Data clustering and noise undressing of correlation matrices

Lorenzo Giada and Matteo Marsili
Istituto Nazionale per la Fisica della Materia (INFM), Trieste Unit, Trieste I-34014, Italy
and International School for Advanced Studies (SISSA), V. Beirut 2-4, Trieste I-34014, Italy
(Received 17 January 2001; published 15 May 2001)

We discuss an approach to data clustering. We find that maximum likelihood leads naturally to an Hamiltonian of Potts variables that depends on the correlation matrix and whose low temperature behavior describes the correlation structure of the data. For random, uncorrelated data sets no correlation structure emerges. On the other hand, for data sets with a built-in cluster structure, the method is able to detect and recover efficiently that structure. Finally we apply the method to financial time series, where the low-temperature behavior reveals a nontrivial clustering.

DOI: 10.1103/PhysRevE.63.061101 PACS number(s): 02.50.Le, 05.40.−a, 64.60.Ak, 89.90.+n

I. INTRODUCTION

Statistical mechanics typically addresses the question of how structures and order arising from interactions in extended systems are dressed, and eventually destroyed, by stochastic—so-called thermal—fluctuations. The inverse problem, unraveling the structure of correlations from stochastic fluctuations in large data sets, has only recently been addressed using ideas of statistical mechanics [1,2]. This is the case of data clustering problems, where the goal is to classify N objects, defined by D dimensional vectors $\{\vec{\xi}_i\}_{i=1}^N$, in equivalence classes. The general idea [1] consists in postulating a cost function that measures how a possible data structure compares with the sample one is studying. This cost function can be considered as an Hamiltonian whose low-energy states correspond to the cluster structures that are most compatible with the data sample. Structures are identified by configurations $\mathcal{S}=\{s_i\}_{i=1}^N$ of class indices, where s_i is the equivalence class to which object i belongs. Regarding s_i as Potts spins, a Potts Hamiltonian $H_q = -\Sigma_{i<j} J_{i,j}\delta_{s_i,s_j}$ has been recently proposed [2] as a cost function, with couplings $J_{i,j}$ decreasing with the distance $d_{i,j}=||\vec{\xi}_i-\vec{\xi}_j||$ between objects i and j. The underlying structure of data sets emerges as the clustering of Potts variables at low temperatures.

In the present work we address the question of data clustering. Rather than postulating the form of the Hamiltonian, we start from a statistical ansatz and invoke maximum likelihood and maximum entropy principles. In this way, the structure of the Hamiltonian arises naturally from the statistical ansatz, without the need of assumptions on its form. The method is particularly suited to study high-dimensional data sets, where each object is characterized by a large number $D \gg 1$ of properties. Time series are an ideal example of high-dimensional objects. The study of the structure of correlations between time series is therefore a crucial benchmark for our method.

First we derive the form of the Hamiltonian in the general case. Then we study the thermal and the ground-state properties of this Hamiltonian by Monte Carlo methods, in three different cases: (1) a synthetic uncorrelated data set, (2) a synthetic data set with a known correlation structure, and finally (3) a data set composed of financial time series with unknown correlations. We find that (1) for random uncorrelated time series no persistent structure emerges at low temperatures; (2) if the time series are generated with some cluster structure $\mathcal{S}^\star$, we find a phase transition to a low-temperature phase that is dominated by cluster configurations close to $\mathcal{S}^\star$. Hence the method does not introduce spurious correlations and is able to recover known correlation structures. The nature of the transition is investigated by a simple mean-field calculation in a simple case. This reveals that the phase transition is of first order.

The financial time series that we will study consists of the returns of the assets composing the Standard and Poor's 500 (S&P 500) index, whose correlations have been the subject of much recent interest [3–5]. On one side it has been observed [3] that the S&P 500 correlation matrix is affected by considerable noise dressing. Indeed its spectral properties are close to those of random, uncorrelated time series. On the other hand, these same correlation matrices have revealed a nontrivial structure of correlations when analyzed by minimal spanning tree methods [4] and by the method [5] of Ref. [2]. These apparently contradictory results raise the issue of disentangling in a systematic way the effects of fluctuations from real correlations in a large but finite data set. This is the main issue we shall focus here.

Quite interestingly, our analysis of the S&P 500 data set reveals a low-temperature behavior dominated by few clusters of correlated assets with scale-invariant properties. We shall not enter into the details of the economic meaning of our findings, which shall be discussed elsewhere [6]. Our aim is rather to address the problem of revealing the structure of *bare* correlations hidden in a finite data set. We show that a thermal average over the relevant cluster structures provides a good fit of the financial correlations, which allows us to estimate the *noise-undressed* correlation matrix. Finally, we discuss several generalizations of our approach to generic data clustering.

