Spectral density of the correlation matrix of factor models: A random matrix theory approach

F. Lillo[1,2,3] and R. N. Mantegna[2,3]
[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
[2]*INFM Unità di Palermo and Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, Viale delle Scienze, I-90128 Palermo, Italy*
[3]*Istituto Nazionale di Fisica Nucleare, Sezione di Catania, Catania, Italy*
(Received 10 September 2003; revised manuscript received 6 December 2004; published 28 July 2005)

We studied the eigenvalue spectral density of the correlation matrix of factor models of multivariate time series. By making use of the random matrix theory, we analytically quantified the effect of statistical uncertainty on the spectral density due to the finiteness of the sample. We considered a broad range of models, ranging from one-factor models to hierarchical multifactor models.

DOI: 10.1103/PhysRevE.72.016219 PACS number(s): 05.45.Tp, 02.10.Yn, 02.50.Ey, 05.40.Ca

I. INTRODUCTION

The current investigation of experimental and real systems often allows the parallel recording of time series, both in experiments and in the monitoring of a wide number of physical, biological, and social systems. The result of this scientific and technological achievement is that the number of multivariate time series describing an experiment, or monitoring a system, is constantly increasing. A natural instrument in the investigation of a multivariate time series is the correlation matrix. In the case of Gaussian multivariate distributions, the matrix is said to belong to the Wishart distribution [1]. The study of the properties of the correlation matrix has a direct relevance in the investigation of mesoscopic physical systems [2], high-energy physics [3], information theory and communication [4–6], physiological data [7,8], investigation of microarray data in biological systems [9–11] and econophysics [12–18].

The extraction of information from a multivariate time series is therefore a central issue in many scientific investigations. Several classical methods have been introduced to this end, ranging from principal component analysis to clustering methods [19,20]. Multivariate analysis methods are designed to extract the information both about the number of main factors characterizing the time dynamics of the investigated system and the composition of the groups (clusters) in which the system is intrinsically organized. Empirical models describing the dynamics of a system, in terms of a finite number of factors, are termed factor models [11,19,21].

Any real experiment or monitoring of a real system is performed by obtaining a finite sample of T records for each variable. The finiteness of the number of sampled records implies that the measured quantities in the analysis of the system behavior present an unavoidable degree of stastitical uncertainty. This fact has been recently expressed as the term "noise dressing" [12]. In the following, we use this term with the meaning of statistical uncertainty due to a finite value of the number of records T of the time series under investigation. In this paper, by using concepts and tools of the random matrix theory [22], we explicitly determine the amount of noise dressing of the eigenvalue spectrum of the correlation matrix of a large system which is described by a factor model and monitored in time by a large number of records.

The paper is organized as follows: In Sec. II, we provide a definition of the class of factor models we investigated and discuss the scientific questions answered in our paper. In Sec. III, we briefly recall some key aspects of the random matrix theory approach to the modeling of the spectral density of correlation matrix eigenvalues. Section IV presents the results obtained for a generic class of factor models driven in time, whereas Sec. V considers the results obtained for a factor model with a sinusoidal time dependence. Section VI briefly summarizes our conclusions. Some technical aspects are discussed in detail in two appendices.

II. OUTLINE OF THE PROBLEM

In this section, we first define the class of factor models we considered and explicitly state the scientific questions answered.

A. Factor models

The simplest and more widespread models of multivariate time series are factor models. In these models, the dynamics of each variable are the linear combination of a given number of factors plus a noise term. A more general multifactor model for N variables $x_i(t)$ $(i=1,\ldots,N)$ can be written as

$$x_i(t) = \sum_{j=1}^{K} \gamma_i^{(j)} f_j(t) + \gamma_i^{(0)} \epsilon_i(t). \tag{1}$$

In this equation, K is the number of factors $f_j(t)$, $\gamma_i^{(j)}$ is a constant describing the weight of factor j in explaining the dynamics of the variable $x_i(t)$, and $\epsilon_i(t)$ is a Gaussian zero-mean noise term with unit variance. The coefficients of the linear combination and the intensity of the noise terms are specific to each variable and assumed, for simplicity, to be time independent. Examples of such models are the Capital Asset Pricing Model (one factor) and the Arbitrage Pricing Theory (multifactor). Both models are widespread in the financial literature [21]. In Eq. (1), we assume that the factors are uncorrelated with each other, i.e., $\langle f_i(t) f_j(t) \rangle = \delta_{ij}$, where the symbol $\langle \ldots \rangle$ indicates an average in time. Also, the noise terms are uncorrelated with each other and with the factors,

i.e., $\langle \epsilon_i(t)\epsilon_j(t)\rangle = \delta_{ij}$ and $\langle f_i(t)\epsilon_j(t)\rangle = 0$. Since, in the rest of this paper, we are interested in studying the linear correlation coefficients, we assume that all the variables x_i have zero-mean and unit variance without loss of generality. These assumptions fix the value $\gamma_i^{(0)}$ through the relation $(\gamma_i^{(0)})^2 = 1 - \Sigma_j(\gamma_i^{(j)})^2$.

Another class of factor models we considered describes the dynamics of the variables driven by one or more sinusoidal signals of a given frequency, with each variable characterized by a different phase. This kind of model has been recently applied to gene expression analysis monitored by microarray data during the cell cycle [9–11]. In this second case, variables follow a common factor which is sinusoidal in time.

B. Noise dressing

Given a set of N time series each recorded for a number T of records, one could ask the question of whether a factor model of the type of Eq. (1) can be used to describe the dynamics of the N variables. In statistics, there are two classes of questions that can be posed: (i) Given a factor model of Eq. (1) with a given number K of factors and with γ parameter known precisely or statistically (i.e., what is known is the probability distribution from which the γs are drawn), can we perform a statistical test of the hypothesis that the empirical data are well described by the model? (ii) What is the "best" choice of the parameters (γ and number of factors K) that describe the data? The two problems are known as *hypothesis testing* and *parameter estimation*, respectively. In this paper, we address only the first question. In most of the hypothesis testing problems, one of the major difficulties of the test comes from the finiteness of the data. In any real experiment, one can record a finite number of data (NT in our case) and the unavoidable statistical fluctuations lead to measured quantities which are different from the ones expected from the model. For a given model, the estimate of a parameter obtained from a sample of finite size T is distributed according to a probability distribution whose dispersion (e.g., standard deviation) tends to zero for $T \to \infty$. For example, the sample mean of an independent identically distributed set of T variables is asymptotically Gaussian distributed around the true mean value and with a standard deviation proportional to $1/\sqrt{T}$.

In the case of a factor model of Eq. (1), one has to choose the parameters to be estimated. We are interested in the correlation matrix of the variables x_i. The correlation coefficient between two variables x_i and x_j is defined as

$$C_{ij} \equiv \frac{\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle}{\sqrt{(\langle x_i^2\rangle - \langle x_i\rangle^2)(\langle x_j^2\rangle - \langle x_j\rangle^2)}}, \tag{2}$$

where again the symbol $\langle \ldots \rangle$ indicates an average in time. The correlation matrix $\mathbf{C}$ is the $N \times N$ symmetric matrix whose C_{ij} element is the linear correlation coefficient between the pair of variables x_i and x_j. Important properties of the correlation matrix are contained in its eigenvalue spectrum. One of the most important multivariate techniques, known as the principal component analysis [19] (also known as singular value decomposition or Karhunen–Loeve transform), is based on the analysis of the eigenvalue spectrum. We make use of asymptotic methods that are rigorously valid in the limit $N \to \infty$, when the number of eigenvalues of the correlation matrix becomes infinite. In this sense, it is useful to introduce the spectral density $\rho(\lambda)$ which is a continuous function describing the eigenvalues distribution. For a matrix having eigenvalues λ_n, $(n = 1, \ldots, N)$, the spectral density is

$$\rho(\lambda) = \sum_{n=1}^{N} \delta(\lambda - \lambda_n), \tag{3}$$

where $\delta(x)$ is the Dirac delta function. This quantity is normalized as $\int_0^\infty \rho(\lambda) d\lambda = N$.

