlation matrices corresponding to the cases shown in (a) and (b), respectively. In all plots, white diagonal blocks correspond to ones, while black ones correspond to zeros. Gray shadings are intermediate values. As already explained in the main text, the gray shadings in (c) and (d) correspond to $\rho = 0.712$ and $\rho = 0.707$, respectively. The presence of unresolved structures in (a) and, less evident, in (b), suggests that the model matrix in (44), depicted in (c) and (d), does not fully capture all empirical features.

comparison to the theoretically expected ones are shown. As can be seen by direct inspection, the correlation matrix we obtain in the FTSE350 Index case is remarkably similar to the one in (44), whereas the one we obtain for the S&P500 Index has some further inner structure as a consequence of the much higher mean correlations.

In Fig. 6 the eigenvalue spectra we obtained from the previously discussed correlation matrices [the ones reported in Figs. 5(a) and 5(b)] are shown. In particular, in Figs. 6(a) and 6(b) we plot the spectra obtained, respectively, from the S&P500 and FTSE350 correlation matrices constructed according to the clustering procedure outlined previously. In both cases, two distinct eigenvalue bulks can be noticed. The smaller bulks on the left are made of six eigenvalues, and, since we have $\bar{N}_{\mathrm{S\&P}} = \bar{N}_{\mathrm{FTSE}} = 7$, this is in agreement with the prediction given by Eq. (29) for $\bar{N} = 7$. Also, in the FTSE350 Index case [Fig. 5(b)] the larger eigenvalue bulk around one is made of $N_{\mathrm{FTSE}} - \bar{N}_{\mathrm{FTSE}} = 21$ eigenvalues, which is again in agreement with (29), while the largest eigenvalue in the spectrum (not shown in the plot) is equal to 5.235, remarkably close to the prediction given by $\bar{N}_{\mathrm{FTSE}}\ \rho_{\mathrm{FTSE}} + (1 - \rho_{\mathrm{FTSE}}) = 5.242$ [see again Eq. (29)]. On the other hand, the spectrum relative to the S&P500 Index yields two large eigenvalues [not shown in Fig. 6(b)] equal to 3.552 and 6.483, and neither value is in agreement to the large eigenvalue prediction $N_{\mathrm{S\&P}}\ \rho_{\mathrm{S\&P}} + (1 - \rho_{\mathrm{S\&P}}) = 5.272$. Such discrepancy is due to the unresolved correlation structure in the empirical S&P matrix [see Fig. 5(a)], which gives rise to additional subclusters.

In Figs. 6(c) and 6(d) the two eigenvalue bulks we just discussed are fitted with the eigenvalue density deriving from the second equation in (38) when applied to the solution of (43), that is, the moment generating function $m_{\mathbf{c}}$ of the noise-dressed version of a correlation matrix $\mathbf{C}$ with degenerate eigenvalues. In both cases we consider correlation matrices with two degenerate eigenvalues in order to try to fit the two main bulks. The smaller eigenvalue Λ_1, responsible for the emergence of the smaller bulks on the left, is assumed to be equal to $1 - \rho$, accordingly to Eq. (29). So in the two different cases we analyzed we have

$$\begin{aligned} \Lambda_{1_{\mathrm{S\&P}}} &= 1 - \rho_{\mathrm{S\&P}} = 0.288, \quad w_{1_{\mathrm{S\&P}}} = \frac{\bar{N}_{\mathrm{S\&P}} - 1}{N_{\mathrm{S\&P}} - 2}, \\ \Lambda_{1_{\mathrm{FTSE}}} &= 1 - \rho_{\mathrm{FTSE}} = 0.293, \quad w_{1_{\mathrm{FTSE}}} = \frac{\bar{N}_{\mathrm{FTSE}} - 1}{N_{\mathrm{FTSE}} - 1}, \end{aligned} \tag{49}$$