II. METHOD

Let the data set $\Xi=\{\vec{\xi}_i\}_{i=1}^N$ be composed of N sets $\vec{\xi}_i = \{\xi_i(d)\}_{d=1}^D$ of D measurements. These are normalized to zero mean $\Sigma_d \xi_i(d)/D=0$ and unit variance $\Sigma_d \xi_i^2(d)/D=1$.

For example, in our application below $\xi_i(d)$ is the normalized daily returns of asset i of the S&P 500 index, in day d. The data set can also refer to a set of N objects that are characterized by D-measured quantities. In this case $\xi_i(d)$ is the ''normalized'' value of property d for object i. We assume that $\xi_i(d)$ are Gaussian variables. The reason is that we want to focus exclusively on pairwise correlations and the Gaussian model is the only one that is completely specified at this level. We shall discuss later how deviations from Gaussian statistics can be accounted for. The key quantity of interest is the matrix

$$C_{i,j}(D) \equiv \frac{1}{D} \sum_{d=1}^{D} \xi_i(d) \xi_j(d). \tag{1}$$

In order to investigate the structure of correlations, let us assume that $\xi_i(d)$ were generated by the equation

$$\xi_i(d) = \frac{\sqrt{g_{s_i}} \eta_{s_i}(d) + \epsilon_i(d)}{\sqrt{1+g_{s_i}}}. \tag{2}$$

Here $g_s > 0$ and s_i are integer variables (so-called Potts spins), $\eta_s(d)$ and $\epsilon_i(d)$ are iid Gaussian variables with zero average and unit variance.

The ansatz of Eq. (2) was proposed by Noh [7] to explain the spectral properties found in Ref. [3]. The idea behind it is that each set i belongs to one cluster s_i and that sets i and j in the same cluster ($s_i = s_j = s$) are correlated [$C_{i,j} \approx g_s/(1+g_s)$] whereas sets in different clusters ($s_i \neq s_j$) are independent. The s^{th} cluster is composed of n_s sets with internal correlation c_s, where

$$n_s = \sum_{i=1}^{N} \delta_{s_i,s}, \quad c_s = \sum_{i,j=1}^{N} C_{i,j} \delta_{s_i,s} \delta_{s_j,s}. \tag{3}$$

In order to allow for totally uncorrelated sets, we allow s_i to take all integer values up to N. Hence $\mathcal{S} = \{s_i\}_{i=1}^{N}$ describes the structure of correlations whereas the parameters $\mathcal{G} \equiv \{g_s\}_{s=1}^{N}$ tune the strength of these correlations.

Note that this ansatz is different from the explicative factor model used in financial applications [8], which is discussed in the Appendix.

The correlation matrix generated by Eq. (2) for $D \to \infty$ is

$$C_{i,j} = \frac{g_{s_i} \delta_{s_i,s_j} + \delta_{i,j}}{1+g_{s_i}}. \tag{4}$$

Its distribution of eigenvalues is simple: To each s with $n_s \geqslant 1$ there correspond one eigenvalue

$$\lambda_{s,0} = \frac{1+g_s n_s}{1+g_s}$$

and $n_s - 1$ eigenvalues $\lambda_{s,1} = 1/(1+g_s)$. Hence, large eigenvalues correspond to groups of many ($n_s \gg 1$) sets. For D finite, we expect noise to lift degeneracies between $\lambda_{s,1}$ but to leave the structure of large eigenvalues unchanged.

In order to fit the data set Ξ with Eq. (2), let us compute the likelihood. This is the probability $P(\Xi|\mathcal{S},\mathcal{G})$ of observing the data Ξ as a realization of Eq. (2) with structure $\mathcal{S}$ and parameters $\mathcal{G} = \{g_s\}_{s=1}^{N}$, and it reads

$$P(\Xi|\mathcal{S},\mathcal{G}) = \prod_{d=1}^{D} \left\langle \prod_{i=1}^{N} \delta\left(\xi_i(d) - \frac{\sqrt{g_{s_i}} \eta_{s_i}(d) + \epsilon_i(d)}{\sqrt{1+g_{s_i}}} \right) \right\rangle,$$

where the average is over all the η's and ϵ's variables and $\delta(x)$ is Dirac's delta function. Gaussian integration and elementary algebra leads to

$$P(\Xi|\mathcal{S},\mathcal{G}) \propto e^{-DH\{\mathcal{S},\mathcal{G}\}} \tag{5}$$

$$H\{\mathcal{S},\mathcal{G}\} = \frac{1}{2} \sum_{s} \left[(1+g_s)\left(n_s - \frac{g_s c_s}{1+g_s n_s} \right) - n_s \ln(1+g_s) + \ln(1+g_s n_s) \right]. \tag{6}$$