The purpose of this paper is twofold. First, we study the problem of calculating the eigenvalue spectrum of a factor model described by Eq. (1). This is done by making assumption on the statistical properties of the parameters γ describing the model. The second purpose of the paper is the quantification of the statistical uncertainty of the spectral density of the correlation matrix of a factor model described by Eq. (1) when the x_i variables are measured in a finite time interval of T records. This quantification is useful for testing the hypothesis that real data under investigation are well described by the considered model. In order to solve this problem we make use of the Random Matrix Theory (RMT) [22]. By considering the assumptions needed to use asymptotic methods of the RMT, most of the results we derive are valid in the limit of $N \to \infty$ and $T \to \infty$, even though we have verified that several of them are very good also for finite large matrices.

The spectrum of a factor model is in general composed of a set of large eigenvalues describing the behavior of the factors (the "signal") and another set of small eigenvalues describing the effect of the noise terms (the "noise"). Most of the applications of the RMT to correlation matrices have focused on the quantification of the noise dressing in the noise part of the spectrum. This has been done either for uncorrelated variables or for correlated variables with the underlying assumption that the variables are homogeneous with respect to the signal. On the contrary, in this paper, we present a method which; (i) Allows one to compute the effect of noise dressing on the part of the spectral density describing the factors, i.e., the signal, and (ii) shows that the heterogeneity of the variables with respect to the factors [the $\gamma_i^{(j)}$ in Eq. (1)] also changes the properties of the noise part of the spectrum.

III. RANDOM MATRIX APPROACH

The application of the RMT to the noise dressing of correlation matrices has recently been addressed [18,23–25] and applied to the study of financial correlation matrices [12–15,17]. The method to obtain the spectral density can be summarized as follows [23]. The $N \times N$ correlation matrix can be thought as the product $\mathbf{M}^T\mathbf{M}$, where $\mathbf{M}^T$ is the $N \times T$ matrix containing the original data $x_i(t)$ from which the mean is subtracted and the result is divided by the standard deviation. It is then useful to introduce the resolvent

$$\mathcal{G}(z) = \mathrm{Tr}[(z - \mathbf{M}^T\mathbf{M})^{-1}] = \sum_{n=1}^{N} \frac{1}{z - \lambda_n}, \tag{4}$$

where $\mathcal{G}(z)$ is a complex function. The resolvent is related to the spectral density through

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0} \mathrm{Im}[\mathcal{G}(\lambda - i\epsilon)]. \tag{5}$$

The resolvent can be rewritten as

$$\mathcal{G}(z) = \partial_z \ln \det(z\mathbf{I_N} - \mathbf{C}). \tag{6}$$

where $\mathbf{I_N}$ is the $N \times N$ identity matrix. These expressions hold for the correlation matrix of the model. One can imagine that the elements of the matrix $\mathbf{M}$ are equal to the sum of an "undressed" part $\mathbf{M}_0$ plus a random part $\mathbf{N}$ describing the effect of the noise (finiteness of the sample). In order to find the effect of noise dressing on the spectral density, one has to average the resolvent $\mathcal{G}(z)$ over the probability distribution of the matrix $\mathbf{M}$, i.e.,

$$G(z) = \langle \mathcal{G}(z) \rangle_{\mathrm{ens}}, \tag{7}$$

where the symbol $\langle \ldots \rangle_{ens}$ indicates an average over the probability distribution of the matrix $\mathbf{M}$. This averaging procedure can be made by making use of the replica trick and by performing a saddle point approximation (see [23] for details).

IV. MODELS WITH TIME UNCORRELATED FACTORS

Let us first consider the general case of a factor model in which the factors $f_j(t)$ are stochastic variables uncorrelated in time, i.e., $\langle f_i(t) f_j(t') \rangle = \delta_{ij}\delta_{tt'}$. In this case, we have also that $\langle x_i(t) x_j(t') \rangle_{\mathrm{ens}} = C_{ij}\delta_{tt'}$. By making use of Eq. (25) of Ref. [23], we direcly show that the equation for the ensemble averaged resolvent is

$$G(z) = \frac{T}{z - \sum_{i=1}^{N} \frac{\lambda_i}{T - \lambda_i G(z)}}. \tag{8}$$

In order to find the effect of noise dressing on the spectral density of this kind of factor model, one needs to; (i) Find the model spectrum $\lambda_1, \ldots, \lambda_N$ of the factor model of Eq. (1), (ii) solve the $N+1$th degree algebraic Eq. (8), and (iii) make use of Eq. (5) with $G(z)$ in place of $\mathcal{G}(z)$ to find $\rho(\lambda)$. The major analytical or computational difficulties are the determination of the spectrum and the solution of Eq. (8). Hereafter, we introduce a scheme able to solve this problem in many cases of interest. In the following paragraphs, we consider factor models of increasing complexity.

A. Zero-factor model

In the absence of any factor, we have a zero-factor model. In the zero-factor model, each variable is described only by a random Gaussian variable $\epsilon_i(t)$. The model correlation matrix is the $N \times N$ identity matrix $\mathbf{I_N}$, thus all the eigenvalues are equal to 1 and the spectral density is $\rho(\lambda) = N\delta(\lambda - 1)$. The noise dressing of the spectrum of the sample correlation matrix has been derived in [23,24]. We report here the results for completeness and for comparison with the results of more complicated models. Moreover, in this paragraph, we assume that all the x_i variables have zero mean and variance σ^2. The reason for considering the more general case of $\sigma \neq 1$ here is that in the following we will need this result. In the limit $T, N \to \infty$, with a fixed ratio $Q = T/N \geqslant 1$, the eigenvalue spectral density of the correlation matrix is given by

$$\rho(\lambda) = \frac{T}{2\pi\sigma^2\lambda} \sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}, \tag{9}$$

where $\lambda_{\min}^{\max} = \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q})$. The spectral density is different from zero in the interval $]\lambda_{\min}, \lambda_{\max}[$.

B. One-factor model

As a first application of this method, we consider a one-factor model in which the dynamics of each variable are controlled by a single factor. The equations describing the one-factor model are given by

$$x_i(t) = \gamma_i f(t) + \gamma_i^{(0)} \epsilon_i(t), \tag{10}$$

i.e., Eq. (1) with $K = 1$. The parameter γ_i^2 gives the fraction of variance explained by the common factor $f(t)$. The model describes the situation when the N variables are essentially controlled by a common factor describing a weighted mean. This type of model is, for example, consistent with the Capital Asset Pricing Model of stock market behavior.

We directly show that the correlation coefficient between variable i and j described by Eq. (10) is $C_{ij} = \gamma_i\gamma_j$. The correlation matrix of the one-factor model can therefore be written as $\mathbf{C} = \mathbf{A} + \mathbf{b}\mathbf{b}^+$, where $\mathbf{A} = \mathrm{diag}(1 - \gamma_i^2)$ is a diagonal $N \times N$ matrix and $\mathbf{b}^+ = (\gamma_1, \ldots, \gamma_N)$ is a row vector. The characteristic equation of $\mathbf{C}$ can be calculated by using the Sherman–Morrison formula [26],

$$\det(\mathbf{A} + \mathbf{b}\mathbf{b}^+) = \det \mathbf{A}(1 + \mathbf{b}^+\mathbf{A}^{-1}\mathbf{b}), \tag{11}$$

and the result is

$$\det(\mathbf{C} - \mathbf{I_N}\lambda) = \prod_{i=1}^{N} (1 - \gamma_i^2 - \lambda) \left[1 + \sum_{i=1}^{N} \frac{\gamma_i^2}{1 - \gamma_i^2 - \lambda} \right] = 0. \tag{12}$$

In the following, we distinguish the case when all the γ_i are the same (*degenerate* case) from the case when the γ_i are extracted from some known probability distribution (*nondegenerate case*). We will show that the spectral properties of the two types of models are quite different.