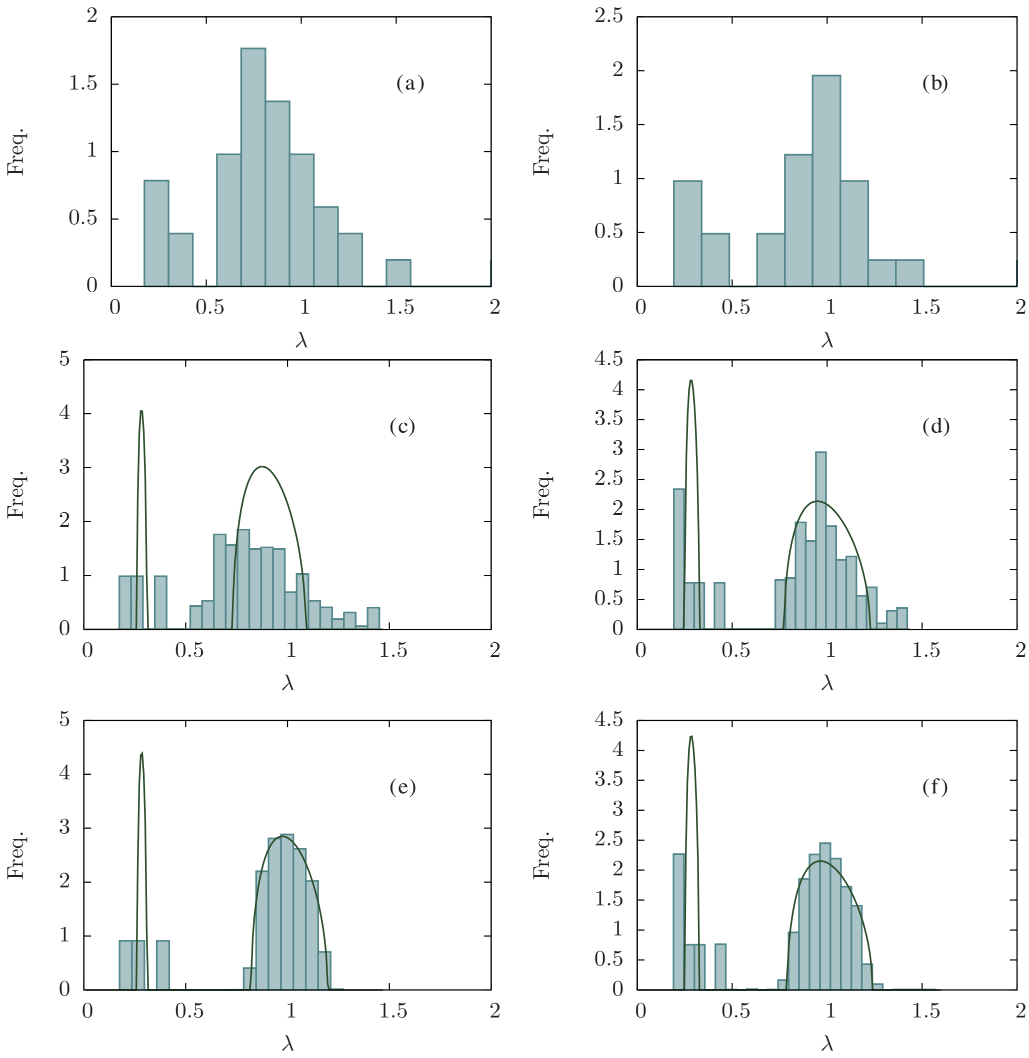

FIG. 6. (Color online) (a) and (b) Eigenvalue densities of the correlation matrices represented in Figs. 5(a) and 5(b). In both cases one can clearly distinguish two well separated bulks, while the largest eigenvalues have not been plotted for better visualization (see main text for further explanation). (c) and (d) Comparison between the theoretically expected spectra derived via Eq. (43) (solid lines) and the empirical ones. The latter have been modified with respect to (a) and (b) according to the following approach. A bootstrap random sampling (100 iterations) has been performed on the weakly correlated subsets of stocks, picking 30 stocks out of 33 in the S&P500 Index case and 18 out of 21 in the FTSE350 Index case. The presence of undetected structures [see Figs. 5(a) and 5(b) and main text] leads to a poor agreement between data and theory. The values and weights of the eigenvalues used to plot the theoretical density obtained from Eq. (38) are detailed in (49). (e) and (f) As in (c) and (d) but with weakly correlated data reshuffled before bootstrap, leading to a much better agreement between data and theory. The reference eigenvalue for the bulks on the right is now assumed to be equal to one (see main text).

where the two slightly different weights are justified by the previously mentioned fact that in the S&P case there are two isolated eigenvalues which separate from the main bulks, while in the FTSE case there is only one such eigenvalue. On the other hand, the larger eigenvalue Λ_2, which according to (29) should be exactly equal to one, is assumed to be equal to the empirical mean value of the main bulks on the right in Figs. 6(c) and 6(d). These are found to be

$$\Lambda_{2_{\mathrm{S\&P}}} = 0.887, \quad \Lambda_{2_{\mathrm{FTSE}}} = 0.997 \tag{50}$$

and one might notice that, again, the value obtained in the FTSE case is in excellent agreement with the theoretically expected one. So, all in all, the curves drawn in Figs. 6(c) and 6(d) are obtained from the values in Eqs. (49) and (50). Such curves, as already mentioned, are fitted to the empirical spectra. However, a bootstrap approach was adopted in order to improve the statistics. More specifically, for each bootstrap iteration a random sampling on the weakly correlated stocks was performed, picking 30 out of 33 in the S&P case and 18 out of 21 in the FTSE case. On the contrary, the stocks forming the highly correlated clusters were always kept (thus keeping the eigenvalue bulks on the left almost unchanged). As can be seen in Figs. 6(c) and 6(d) the agreement between theory and prediction is very poor. This is essentially due to the additional correlation structures in the empirical correlation matrices (see Fig. 5), which are neglected in the model matrix (44) and in its eigenvalue spectrum (29). All the bulks displayed in