For any given structure $\mathcal{S}$ and $D \gg 1$, the likelihood $P(\Xi|\mathcal{S},\mathcal{G})$ is maximal for $g_s = \hat{g}_s$, where

$$\hat{g}_s = \frac{c_s - n_s}{n_s^2 - c_s} \tag{7}$$

for $n_s > 1$ and $\hat{g}_s = 0$ for $n_s \leqslant 1$. Inverting Eq. (7) gives $c_s = (\hat{g}_s n_s^2 + n_s)/(\hat{g}_s + 1)$ that is exactly what one would get combining Eqs. (3) and (4). Hence the maximum likelihood estimators $\hat{g}_s$ are consistent with our ansatz (2).

Note that for uncorrelated sets $C_{i,j} = \delta_{i,j}$, we have $c_s = n_s$ for each s and hence $\hat{g}_s = 0$. The coupling strength $\hat{g}_s$ instead diverges for totally correlated sets ($C_{i,j} = 1$) because $c_s = n_s^2$.

An expansion to second order in $g_s - \hat{g}_s$ of Eq. (6) shows that the likelihood quickly vanishes for $|g_s - \hat{g}_s| \gg 1/\sqrt{D}$. Hence, for $D \gg 1$, we can simplify things considerably by setting $g_s = \hat{g}_s$ in Eq. (6). The likelihood of structure $\mathcal{S}$ under ansatz (2) then takes the form $P(\Xi|\mathcal{S}) \propto e^{-DH_c}$, where

$$H_c\{\mathcal{S}\} = \frac{1}{2} \sum_{s:n_s>0} \left[\ln \frac{c_s}{n_s} + (n_s - 1) \ln \frac{n_s^2 - c_s}{n_s^2 - n_s} \right]. \tag{8}$$

The ground state $\mathcal{S}_0$ of H_c yields the maximum likelihood fit with Eq. (2). This would probably take the ansatz (2) too seriously. In general, it is preferable to consider probabilistic solutions $P\{\mathcal{S}\}$ and, following Ref. [1], we invoke the maximum entropy principle: Among all distributions $P\{\mathcal{S}\}$ with the same average log likelihood, we select that one which has maximal entropy. This, as usual, leads to the Gibbs distribution $P\{\mathcal{S}\} \propto e^{-\beta H_c\{\mathcal{S}\}}$ where the inverse temperature β arises as a Lagrange multiplier.

The Hamiltonian H_c depends implicitly on the Potts spins s_i through the cluster variables n_s and c_s of Eq. (3). Unlike the Potts Hamiltonian H_q, the dependence on δ_{s_i,s_j} is nonlinear and it is modulated by $C_{i,j}$. For $s_i \neq s_j$ for all $i \neq j$ we

have $n_s = c_s = 1$ for all s and hence $H_c = 0$. This state is representative of the high-temperature ($\beta \to 0$) limit. The low-temperature physics of H_c is instead nontrivially related to the correlation matrix $C_{i,j}$. Note, that the ferromagnetic state $s_i = 1$ for each i, which dominates as $\beta \to \infty$ in clustering methods based on Potts models [2], is in general not the ground state of H_c. Intuitively we expect that, if the model of Eq. (2) is reasonable, H_c should have a well-defined ground state and low-temperature phase that is energetically dominated by this state. In these cases, as in Ref. [2], we expect a thermal phase transition [9].

The form of the Hamiltonian H_c clearly depends on the ansatz (2). For example if one takes a factor model for the correlations one finds a different Hamiltonian that depends on different variables, as discussed in the Appendix. Also note that the present model only describes positive correlations. It is straightforward to generalize this approach to the case where a sizeable number of matrix elements $C_{i,j}$ are negative and not small. The idea is to introduce spin variables $\sigma_i = \pm 1$ for each set and modify Eq. (2) by multiplying the right-hand side by σ_i. This leads us to the analysis of a system where the Potts variables s_i and the spin variables σ_i interact. An account of this method shall be given elsewhere [6].

III. DATA

We consider three different data sets, i.e., three different matrices $C_{i,j}$. For all of them we took $N = 443$ time series of length $D = 1599$. The results with shorter time series will also be discussed below.