1. Degenerate model

In the case of a degenerate one-factor model, i.e., when $\gamma_i = \gamma$ for all values of i, the characteristic equation (12) can be solved and the spectrum is composed by a large eigenvalue $\lambda_1 = 1 + (N-1)\gamma^2 \simeq N\gamma^2$ and $N-1$ degenerates eigenvalues $\lambda_0 = 1 - \gamma^2$ [17].

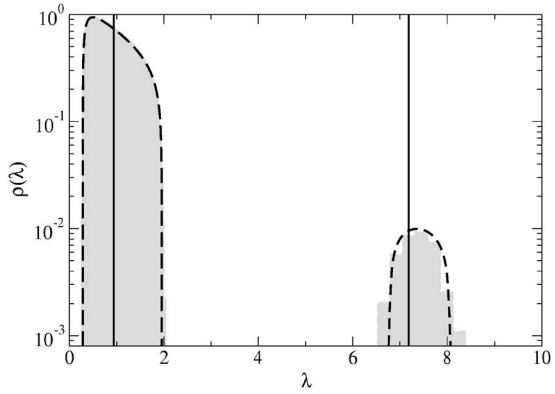

FIG. 1. Spectral density of a degenerate one-factor model. The gray areas are the average over 1000 numerical simulations of a one-factor model of $N=100$ variables for $T=500$ time steps. The value of γ is 0.25. The dashed line is the theoretical prediction and the vertical lines indicate the spectrum of the model.

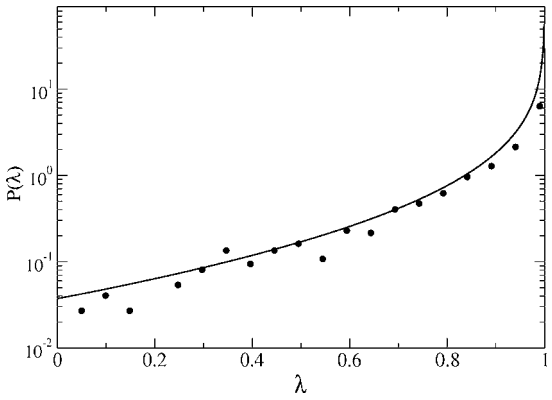

FIG. 2. The low part of the spectrum of the correlation matrix of a nondegenerate one-factor model (circle). In this case, $N=2000$ and the γ_i are extracted from an exponential distribution with $\bar{\gamma}=0.25$. The continuous line is the prediction based on the ansatz discussed in the text.

Once we have the spectrum of the model, we turn into the solution of Eq. (8) in order to find the related noise dressed spectrum. Equation (8) for the resolvent becomes a third-degree algebraic equation

$$z\lambda_0\lambda_1 G^3(z) + (-Tz\lambda_0 - Tz\lambda_1 + N\lambda_0\lambda_1 - T\lambda_0\lambda_1)G^2(z) + (T^2 z + T\lambda_0 - NT\lambda_0 + T^2\lambda_0 - T\lambda_1 + T^2\lambda_1)G(z) - T^3 = 0. \quad (13)$$

The spectral density can be obtained analytically from Eq. (13), even if the expression is quite long. As expected, the spectral density is different from zero in two intervals, one for the $N-1$ small eigenvalues and one for the large eigenvalue λ_1. Numerical calculations and analytical considerations show that the low part of the spectrum is well fitted by the functional form of Eq. (9). The width of the two intervals scale with the parameter of the model as $\Delta_0 \sim (1-\gamma^2)\sqrt{N/T}$ and $\Delta_1 \sim N\gamma^2/\sqrt{T}$, where $\Delta_0(\Delta_1)$ is the width of the low (high) part of the spectrum. Figure 1 shows the comparison of theoretical prediction and numerical simulations of a degenerate one-factor model. The agreement is very good in the whole range of eigenvalues. It is worth noting that such an agreement is obtained also when $T<N$. Figure 1 also shows the spectrum of the model $\rho(\lambda)=(N-1)\delta(\lambda-\lambda_0) + \delta(\lambda-\lambda_1)$ as vertical lines.

Finally, it is worth mentioning that for the degenerate one-factor model, one can obtain the functional dependence of the low part of the spectrum in a simpler way. A reasonable idea [12] is that the components of the correlation matrix which are orthogonal to the eigenspace associated with the largest eigenvalue are described by pure noise. This idea amounts to subtracting the contribution of λ_1 from the variance of the variables. In other words, one could use the equation of the zero-factor model, Eq. (9), with $\sigma^2=1-\lambda_1/N$. The corresponding spectral density $\rho(\lambda)$ is essentially indistinguishable from the low part of the exact result obtained from Eq. (13) and, for this reason, we do not show the corresponding result in Fig. 1. On the other hand, our method is able to give information on the distribution of the highest eigenvalue λ_1. It is worth noting that this last information cannot be obtained by using the simpler approximate approach discussed above.

2. Nondegenerate model

We come now to the more complicated case of a nondegenerate one-factor model. In this case, we suppose that the γ parameters are drawn from a known probability distribution $P(\gamma_i)$. In order to find the spectrum, one should solve Eq. (12) for all the eigenvalues and then solve the $(N+1)$th degree polynomial of Eq. (8). This task is too complicated even numerically. We are not able to solve Eq. (12) in the nondegenerate case, but we are able to provide an approximate form of the spectrum when N is large. In the nondegenerate case, we can still expect a spectral density composed by a large eigenvalue and $N-1$ small eigenvalues. The large eigenvalue can be obtained by equating to zero the term in square brackets in Eq. (12). In fact, since the largest eigenvalue is much larger than $1-\gamma_i^2$ for any i, we can approximate the term in the square bracket as $\Sigma\gamma_i^2/\lambda_1 \simeq 1$, i.e., $\lambda_1 \simeq \Sigma_{i=1}^N \gamma_i^2 = N\langle\gamma_i^2\rangle_{\text{par}}$, where the symbol $\langle\ldots\rangle_{\text{par}}$ indicates an average over the probability distribution $P(\gamma_i)$ of the γ parameters. In order to have insight into the spectral density for the other $N-1$ we make the ansatz that the distribution of λ_i is given by $P(\lambda)=P(\gamma_i)d\gamma_i/d\lambda$ where $\gamma_i=\sqrt{1-\lambda}$. The idea behind this ansatz is that the relation between eigenvalues and γ_i is the same as in the degenerate case. For example, if γ_i is distributed uniformly in a subinterval $[m-d,m+d]$ of $[0, 1]$, the distribution of the $N-1$ eigenvalues is given by $P(\lambda) = (4d)^{-1}(1-\lambda)^{-1/2}$. Another case of interest for the following is when γ_i is distributed exponentially $P(\gamma_i) \propto \exp(-\gamma_i/\bar{\gamma})$ with $0<\gamma_i<1$. In this case, the distribution of the $N-1$ eigenvalues is proportional to $\exp\left(-\sqrt{1-\lambda}/\bar{\gamma}\right)(1-\lambda)^{-1/2}$. Note that under our ansatz, the low part of the spectrum is bounded from above by the value $\lambda=1$. In Fig. 2, we show the low part of the eigenvalue spectral density of a one-factor model where the γ_i are distributed exponentially. The line is the theoretical prediction based on our ansatz. The agreement between data and the ansatz is quite good.