Figs. 6(c) and 6(d) appear to be "smeared" versions of their theoretical counterparts, even the small ones relative to the eigenvalues in (49). Interestingly this shows that inhomogeneities in correlation structures have quite an impact on eigenvalue spectra even on a "small scale" (let us recall that $\bar{N}_{\text{S\&P}} = \bar{N}_{\text{FTSE}} = 7$).

In Figs 6(e) and 6(f) the same fit as the one just discussed is performed, the only difference being that an additional random reshuffling of the returns is performed on the bootstrapped assets. Such an operation is meant to destroy all possible correlations, and this leads to a quite good agreement between data and predictions on the bulks on the right [the theoretical densities being now computed with $\Lambda_2 = 1$, accordingly to Eq. (29)]. This essentially confirms that the substantial deviations shown in Figs. 6(c) and 6(d) can entirely be imputed to the unresolved cluster structures in the empirical correlation matrices. The same kind of analyses (bootstrap and reshuffling) were not performed on the stocks belonging to the correlated clusters because of their very small number.

All in all, the previous observations definitely suggest that the empirically observed eigenvalue bulks cannot be regarded as a consequence of the noisiness in financial correlation matrices. On the contrary, in the light of the previous discussions it could be conjectured that bulks emerge from the interplay of several cluster structures like the ones we isolated (see Fig. 5) [12,13].

IV. SUMMARY AND CONCLUSIONS

Let us now summarize the main messages in the paper.

(1) Several rough but useful results about spectral properties of financial correlation matrices, such as the position of large nondegenerate eigenvalues, can be inferred by a clever application of the direct problem (see Sec. II A). This only involves algebraic calculations, namely the solution of suitable secular equations. This approach can be used either when the cluster structure is known *a priori*, or when there are good reasons to assume a certain correlation structure. Combining the direct analysis with Monte Carlo simulations can provide a clear picture in a number of situations, avoiding the analytical difficulties of random matrix theory, and keeping the finite-sized nature of the problem. Typically, one wishes to reproduce observed spectra starting from a factor model, and this can be done as follows.

(a) Identify the cluster structure in the data set under analysis, using clustering algorithms [34].

(b) Estimate the average correlations within clusters.

(c) Build a theoretical, "mean field," matrix model **C** from the above estimates.

(d) Run Monte Carlo simulations of the matrix model.

(e) Compare the outcome of the simulation to the empirical spectrum.

If the comparison is statistically satisfactory, the matrix model **C** can be retained and used for further analyses, such as portfolio selection. If not, the model is to be refined, for example, by studying more carefully the correlation structure of the data or by abandoning the mean field assumption, at least allowing for some cluster interactions.

(2) As far as the largest eigenvalue is concerned, its distribution is not Tracy-Widom, but Normal [30,31]. Moreover, this distribution cannot be derived from the thermodynamic limit formula (43). In fact, such an eigenvalue is typically nondegenerate and its weight in a diagrammatic expansion of the Green's function would vanish as $1/N$, for $N \to \infty$.

(3) For factor models the bulks in empirical eigenvalue spectra come as the noise-dressed version of degenerate eigenvalues. Thus such bulks encode the information on the cluster structure of the empirical correlation matrix **c**, and this can be evidenced by means of proper clustering methods, as done in Sec. III.

(4) The results we obtain in Sec. III by means of our filtering procedure suggest that empirically observed eigenvalue bulks in financial correlation matrix spectra emerge as a consequence of the subtle interplay between factor eigenvalues and noise. More specifically, the whole eigenvalue bulks seem to arise as superpositions of smaller structures, such as the ones shown in Figs. 6(a) and 6(b), which merge together according to the mechanism shown in Fig. 3. So one could safely state that the empirically observed eigenvalue bulks are not a mere consequence of noise. As a matter of fact they arise as the noise dressing of degenerate eigenvalues which do carry information on the correlation structure of the data under study.