The first data set refers to N totally uncorrelated time series of length D. This is obtained, for example, from Eq. (2) with $s_i = i$. The second also is obtained from Eq. (2), but this time with preassigned structure $\mathcal{S}^\star$ and coupling strengths $\mathcal{G}^\star$. We shall discuss below how the specific structure was chosen. These two data sets help us to understand how the method performs when no structure is present at all and to check whether a predefined structure can be recovered.

Our third data set is made of financial time series of asset prices relative to the S&P 500. More precisely $\xi_i(d)$ is the normalized daily returns of asset i of the S&P 500, in day d; this is defined as

$$\xi_i(d) = \frac{\ln[p_i(d)/p_i(d-1)] - r_i}{\sigma_i}, \tag{9}$$

where $p_i(d)$ is the price of asset i in day d. The parameters r_i and σ_i are determined in order to have zero mean and unit variance for all i.

Correlation matrices of financial time series are of great practical interest. Indeed they are at the basis of risk minimization in the modern portfolio theory [8]. This states that, in order to reduce risk, the investment needs to be diversified (i.e., divided) on many uncorrelated assets, so that price fluctuations are averaged out. However the measure of correlation in samples with a number of observation times comparable to the number of assets was recently found to be affected by considerable noise dressing [3]. For example, the S&P 500 is composed of $N = 500$ assets and considering daily data from July 3, 1989 to October 27, 1995, one has $D = 1600$ data points for each asset. This data set is then an ideal instance of a problem where our method applies.

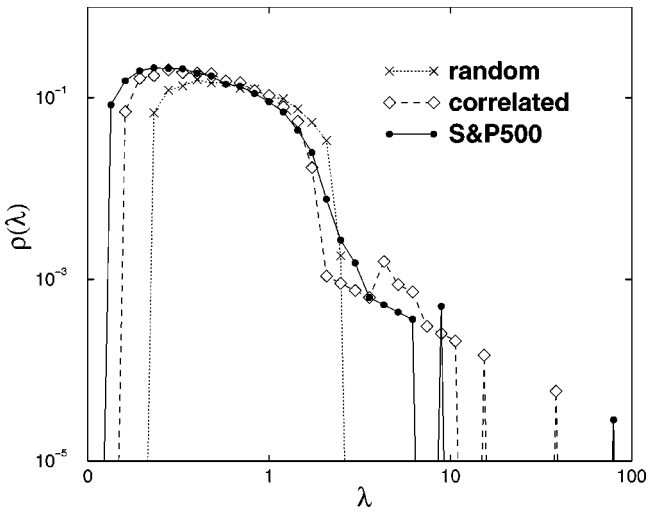

FIG. 1. Distribution of eigenvalues of the correlation matrices of S&P 500 (solid line ●), random (dotted ×), and synthetic correlated (dashed ◇) time series.

In addition, this data set has been studied by other authors with several methods including spectral analysis [3], minimal spanning tree [4] and super-paramagnetic clustering [5]. This allows us to compare the results of our method with those of other methods. Finally, in order to better appreciate the performance of our method in a real application, we choose the synthetic correlated data set $\{\mathcal{S}^\star, \mathcal{G}^\star\}$ as a ''large likelihood'' structure of the S&P 500 data set. In other words we performed a simulated annealing experiment on the S&P 500 data set, where the fictitious temperature $1/\beta$ was gradually decreased to 0. This allowed us to compare how well the real S&P 500 data set can be described by a maximum likelihood structure.[1]

Figure 1 shows a comparison of the spectral properties of the three correlation matrices. The spectrum of eigenvalues for uncorrelated time series are known exactly [10]. It extends over an interval of size $\sim N/D$ around $\lambda = 1$, as shown in Fig. 1. The spectrum of eigenvalues of the S&P 500 correlation matrix has a similar shape for $\lambda \approx 1$. This suggests that significant noise dressing due to finite D occurs [3]. The tail of the distribution ($\lambda \gg 1$) implies that some correlation is however present. Within our framework, large eigenvalues are associated with large clusters. Indeed the synthetic correlated data set has a broad distribution of cluster sizes (see Fig. 3) and a correspondingly fat tail in the distribution of eigenvalues.

[1]The maximum likelihood structure may be computationally hard to find and simulated annealing may get trapped into a local minimum. Indeed in our case we found slightly different structures in different experiments. The structure $\mathcal{S}^\star$ was that with largest likelihood.