Once we have determined the distribution of small eigenvalues, we turn to the determination of the noise dressing. The sum in the denominator of Eq. (8) is split in a term for λ_1 plus a sum over the remaining $N-1$ small eigenvalues. This last term can be computed as $N-1$ times the average of $\lambda/[T-\lambda G(z)]$ over $P(\lambda)$ introduced in our ansatz. In other words, Eq. (8) for the resolvent becomes

$$G(z) = \frac{T}{z - \dfrac{\lambda_1}{T-\lambda_1 G(z)} - (N-1)\left\langle \dfrac{\lambda_i}{T-\lambda_i G(z)} \right\rangle_{\text{par}}}, \quad (14)$$

where $\lambda_1 \simeq 1+(N-1)\langle \gamma_i^2 \rangle_{\text{par}}$. In general, the average term in Eq. (14) is not a rational function and, therefore, Eq. (14) cannot be reduced to an algebraic equation in G. Thus the equation for $G(z)$ is no more algebraic, but it becomes in general trascendental. In order to solve the complex transcendental Eq. (14) we introduce a simple strategy. The average term in Eq. (14) depends typically on the dispersion of the $P(\gamma_i)$ at the second order. Therefore, the low part of the spectrum of a nondegenerate one-factor model with a small dispersion in γ_i should not be very different from a degenerate one-factor model with $\gamma_{\text{eff}} = \sqrt{\langle \gamma_i^2 \rangle_{\text{par}}}$. We can therefore use the value of the resolvent of this effective degenerate one factor model as the starting point for the numerical search of the solution of Eq. (14). Quite surprisingly, this method also works well when the dispersion of the γ_i is high.

As an example, we observe that when γ_i is distributed uniformly in a subinterval $[m-d, m+d]$ of $[0,1]$ the average term in Eq. (14) is

$$\left\langle \frac{\lambda_i}{T-\lambda_i G} \right\rangle_{\text{par}} = -\frac{1}{G} + \frac{T}{2d\, G^{3/2}\sqrt{G-T}} \times \left(\operatorname{arctanh}\left[\frac{\sqrt{G}(m-d)}{\sqrt{G-T}} \right] - \operatorname{arctanh}\left[\frac{\sqrt{G}(m+d)}{\sqrt{G-T}} \right] \right), \quad (15)$$

and $\lambda_1 = 1+(N-1)\left(m^2+\frac{d^2}{3}\right)$. For each value of z, we solve the trascendental Eq. (14) for $G(z)$ and by taking its imaginary part [see Eq. (5)], we find $\rho(\lambda=z)$.

In Fig. 3, we show the low part of the spectrum for a one-factor model in which γ_i is uniformly distributed between 0 and 1. We see that the agreement between the theory and the simulations is very good.

It is worth noting that in the general case of a nondegenerate one-factor model, the low part of the spectrum is not compatible with the form of Eq. (9). For example, let us consider the nondegenerate one-factor model with γ_i uniformly distributed between 0 and 1 in which $\sigma^2 = 1-\lambda_1/N = 0.667$. For this model, the low part of the spectral density predicted by the adjusted zero-factor model is shown in Fig. 3 as a dotted line. It is evident that the adjusted zero-factor model is unable to describe the simulation data, whereas the theory based on our approach fits very well the simulations. Specifically the adjusted zero factor model overestimates (underestimates) the lower (upper) edge of the low part of the spectral density. Suppose that one wants to compare an empirical eigenvalue spectrum with the spectral density of a one-factor model computed by using the adjusted zero-factor model. Figure 3 shows that an eigenvalue slightly larger than the upper bound of the spectral density of the adjusted zero factor model would erroneously be interpreted as a signal eigenvalue not explained by the one-factor model. The full theory of noise dressing of one-factor model developed here shows that a distribution in the γ_i parameters gives a larger interval of predicted noise eigenvalues.

The inset of Fig. 3 shows the comparison between the spectrum of the model expected for $T \to \infty$ and the spectrum predicted by our method for a realization of the nondegenerate one-factor model with finite T. In the low part of the spectrum, the difference between the two curves is quite significant. Specifically, the model spectrum is bounded from above by $\lambda=1$ and in it diverges at this value. On the other hand, when the time series is finite it is possible to observe eigenvalues larger than 1 in the low part of the spectrum. This example shows the importance of a careful characterization of the effect of finite values of T on the shape of eigenvalue spectrum.

A similar good agreement is observed for exponentially distributed γ_i. If γ_i is distributed exponentially, $P(\gamma_i) \propto \exp(-\gamma_i/\bar{\gamma})$ in $0 < \gamma_i < 1$, the average term becomes

$$\left\langle \frac{\lambda_i}{T-\lambda_i G(z)} \right\rangle_{\text{par}} = \Bigg[e^{-\sqrt{G(z)-T}/\sqrt{G(z)}\,\bar{\gamma}} \Bigg(-2e^{\sqrt{G(z)-T}/\sqrt{G(z)}\,\bar{\gamma}} \left(-1+e^{1/\bar{\gamma}}\right) \sqrt{G(z)}\,\bar{\gamma} G(z) - T + e^{1/\bar{\gamma}} T \operatorname{Ei}\left(\frac{-1+\dfrac{\sqrt{G(z)-T}}{\sqrt{G(z)}}}{\bar{\gamma}} \right) + e^{2\sqrt{G(z)-T}/\sqrt{G(z)}\,\bar{\gamma}} \left[-\operatorname{Ei}\left(-\frac{1+\dfrac{\sqrt{G(z)-T}}{\sqrt{G(z)}}}{\bar{\gamma}} \right) + \operatorname{Ei}\left(-\frac{\sqrt{G(z)-T}}{\sqrt{G(1z)}\,\bar{\gamma}} \right) \right] - \operatorname{Ei}\left(\frac{\sqrt{G(z)-T}}{\sqrt{G(z)}\,\bar{\gamma}} \right) \Bigg) \Bigg] \Bigg/ \left[2\left(-1+e^{1/\bar{\gamma}}\right) G^{3/2}(z)\,\bar{\gamma} \sqrt{G(z)-T} \right], \quad (16)$$

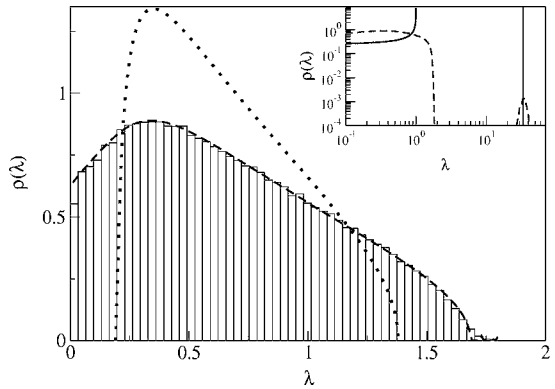

FIG. 3. Low part of the spectral density of a one-factor model in which γ_i is distributed uniformly between 0 and 1. We performed 1000 numerical simulations of a one-factor model of $N=100$ variables for $T=500$ time steps. The dashed line is the theoretical result obtained through the theory developed in the text. The dotted line is the spectral density predicted by assuming that the components of the correlation matrix—which are orthogonal to the eigenspace associated with the largest eigenvalue—are described by statistical fluctuations and are described by Eq. (9). The inset shows the model spectrum (continuous lines) and the noise dressed (dashed line) spectral densities. The vertical line indicates the largest eigenvalue of the spectrum of the model.

where $\mathrm{Ei}(x)$ is the exponential integral function.