While there would be no difficulty in studying non-Gaussian multivariate models by means of Monte Carlo simulations, the analytical results presented in Sec. II B cannot be easily generalized. In fact, the integrals needed to calculate $\mathbf{g_c}$ in (34) in the Gaussian case can be exactly obtained by virtue of Wick's theorem, whereas different stochastic models would require painful calculations.

The diagrammatic method outlined in [28] allows, in principle, for the exact evaluation of the Green's function $\mathbf{g_c}$ for any finite size $N \times T$, as a function of N and T. Nevertheless, this is a series of $1/z$ powers, whose convergence properties would be interesting to investigate in the near future.

ACKNOWLEDGMENTS

Simone Alfarano acknowledges financial support by the Spanish Ministry of Science and Innovation from research project ECO2008-00510, and by Universitat Jaume I - Bancaixa from research project P11A2009-09. Enrico Scalas is grateful to Universitat Jaume I for the financial support received from their Research Promotion Plan during his scientific visit in Castellón, where the authors met to work on this paper. Giacomo Livan wishes to thank G. Montagna and O. Nicrosini for their kind support and helpful suggestions. The authors are indebted to Z. Burda for useful discussions and to P. Vivo for pointing out reference [33].

[1] E. P. Wigner, Ann. Math. **62**, 548 (1955); **67**, 325 (1958).
[2] M. Mehta, *Random Matrices* (Elsevier, Amsterdam, 2004).
[3] J. Wishart, Biometrika **20**, 32 (1928).
[4] V. A. Marčenko and L. A. Pastur, Math. USSR-Sb **1**, 457 (1967).
[5] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).

[6] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[7] A. N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, New York, 1956).

[8] P. Billingsley, *Probability and Measure* (John Wiley & Sons, New York, 1995).

[9] J. Y. Campbell, A. W. Lo, and A. Craig MacKinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, 1997).

[10] Z. Burda, A. Jarosz, J. Jurkiewicz, M. A. Nowak, G. Papp, and I. Zahed, e-print arXiv:physics/0603024.

[11] J. W. Silverstein and Z. D. Bai, J. Multivariate Anal. **54**, 175 (1995).

[12] Z. Burda, A. Görlich, A. Jarosz, and J. Jurkiewicz, Physica A **343**, 295 (2004).

[13] Z. Burda and J. Jurkiewicz, Physica A **344**, 67 (2004).

[14] H. Markowitz, J. Financ. **7**, 77 (1952).

[15] W. Sharpe, J. Financ. **19**, 425 (1964).

[16] F. Lillo and R. N. Mantegna, Phys. Rev. E **72**, 016219 (2005).

[17] G. Raffaelli and M. Marsili, J. Stat. Mech.-Theory E (2006) L08001.

[18] M. Marsili, G. Raffaelli, and B. Ponsot, J. Econ. Dyn. Control **33**, 1170 (2009).

[19] S. Pafka and I. Kondor, Eur. Phys. J. B **27**, 277 (2002).

[20] J. Shlens, *A Tutorial on Principal Component Analysis: Derivation, Discussion and Singular Value Decomposition*, available online at [http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf].

[21] J. Kwapien, S. Drożdż, and P. Oswiecimka, Physica A **359**, 589 (2006).

[22] G. Akemann, J. Fischmann, and P. Vivo, Physica A **389**, 2566 (2010).

[23] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).

[24] T. Guhr and B. Kalber, J. Phys. A **36**, 3009 (2003).

[25] O. E. Barndorff-Nielsen and S. Thorbjørnsen, Proc. Natl. Acad. Sci. USA **99**, 16568 (2002).

[26] M. Politi, E. Scalas, D. Fulger, and G. Germano, Eur. Phys. J. B **73**, 13 (2010).

[27] C. Itzykson and J.-M. Drouffe, *Statistical Field Theory* (Cambridge University Press, Cambridge, 1989).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.84.016113 for an outline of the diagrammatic method employed to derive Eq. (39) and for a Matlab® program giving the solution of Eq. (43) for $L = 2$ [and the corresponding eigenvalue density via Eq. (38)].

[29] Z. Burda, A. Görlich, J. Jurkiewicz, and B. Wacław, Eur. Phys. J. B **49**, 319 (2006).

[30] C. A. Tracy and H. Widom, *The Distributions of Random Matrix Theory and Their Applications*, available online at [http://www.math.ucdavis.edu/∼tracy/talks/SITE7.pdf].

[31] D. Paul, Stat. Sinica **17**, 1617 (2007).

[32] M. A. Stephens, J. Am. Stat. Assoc. **69**, 730 (1974).

[33] A. Edelman and P. O. Persson, e-print arXiv:math-ph/0501068.

[34] M. S. Aldenderfer and R. K. Blashfield, *Cluster Analysis* (SAGE, Newbury Park, California, 1995).