IV. CLUSTERING BY MONTE CARLO SIMULATIONS

In order to study the properties of H_c we resort to Monte Carlo (MC) method with Metropolis algorithm [11]. This, at equilibration, allows us to sample the Gibbs distribution $P\{\mathcal{S}\}$ and compute average quantities, such as the internal energy $E_\beta = \langle H_c \rangle_\beta$ where $\langle \ldots \rangle_\beta$ stands for thermal average. To detect the occurrence of spontaneous magnetization—which correspond to the s_i remaining locked into energetically favorable configurations at low temperature—we measure the autocorrelation function

$$\chi(t,\tau) = \frac{\sum_{i<j} \delta_{s_i(t),s_j(t)} \delta_{s_i(t+\tau),s_j(t+\tau)}}{\sum_{i<j} \delta_{s_i(t),s_j(t)}}. \qquad (10)$$

This quantity tells us how many pairs of sites belonging to the same cluster at time t are still found in the same cluster after τ MC steps. For t large enough, χ becomes a function of τ only. This function decreases rapidly to a plateau value $\chi_\beta = \langle \chi(t,\tau) \rangle_\beta$ for $t \gg \tau \gg 1$. Clearly $\chi_\beta \simeq 0$ implies that no persistent structure is present whereas, at the other extreme, $\chi_\beta = 1$ implies that all sites are locked in a persistent structure of clusters.

We monitored these quantities for the three data sets. Let us start with a truly uncorrelated time series. We generate the time series $\xi_i(d)$ and then compute $C_{i,j}$ by Eq. (1). With this we compute the Hamiltonian H_c and study its thermal properties by the MC method. We do not expect any clustering to emerge in this case. Indeed, the internal energy E_β stays very close to 0 [see Fig. 2(a)] for all values of β investigated up to $\beta = 512$. Correspondingly no persistent cluster arises, i.e., $\chi_\beta \simeq 0$.

The results change turning to correlated data. Let us first discuss the S&P 500 data (for $D = 1599$). As Fig. 2(a) shows, for $\beta \approx 20$ the energy E_β starts deviating significantly from zero. For $\beta > 20$ persistent clusters are present: χ_β rapidly raises from zero and it has a maximum at $\beta \approx 40$ [see Fig. 2(b)]. The energy fluctuations reported in the inset shows a broad peak of intensity marking the onset of an ordered low-temperature phase. As β increases the dynamics is significantly slowed down. At $\beta \approx 200$ the energy reaches a minimal value $E_\beta \simeq -0.11N$ and does not decrease significantly increasing β at least up to $\beta = 4095$. This energy is smaller than that of the ferromagnetic state ($E_f = -0.086N$), with all sets in the same cluster. The system in this range of temperatures visits only few configurations.

The statistical properties of cluster configurations, as β varies, are shown in Fig. 3. For small β only small clusters survive to thermal fluctuations. As β increases a distribution of cluster sizes develops. At low temperatures the rank order plot of n_s reveals a broad distribution of clusters with the largest aggregating more than 190 sets. By a power-law fit of this distribution, we find that the number of clusters with more than n sets decays as $n^{-0.83}$. The scatter plot of c_s versus n_s also reveals a nontrivial power-law dependence $c_s \sim n_s^{1.66}$. This gives a statistical characterization of the