The largest eigenvalue in Eq. (14) is in this case

$$\lambda_1 = 1 + \frac{[-1-2\bar{\gamma}+2\left(-1+e^{1/\bar{\gamma}}\right)\bar{\gamma}^2](-1+N)}{-1+e^{1/\bar{\gamma}}}. \quad (17)$$

C. Multifactor models

The results obtained for the one-factor model can be extended to multifactor models. When the factors are stochastic and uncorrelated to each other, the structure of the correlation matrix is given by the composition of the groups of variables correspondent to the factors. To be more precise, we define a *group* of variables as Subset A of the N variables x_i that are influenced in their dynamics by a factor $f_j(t)$, i.e., $\gamma_i^{(j)} \neq 0$ for $x_i \in A$ and $\gamma_i^{(j)}=0$ otherwise. The existence of groups of variables does not in general imply that each variable is determined only by one factor, i.e., that each variable belongs only to one group. In fact, groups can partially overlap, or a group can be a subset of a larger group. In the following, we consider two classes of multifactor models, that can be relevant as a first approximation for several applications. In the first class (termed block model), each variable belongs to one and only one group, i.e., the groups are a partition of the set of variables. The spectral properties of this class of models has been previously studied in Ref. [27] in the context of financial markets. In the second class (termed the hierarchical model), there is a hierarchy of factors composed by H layers. More specifically, there is a common factor influencing all of the variables, then a second layer composed by a set of nonoverlapping groups that partition the set of variables, then a third layer in which each group of the second layer is partitioned in nonoverlapping groups and so on until the layer H. The noise dressing of the spectral density of hierarchical models has not been studied in the literature.

1. Block models

The simplest case is when each variable belongs to one and only one group, i.e., its dynamics is determined by only one factor and by the idiosyncratic noise. There are K groups each composed by $n_1, n_2, \ldots n_K$ variables such that $n_1+n_2+\ldots+n_K=N$. In this case, the correlation matrix of the model is block diagonal. The correlation coefficient between variables belonging to different groups is zero, while, when the variables i and j belong to the same group k, the correlation coefficient is $\gamma_i^{(k)}\gamma_j^{(k)}$. The spectral density of this kind of models is simply given by the superposition of the spectral densities of K one-factor models. For example, if the model is degenerate, i.e., for each group $\gamma_i^{(k)} \equiv \gamma^{(k)}$ does not depend on i, the spectrum is composed by K large eigenvalues $1-(n_j-1)(\gamma^{(j)})^2$ $(j=1,\ldots,K)$ and $N-K$ small eigenvalues $1-(\gamma^{(j)})^2$ each with degeneracy n_j-1 $(j=1,\ldots,K)$.

The noise dressing of this spectrum follows directly from Eq. (8) in which the number of distinct eigenvalues is $2K$. The equation for $G(z)$ is therefore an algebraic equation of degree $2K+1$.

When the model is nondegenerate and the probability distributions of $\gamma_i^{(k)}$ are given, one can solve the nondegenerate case by using the same arguments of the onefactor model. Clearly, the computational task increases with the number of factors.

2. Hierarchical models

An interesting generalization of multifactor models occurs when there is a hierarchical overlap between different groups described above. To give a concrete example, let us consider a portfolio of stocks. As a first approximation, we can consider the portfolio as composed of a large group following a common factor, e.g., the market factor in the Capital Asset Pricing Model, and a certain number of groups homogeneous in economic activity following a sectorial factor, such as, for example, the oil companies or the technological stocks. In this case, the composition of the groups induces a hierarchical structure to the correlation matrix.

We present here a simple example in which the N variables follow a common factor with a constant Γ. Moreover, the set of variables is divided in two groups. There are n_1 variables following the first subfactor with constant γ_1 and $n_2=N-n_1$ variables following the second subfactor with constant γ_2. Therefore, the equation of the model is

$$x_i(t) = \Gamma f(t) + \gamma_i^{(1)} f_1(t) + \gamma_i^{(2)} f_2(t) + \gamma_i^{(0)} \epsilon_i(t), \quad (18)$$

where $\gamma_i^{(1)}=\gamma_1$ for $i=1,\ldots,n_1$ and $\gamma_i^{(1)}=0$ for $i=n_1+1,\ldots,N$. Analogously $\gamma_i^{(2)}=0$ for $i=1,\ldots,n_1$ and $\gamma_i^{(2)}=\gamma_2$ for $i=n_1+1,\ldots,N$.

The correlation matrix of this model is a block matrix

$$\mathbf{C} = \begin{pmatrix} 1 & \cdots & \Gamma^2+\gamma_1^2 & \Gamma^2 & \cdots & \Gamma^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \Gamma^2+\gamma_1^2 & \cdots & 1 & \Gamma^2 & \cdots & \Gamma^2 \\ \Gamma^2 & \cdots & \Gamma^2 & 1 & \cdots & \Gamma^2+\gamma_2^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \Gamma^2 & \cdots & \Gamma^2 & \Gamma^2+\gamma_2^2 & \cdots & 1 \end{pmatrix}. \tag{19}$$

The spectrum of this matrix is composed by two large eigenvalues given by

$$\lambda_\pm = \frac{1}{2}(2+\gamma_1^2(n_1-1)+\gamma_2^2(n_2-1)+\Gamma^2(n_1+n_2-2) \pm \sqrt{A^2+\Gamma^4(n_1+n_2)^2+2A\Gamma^2(n_1-n_2)}), \tag{20}$$

where $A=(\gamma_1^2(n_1-1)-\gamma_2^2(n_2-1))$ and n_1-1 eigenvalues equal to $\lambda_{10}\equiv 1-\Gamma^2-\gamma_1^2$ and n_2-1 eigenvalues equal to $\lambda_{20}\equiv 1-\Gamma^2-\gamma_2^2$. The derivation of Eq. (20) and the general approach that can be used to calculate the spectrum of a more complicated hierarchical model are given in Appendix A.

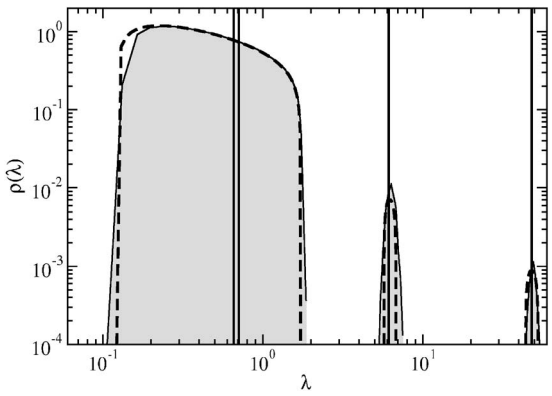

FIG. 4. Spectral density of a two-level hierarchical factor model. The parameters are $n_1=100$, $n_2=70$, $\Gamma=0.5$, $\gamma_1=0.2$, $\gamma_2=0.3$, and $T=500$. The gray area is based on the average over 1000 simulations and the dashed line is the theoretical result. The vertical lines indicate the position of the eigenvalues of the spectrum of the model.

Again by making use of Eq. (8), it is possible to find the effect of noise dressing by solving the corresponding fifth degree algebraic equation

$$\begin{aligned} &-T^5+G(z)(T^4z+T^3\lambda_{10}-n_1T^3\lambda_{10}+T^4\lambda_{10}+T^3\lambda_{20}-n_2T^3\lambda_{20}+T^4\lambda_{20}-T^3\lambda_-+T^4\lambda_--T^3\lambda_++T^4\lambda_+)+G^2(z)(-T^3z\lambda_{10}-T^3z\lambda_{20}\\ &-2T^2\lambda_{10}\lambda_{20}+n_1T^2\lambda_{10}\lambda_{20}+n_2T^2\lambda_{10}\lambda_{20}-T^3\lambda_{10}\lambda_{20}-T^3z\lambda_-+n_1T^2\lambda_{10}\lambda_--T^3\lambda_{10}\lambda_-+n_2T^2\lambda_{20}\lambda_--T^3\lambda_{20}\lambda_--T^3z\lambda_+\\ &+n_1T^2\lambda_{10}\lambda_+-T^3\lambda_{10}\lambda_++n_2T^2\lambda_{20}\lambda_+-T^3\lambda_{20}\lambda_++2T^2\lambda_-\lambda_+-T^3\lambda_-\lambda_+)+G^3(z)(T^2z\lambda_{10}\lambda_{20}+T^2z\lambda_{10}\lambda_-+T^2z\lambda_{20}\lambda_-+T\lambda_{10}\lambda_{20}\lambda_-\\ &-n_1T\lambda_{10}\lambda_{20}\lambda_--n_2T\lambda_{10}\lambda_{20}\lambda_-+T^2\lambda_{10}\lambda_{20}\lambda_-+T^2z\lambda_{10}\lambda_++T^2z\lambda_{20}\lambda_++T\lambda_{10}\lambda_{20}\lambda_+-n_1T\lambda_{10}\lambda_{20}\lambda_+-n_2T\lambda_{10}\lambda_{20}\lambda_+\\ &+T^2\lambda_{10}\lambda_{20}\lambda_++T^2z\lambda_-\lambda_+-T\lambda_{10}\lambda_-\lambda_+-n_1T\lambda_{10}\lambda_-\lambda_++T^2\lambda_{10}\lambda_-\lambda_+-T\lambda_{20}\lambda_-\lambda_+-n_2T\lambda_{20}\lambda_-\lambda_++T^2\lambda_{20}\lambda_-\lambda_+)\\ &+G^4(z)(-Tz\lambda_{10}\lambda_{20}\lambda_--Tz\lambda_{10}\lambda_{20}\lambda_+-Tz\lambda_{10}\lambda_-\lambda_+-Tz\lambda_{20}\lambda_-\lambda_++n_1\lambda_{10}\lambda_{20}\lambda_-\lambda_++n_2\lambda_{10}\lambda_{20}\lambda_-\lambda_+-T\lambda_{10}\lambda_{20}\lambda_-\lambda_+)\\ &+G^5(z)z\lambda_{10}\lambda_{20}\lambda_-\lambda_+=0. \end{aligned} \tag{21}$$

Figure 4 shows the comparison between numerical simulations and theoretical results for this model. The agreement between simulations and analytical calculations is quite good. The vertical lines indicate the position of the spectrum of the model which is equal to $\rho(\lambda)=(n_1-1)\delta(\lambda-\lambda_{10})+(n_2-1)\delta(\lambda-\lambda_{20})+\delta(\lambda-\lambda_+)+\delta(\lambda-\lambda_-)$. The parameters of the model are indicated in the caption. It is worth noting that the number of large eigenvalues of the two-layer hierarchical model described by Eq. (18) is two (see also Fig. 4). This is the same number of a block model with two blocks, even if the hierarchical is generated by three factors. This observation suggests that the difference between the two models can be observed in the structure of the eigenvectors, rather than from the difference in the eigenvalue spectrum.

To summarize the results presented in this section, we note that when the factors are uncorrelated in time one can find the spectrum of the model of the degenerate model by a block diagonalization of the correlation matrix. Furthermore, by solving the algebraic Eq. (8) one may obtain the noise dressed spectral density. When the model is non degenerate and the distribution of $\gamma_i^{(k)}$ is known, one can apply the above discussed strategy to obtain the spectral density numerically starting from their numerical solution.

V. FACTOR MODELS WITH SINUSOIDAL FACTOR

Our method is also applicable when the factors are correlated in time. To give a concrete example, let us consider a model where the dynamics of the variables are described by the equation

$$x_i(t) = \gamma\sqrt{2}\sin(\omega t+\phi_i)+\gamma^{(0)}\epsilon_i(t). \tag{22}$$

The model described in Eq. (22) is the first approximation of the dynamics of the level of expression of genes during cell cycle as detected in microarray experiments [9–11]. In this

case, the frequency ω is related to the duration of the cell cycle. Microarray experiments usually have a very small number of time records compared with the number of variables (genes). This fact leads to a heavy dressing of the correlation matrix by noise, and hence a careful characterization of the noise dressing might be even more important in this case.

The equal time correlation coefficient of this model is $C_{ij}=\gamma^2\cos(\phi_i-\phi_j)$ for $i\neq j$ and $C_{ii}=1$. Thus the correlation matrix is

$$\mathbf{C}=\begin{pmatrix} 1 & \cdots & \gamma^2\cos(\phi_1-\phi_j) & \cdots & \gamma^2\cos(\phi_1-\phi_N) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma^2\cos(\phi_1-\phi_j) & \cdots & 1 & \cdots & \gamma^2\cos(\phi_j-\phi_N) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma^2\cos(\phi_1-\phi_N) & \cdots & \gamma^2\cos(\phi_j-\phi_N) & \cdots & 1 \end{pmatrix}. \tag{23}$$

In Appendix B, we show that the distinct eigenvalues of $\mathbf{C}$ are three, specifically the spectrum is composed by $\lambda_0=1-\gamma^2$ with multiplicity $N-2$ and two large eigenvalues $\lambda_\pm=1-\gamma^2+\frac{\gamma^2}{2}(N\pm\sqrt{|g|})$, where $g\equiv\Sigma_{j=1}^{N}e^{2i\phi_j}$. The theory of noise dressing of this correlation matrix *cannot* be performed by following the lines we used in the previous sections for factor models which are time uncorrelated, i.e., by using Eq. (8). This is because variables x at different times are correlated, whereas Eq. (8) is obtained with the assumption that $\langle x_i(t)x_j(t')\rangle_{\text{ens}}\propto\delta_{tt'}$. A way to solve this problem is to assume that ϕ_i is a random phase distributed according to a uniform distribution in $[0,2\pi]$. Thus, the ensemble average consists in an averaging over the distribution of the phases and of the noise terms ϵ_i. In this case, the matrix $\mathbf{C}$ becomes the identity matrix and the ensemble average of the product of two variables in two distinct instants of time can be written as $\langle x_i(t)x_j(t')\rangle_{\text{ens}}=\bar{C}_{ij}D_{tt'}$, where $\bar{\mathbf{C}}=\mathbf{I}_N$ and

$$\mathbf{D}=\begin{pmatrix} 1 & \cdots & \gamma^2\cos[\omega(t_1-t_k)] & \cdots & \gamma^2\cos[\omega(t_1-t_T)] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma^2\cos[\omega(t_1-t_k)] & \cdots & 1 & \cdots & \gamma^2\cos[\omega(t_j-t_T)] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma^2\cos[\omega(t_1-t_T)] & \cdots & \gamma^2\cos[\omega(t_j-t_T)] & \cdots & 1 \end{pmatrix}. \tag{24}$$