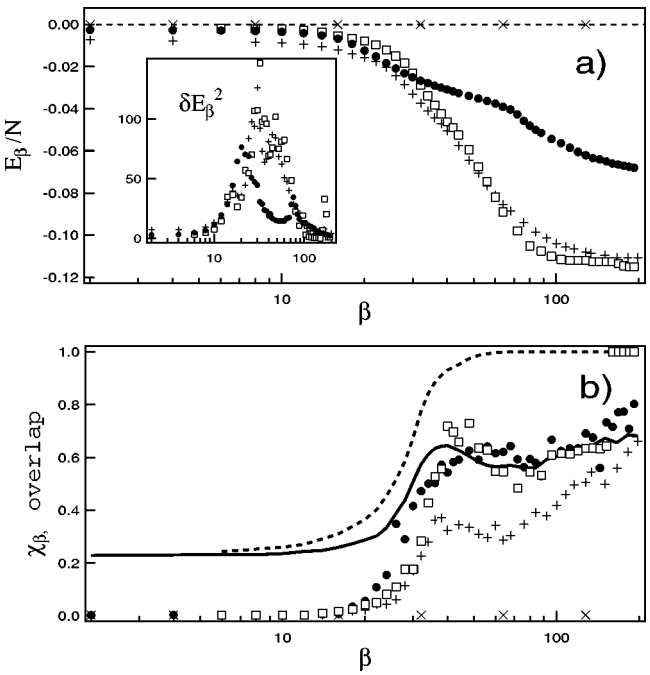

FIG. 2. (a) Energy E_β as a function of β for random ($\times$), S&P 500 ($+$), and correlated ($\square$) data sets of length $D = 1599$, respectively. The results for the S&P 500 data set over the last $D = 400$ days are also shown ($\bullet$). Inset: square energy fluctuation δE_β^2 vs β for the same data sets (same symbols). (b) Autocorrelation χ_β as a function of β for the same data sets (same symbols). The solid (dashed) line refers to the overlap with the configuration $s^\star$ for the S&P 500 (correlated) data set with $D = 1599$.

dominant configurations of clusters at low energy. The cluster structure we obtain is reasonable from the economic point of view: companies in the same economic sector belong to the same cluster. These issues will be discussed in detail elsewhere [6]. Here we restrict our attention to the clustering method.

It is instructive to compare these results with those obtained form shorter time series. We performed a second set

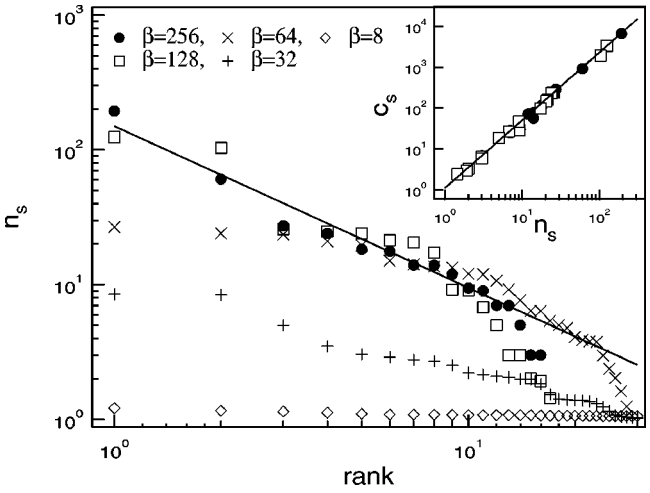

FIG. 3. Rank plot of n_s for several values of β. The line corresponds to $n \sim \text{rank}^{-1.2}$. Inset: c_s versus n_s for $\beta = 256$ ($\bullet$) and β ($\square$). The line corresponds to $c \sim n^{1.66}$.

of simulations with $C_{i,j}$ computed using the quotes of the S&P 500 assets for the last $D=400$ days. We find two inverse temperatures $\beta_1 \approx 20$ and $\beta_2 \approx 80$ that separates three regimes. This is signaled by the bending in the E_β curve and by peaks in the δE_β^2 versus β plot. At the first temperature clusters start to appear. For $\beta < \beta_2$ the largest cluster groups less than 30 sets and for $\beta > \beta_2$ larger clusters $n_s \approx 100$ appear. This hints at a time dependence of correlations, which are averaged in the $D=1599$ data set. For even shorter time series we found that sampling errors, acting like a temperature, destroy large clusters and only relatively small clusters ($n_s < 40$ for $D=60$) were found.

Finally let us discuss the results for the synthetic correlated data set (for $D=1599$). As already mentioned, the structure $\mathcal{S}^\star$ is a typical low-energy configuration for the S&P 500 data set extracted from the previous simulations (with $D=1599$). The parameters $g_s^\star$ where deduced from the n_s and c_s of this configuration, via Eq. (7). We recall that this data set is useful for two reasons: first it allows one to understand to what extent a structure of correlation put by hand with the form dictated by Eq. (2) can be correctly recovered. Secondly it allows us to compare the results found for the S&P 500 data with those of a time series with correlations described by Eq. (2).

For $\beta < 150$, the behaviors of E_β, δE_β^2, and χ_β are similar to those found for the S&P 500 data (see Fig. 2). A second, sharp peak in δE_β^2 at $\beta \approx 170$ signals a new clustering transition. Below this temperature, as shown by the plot of χ_β [Fig. 2(b)], the MC dynamics freezes into the original structure $\mathcal{S}^\star$. The overlap with the configuration $\mathcal{S}^\star$, defined as in Eq. (10) as the fraction of ''bonds'' $s_i = s_j$ for which $s_i^\star = s_j^\star$, quickly converges to 1 [see Fig. 2(b)] for the synthetic time series, whereas it remains around 60% for the S&P 500 data set. This, on one hand means that the original structure $\mathcal{S}^\star$ can be recovered quite efficiently. On the other hand, it suggests that several cluster configurations compete at low temperatures in the S&P 500 data set.

V. MEAN-FIELD MODEL

In this section we would like to determine the nature of the clustering transition that takes place in our system. We apply our method to an unrealistically simple situation that will allow us to extract analytical expressions for the order parameter associated with the phase transition. Our analysis is rather similar to that in Ref. [2].

We take N time series that belong to M clusters of the same size $n=N/M$. Let $\overline{s_i}$ be the cluster index of the ith time series, and let

$$C_{i,j} = \gamma \delta_{\overline{s_i},\overline{s_j}} + (1-\gamma)\delta_{i,j} \tag{11}$$

be the correlation matrix for $D=\infty$. This means that time series with $\overline{s_i} = \overline{s_j}$ have correlation γ whereas $\overline{s_i} \neq \overline{s_j}$ have $C_{i,j}=0$. The matrix $C_{i,j}$ has N/M blocks of size M along the diagonal and is zero elsewhere. A finite D sample of this problem is generated with Eq. (2) with $g_s = \gamma/(1-\gamma)$ and $s_i = \overline{s_i}$.

To fix ideas we can imagine to have the problem of putting N balls of M different colors in M boxes. Colors represent the original structure $\overline{s_i}$ whereas boxes represent the actual clustering configuration s_i. When the balls contained in each box have the same color, the original cluster structure has been recovered. Let m_s^c be the number of balls of color c in box (or cluster) s. Now $\Sigma_s m_s^c = n = N/M$ is the overall number of balls of color c, assumed equal for all colors, and $\Sigma_c m_s^c = n_s$ is the number of balls in box s, as in Eq. (3). With the above choice of parameters the internal correlation of box s for a given configuration $\{m_s^c\}$ of clusters is

$$c_s = (1-\gamma) n_s + \gamma \sum_c m_s^{c2}. \tag{12}$$

To compute the free energy $F=U-TS$ of the system we use the energy H_c as in Eq. (8) and we estimate the configuration entropy in the following way: The number of ways in which one can distribute the balls of color c by putting m_s^c of them in box s is

$$\frac{\left(\sum_s m_s^c\right)!}{\prod_s (m_s^c!)} = \frac{n!}{\prod_s (m_s^c!)}, \tag{13}$$

and the total number of configurations for all the colors is the product over c of this expression. Taking the logarithm of this product we obtain the configuration entropy

$$S = \ln\left(\prod_c \frac{\left(\sum_s m_s^c\right)!}{\prod_s (m_s^c!)} \right) = \sum_{s,c} m_s^c \ln\left(\frac{m_s^c}{n}\right), \tag{14}$$

where we have approximated in the usual way the logarithm of the factorial. Finally the free energy is

$$F = \sum_s F_s$$

$$F_s = \frac{1}{2}\left[\ln\frac{c_s}{n} + (n-1)\ln\left(\frac{n^2-c_s}{n^2-n}\right)\right] + T\sum_c m_s^c \ln\left(\frac{m_s^c}{n}\right). \tag{15}$$

After substitution of Eq. (12) we find an expression that depends on the occupation variables m_s^c. The occupations in different boxes are related by the overall constraints $\Sigma_s m_s^c = N/M$. We take the mean-field approximation, which is legitimate in cases like this, where we neglect these effects. In other words, we minimize each of the F_s independently and we omit therefore the subscript s from now on.

We can then focus on just one box and look for solutions of the form

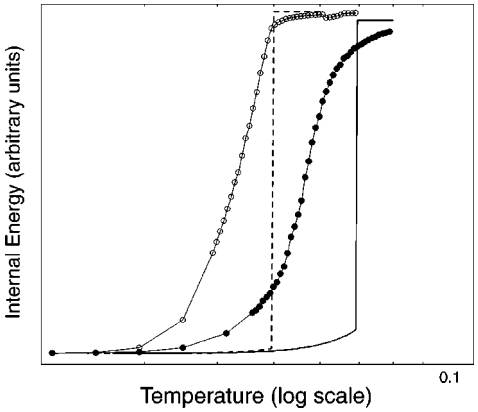

FIG. 4. Ferromagnetic clustering: filled dots and full lines correspond to Monte Carlo and analytical results for a system with $N=150$, $M=6$; empty dots and dashed lines to $N=2400$, $M=24$. γ is always 0.3.

$$m^c=\frac{N}{M}\begin{cases}\phi, & c=1\\ \dfrac{1-\phi}{M-1}, & c\neq 1,\end{cases} \tag{16}$$

with $0\leqslant\phi\leqslant 1$. In this ansatz, the balls of color $c=1$ are more (or less) numerous that those of other colors $c>1$. In the spirit of mean-field approximation, we neglect the possibility that the number of balls of colors $c>1$ may be unevenly distributed. Hence the parameter ϕ plays the role of the order parameter.