The case of factorization of $\langle x_i(t)x_j(t')\rangle_{\text{ens}}$ in a variable and time component is treated in Ref. [23]. Specifically, the equation for the resolvent is in this case [cfr. Eqs. (21)–(23) of Ref. [23]]

$$G(z)=\sum_{i=1}^{T}\frac{1}{z-Q(z)d_i}, \tag{25}$$

where $Q(z)$ is the solution of the equations

$$Q(z)=\frac{1}{T}\sum_{j=1}^{N}\frac{c_j}{1-R(z)c_j},\quad R(z)=\frac{1}{T}\sum_{i=1}^{T}\frac{d_i}{z-d_iQ(z)}. \tag{26}$$

and c_j, $(j=1,\ldots,N)$ and d_i, $(i=1,\ldots,T)$ are the eigenvalues of $\bar{\mathbf{C}}$ and $\mathbf{D}$, respectively. Because of the averaging procedure, the eigenvalues of $\bar{\mathbf{C}}$ are $c_j=1$ and the first equation in Eq. (26) becomes

$$Q(z)=\frac{N}{T}\frac{1}{1-R(z)}. \tag{27}$$

In Appendix B, we show how to diagonalize the $\mathbf{D}$ matrix assuming that the sampling times $t_1,\ldots,t_T$ are equispaced, i.e., $t_j-t_{j-1}=\tau$. The spectrum of $\mathbf{D}$ consists of two large eigenvalues

$$d_{1,2}=1-\gamma^2+\frac{\gamma^2}{2}\left(T\pm\frac{\sin\omega T\tau}{\sin\omega\tau}\right), \tag{28}$$

and $T-2$ eigenvalues equal to $d_i=(1-\gamma^2)\equiv d_0$, where $i=3,\ldots,T$.

From Eq. (26), one thus obtains the equation for $Q(z)$, that is

$$\begin{aligned} &Q^4(z)d_0d_1d_2T+Q^3(z)(-d_0d_1d_2N+d_0d_1d_2T-d_0d_1Tz \\ &\quad-d_0d_2Tz-d_1d_2Tz)+Q^2(z)(d_0d_1z+d_0d_2z-2d_1d_2z \\ &\quad+d_0d_1Nz+d_0d_2Nz+d_1d_2Nz-d_0d_1Tz-d_0d_2Tz+d_0Tz^2 \\ &\quad+d_1Tz^2+d_2Tz^2)+Q(z)(-2d_0z^2+d_1z^2+d_2z^2-d_0Nz^2 \\ &\quad-d_1Nz^2-d_2Nz^2+d_0Tz^2-Tz^3)+Nz^3=0. \end{aligned} \tag{29}$$

From the solutions of this fourth-order algebraic equation, one can compute the resolvent $G(z)$ by using Eq. (25) Figure 5 shows the comparison between numerical simulations and theoretical results for this model. The parameters of the model are indicated in the caption. The obtained results al-

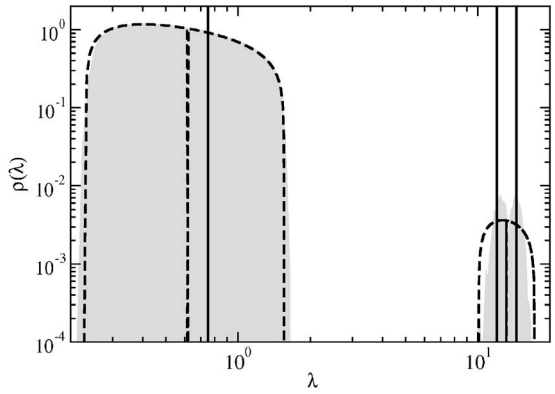

FIG. 5. Spectral density of the one-factor model in with a sinusoidal factor described by Eq. (22). The parameters are $N=100$, $T=500$, $\gamma=0.5$, $\omega=1$, $\tau=1$, and the phase ϕ is distributed uniformly between 0 and 2π. The dashed line is the analytical result based on the RMT.

low one to quantitatively characterize the noise dressing of the signal part of the spectrum of the factor model with a sinusoidal factor and random phases.

VI. CONCLUSIONS

In conclusion, we have shown that the application of RMT allows one to quantitatively solve the problem of the modeling of the spectral density of eigenvalues of the correlation matrix of a large class of factor models in the presence of the statistical uncertainty due to the finiteness of the number of records of time series. This class includes factor models with factors uncorrelated in time or sinusoidally time dependent factors with random phases. We have shown that a careful modeling of the effects of statistical uncertainty requires more than just the simple assumption that the components of the correlation matrix which are orthogonal to the eigenspace—associated with the largest eigenvalues—are fully controlled by statistical uncertainty. In fact, it turns out that the precise profile of both the low part of the spectral density and the computed value of the largest eigenvalues are related to the details of the considered factor model in a way that can be precisely quantified.

Our results can be applied to the modeling of several systems belonging to many different disciplines including physics, information theory and communication, economics, finance, molecular biology, and in general to any study in which factor models can constitute a good starting point for modeling the simultaneous dynamics of many variables.

ACKNOWLEDGMENTS

Authors acknowledge support from the research Project No. MIUR 449/97, "High frequency dynamics in financial markets;" Project No. MIUR-FIRB RBNE01CW3M, "Cellular self-organizing nets and chaotic nonlinear dynamics to model and control complex system;" and from the European Union STREP Project No. 012911, "Human behavior through dynamics of complex social networks: an interdisciplinary approach."

APPENDIX A

For the eigenvalue spectrum, we derive the expression of Eq. (20) the correlation matrix of a (two-layer) hierarchical factor model described in Eq. (18). In order to find the eigenvalues of the correlation matrix, we need to put equal to zero the determinant

$$\begin{vmatrix} 1-\lambda & \cdots & \Gamma^2+\gamma_1^2 & \Gamma^2 & \cdots & \Gamma^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \Gamma^2+\gamma_1^2 & \cdots & 1-\lambda & \Gamma^2 & \cdots & \Gamma^2 \\ \Gamma^2 & \cdots & \Gamma^2 & 1-\lambda & \cdots & \Gamma^2+\gamma_2^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \Gamma^2 & \cdots & \Gamma^2 & \Gamma^2+\gamma_2^2 & \cdots & 1-\lambda \end{vmatrix}. \tag{A1}$$

The matrix is block diagonal and the determinant can be computed by using the formula [19]

$$\begin{vmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{vmatrix} = |\mathbf{A_{22}}||\mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}}|. \tag{A2}$$

In order to compute $\mathbf{A_{22}^{-1}}$, we note that $\mathbf{A_{22}}$ is of the form $a\mathbf{I}_{n_2}+b\mathbf{J}_{n_2n_2}$, where $\mathbf{I}_{n_2}$ is the $n_2\times n_2$ identity matrix and $\mathbf{J}_{n_2n_2}$ is the $n_2\times n_2$ unit matrix (i.e. a matrix consisting of all ones). Moreover $a=1-\lambda-\Gamma^2-\gamma_2^2$ and $b=\Gamma^2+\gamma_2^2$. The matrices of this type have the properties

$$\det(a\mathbf{I}_n + b\mathbf{J}_{nn}) = a^{n-1}(a+bn), \tag{A3}$$

$$(a\mathbf{I}_n + b\mathbf{J}_{nn})^{-1} = \frac{1}{a}\left(\mathbf{I}_n - \frac{b}{a+bn}\mathbf{J}_{nn}\right), \tag{A4}$$

and they are closed under sum [i.e. $(a\mathbf{I}_n+b\mathbf{J}_{nn})+(a'\mathbf{I}_n+b'\mathbf{J}_{nn})=(c\mathbf{I}_n+d\mathbf{J}_{nn})$]. The off diagonal matrices in the determinant (A1) are $\mathbf{A}_{12}=\Gamma^2\mathbf{J}_{n_1n_2}$ and $\mathbf{A}_{21}=\Gamma^2\mathbf{J}_{n_2n_1}$, where $\mathbf{J}_{mn}$ is a $m\times n$ matrix consisting of all 1s. The $\mathbf{J}_{mn}$ matrices form a closed algebra under multiplication, for example $\mathbf{J}_{nm}\mathbf{J}_{mp}=m\mathbf{J}_{np}$. These properties allow to conclude that both $\mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}}$ and $\mathbf{A_{11}}-\mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}}$ are of the form $a\mathbf{I}_{n_1}+b\mathbf{J}_{n_1n_1}$. It is therefore possible to calculate explicitly the two determinants in the right side of Eq. (A2). The direct application of Eq. (A3) allows to find the determinant (A1) and, by solving the characteristic equation, to find the eigenvalues of Eq. (20).

APPENDIX B

In this appendix, we find the eigenvalues of the **C** and **D** matrices [Eqs. (23) and (24)] of the random phase model.

1. C matrix

The matrix $\mathbf{C}-\lambda\mathbf{I}_N$, whose determinant serves to find the eigenvalues of Eq. (23), can be rewritten as

$$(1-\lambda-\gamma^2)\mathbf{I}_N + \mathbf{u}\times\mathbf{v}^T + \mathbf{v}\times\mathbf{u}^T, \tag{B1}$$

where $\mathbf{u}^T=\gamma/\sqrt{2}(e^{i\phi_1},\ldots,e^{i\phi_N})$ and $\mathbf{v}$ is the complex conjugate of $\mathbf{u}$. We call $\mathbf{A}=(1-\lambda-\gamma^2)\mathbf{I}_N+\mathbf{u}\times\mathbf{v}^T$ and we make use

of the Sherman–Morrison formula for determinants

$$|\mathbf{A}+\mathbf{v}\times\mathbf{u}|^T=|\mathbf{A}|(1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}). \tag{B2}$$

Also, the determinant and the inverse of the matrix **A** can be computed by using the Sherman–Morrison formula for determinants Eq. (B2) and for inverse

$$(\mathbf{A}+\mathbf{u}\times\mathbf{v}^T)^{-1}=\mathbf{A}^{-1}-\frac{(\mathbf{A}^{-1}\mathbf{u})\times(\mathbf{v}\mathbf{A}^{-1})}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}. \tag{B3}$$

The direct application of these formulas gives

$$|\mathbf{A}|=(1-\lambda-\gamma^2)^n\left[1+\frac{\gamma^2 n}{2(1-\lambda-\gamma^2)}\right], \tag{B4}$$

$$\mathbf{A}^{-1}=\frac{\mathbf{I}_N}{1-\lambda-\gamma^2}-\frac{\frac{\mathbf{u}\times\mathbf{v}^T}{(1-\lambda-\gamma^2)^2}}{1+\frac{\gamma^2 n}{2(1-\lambda-\gamma^2)}}. \tag{B5}$$

By substituting these expression in Eq. (B2), we obtain for $|\mathbf{C}-\lambda\mathbf{I}_N|$ the expression

$$\frac{(1-\lambda-\gamma^2)^{N-2}}{4}([2(1-\lambda-\gamma^2)+\gamma^2 N]^2-\gamma^4|g|^2), \tag{B6}$$

where the complex function g is

$$g\equiv\sum_{j=1}^{N}e^{2i\phi_j}=\frac{2}{\gamma^2}\mathbf{u}^T\mathbf{u}. \tag{B7}$$

By putting Eq. (B6) equal to zero, we find the eigenvalues of **C** that is composed by two large eigenvalues

$$\lambda_\pm=1-\gamma^2+\frac{\gamma^2}{2}(N\pm|g|), \tag{B8}$$

and one eigenvalue $\lambda_0=1-\gamma^2$ with multiplicity $N-2$.

The remaining task is to calculate the parameter g of Eq. (B7). Its absolute value is given by

$$|g|=\sqrt{N+\sum_{i\neq j}e^{2i(\phi_j-\phi_i)}}. \tag{B9}$$

If the ϕ_j are uniformly distributed in $[0,2\pi]$, the second term in the square root vanishes so that $|g|=\sqrt{N}$ and

$$\lambda_\pm=1-\gamma^2+\frac{\gamma^2}{2}(N\pm\sqrt{N}). \tag{B10}$$

2. *D* matrix

The matrix **D** in Eq. (24) is a special case of the matrix **C** of Eq. (23) where $\phi_j=\omega t_j$, $(j=1,\ldots,T)$. Suppose that the sampling times $t_1,\ldots,t_T$ are equispaced, i.e., $t_j-t_{j-1}=\tau$. In this case, one can write $t_j=(j-1)\tau$ and the g parameter of Eq. (B7) is

$$g=\sum_{j=1}^{T}e^{2i\omega t_j}=\sum_{j=1}^{T}e^{2i\omega\tau(j-1)}=\frac{1-e^{2i\omega\tau T}}{1-e^{2i\omega\tau}}, \tag{B11}$$

and therefore

$$|g|^2=\frac{\sin^2\omega T\tau}{\sin^2\omega\tau}. \tag{B12}$$

From the proof in the previous subsection, we conclude that the spectrum of **D** consists of two large eigenvalues

$$d_{1,2}=1-\gamma^2+\frac{\gamma^2}{2}\left(T\pm\frac{\sin\omega T\tau}{\sin\omega\tau}\right), \tag{B13}$$

and $T-2$ eigenvalues equal to $d_i=(1-\gamma^2)\equiv d_0$, where $i=3,\ldots,T$.

[1] J. Wishart, Biometrika **A20**, 32 (1928).
[2] P. J. Forrester, and T. D. Hughes, J. Math. Phys. **35**, 6736 (1994).
[3] Y. Demasure, and R. A. Janik, Phys. Lett. B **553**, 105 (2003).
[4] A. L. Moustakas *et al.*, Science **287**, 287 (2000).
[5] J. Tworzydlo, and C. W. J. Beenakker, Phys. Rev. Lett. **89**, 043902 (2002).
[6] S. E. Skipetrov, Phys. Rev. E **67**, 036621 (2003).
[7] J. Kwapien, S. Drozdz, and A. A. Ioannides, Phys. Rev. E **62**, 5557 (2000).
[8] P. Seba, Phys. Rev. Lett. **91**, 198104 (2003).
[9] N. S. Holter *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 8409 (2000).
[10] O. Alter, P. O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. U.S.A. **97**, 10101 (2000).
[11] N. S. Holter *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 1693 (2001).
[12] L. Laloux *et al.*, Phys. Rev. Lett. **83**, 1467 (1999).
[13] V. Plerou *et al.*, Phys. Rev. Lett. **83**, 1471 (1999).
[14] S. Maslov, and Y.-C. Zhang, Phys. Rev. Lett. **87**, 248701 (2001).
[15] S. Pafka, and I. Kondor, Eur. Phys. J. B **27**, 277 (2002).
[16] J. Kwapien *et al.*, Physica A **309**, 171 (2002).
[17] Y. Malevergne, and D. Sornette, Physica A **331**, 660 (2004).
[18] Z. Burda *et al.*, Physica A **343**, 295 (2004).
[19] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Academic, San Diego, 1979).
[20] M. R. Anderberg, *Cluster Analysis for Applications* (Academic, New York, 1973).
[21] Y. J. Campbell, A. W. Lo, and A. C. Mackinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, New Jersey, 1997).
[22] M. Metha, *Random Matrices* (Academic, New York, 1995).
[23] A. M. Sengupta, and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).
[24] L. Denby, and C. L. Mallows, *Computing Sciences and Statistics: Proceedings of the 23rd Symposium on the Interface*, edited by E. M. Keramidas, (Interface Foundation, Fairfax Station, VA, 1991), pp.54–57.
[25] A. Soshnikov, J. Stat. Phys. **108**, 1033 (2002).
[26] W. H. Press *et al.*, *Numerical Recipes* (Cambridge University Press, Cambridge, U.K, 1992).
[27] J. D. Noh, Phys. Rev. E **61**, 5981 (2000).