The paramagnetic solution $\phi=1/M$, which corresponds to an uniform distribution of colors, is always a solution of the saddle point equations $\partial F/\partial\phi=0$. This state is expected to be stable (the minimum of F) at high temperature. A second solution of $\partial F/\partial\phi=0$, which corresponds to the clustered ''ferromagnetic'' state, appears at intermediate temperatures with $\phi\approx 1$. For $T=T_c$ the values of the free energy corresponding to the two states are equal and a first-order phase transition to a ferromagnetic state takes place. The order of the transition is independent of the values of the parameters, while the critical temperature is determined mainly by the number of time series N.

We checked that this result is compatible with that obtained from Monte Carlo simulations. We find that the mean-field approach provides a good qualitative picture of the transition and a reliable estimate of the critical temperature at which it takes place (see Fig. 4).

VI. NOISE UNDRESSING

Equation (2) with a single cluster configuration ($\beta\to\infty$), is inadequate to capture the full complexity of the correlations in the S&P 500 data set. Probabilistic clustering, where several cluster structures $\mathcal{S}$ are allowed with their Gibbs probability $P\{\mathcal{S}\}$, provides an alternative approximation. In this approach the parameter β can be tuned to determine the optimal spread in configuration space, which best describes the correlation structure built in the original data set. This line of reasoning will lead us to a method to ''fit'' the correlation structure of a data set with a single parameter β. This will finally allow us to undress the correlation matrix $C_{i,j}(D)$ of its noise dressing and to reveal the bare correlations.

Let us start by remarking that the problem with Eq. (2) is that it stipulates that a set i can belong to only one cluster. This suggests to consider the generalized model

$$\xi_i(d)=\frac{\sum_s \sqrt{g_{s,i}}\,\eta_s(d)+\epsilon_i(d)}{\sqrt{1+\sum_s g_{s,i}}}, \tag{17}$$

where each set i can belong to any cluster s. Equation (17) has, on the other side, the disadvantage that it depends on too many variables, and it leads to overfitting stochastic fluctuations.

The finite temperature distribution $P\{\mathcal{S}\}$ provides a natural way out of this situation. Indeed at finite β each set i visits different clusters s and we can define

$$g_{s,i}(\beta)=\left\langle \frac{c_s-n_s}{n_s^2-c_s}\,\delta_{s,s_i}\right\rangle_\beta . \tag{18}$$

The parameters $g_{s,i}(\beta)$ can be measured in a MC simulation and provide us with a measure of the strength of the correlation between set i and cluster s.

These parameters and Eq. (17) also allow us to generate synthetic data sets $\tilde{\xi}_i(d)$, whose statistical properties can be compared to those of the original data set. We make this comparison using the spectral properties of the correlation matrix. In other words, with Eq. (1) and $\tilde{\xi}_i(d)$ we compute a ''β-synthetic'' correlation matrix; we determine the spectrum of eigenvalues and compare it to that of the original matrix. The parameter β can be tuned in order to get the best fit.

This procedure was carried out for the S&P 500 data set. The eigenvalue spectra of the two matrices are compared in Fig. 5 for $\beta=48$. The value of β was chosen by visual inspection as that giving the best fit. The curves are remarkably close, suggesting that Eq. (17) provides a good statistical description of the correlations among assets.

Once the value β^* that gives the best fit is found, we can compute the *noise undressed* correlation matrix

$$C^*_{i,j}(\infty)=\frac{\delta_{i,j}+\sum_s \sqrt{g^*_{s,i}g^*_{s,j}}}{\sqrt{1+\sum_s g^*_{s,i}}\sqrt{1+\sum_s g^*_{s,j}}} \tag{19}$$

from the parameters $g^*_{s,i}=g_{s,i}(\beta^*)$. This is the correlation matrix of a synthetic data set obtained from Eq. (17) with $D\to\infty$. Figure 5 shows the eigenvalue distribution of the noise undressed matrix $C^*_{i,j}(\infty)$. This allows one to appreciate the effect of noise dressing. As expected, noise mainly affects small eigenvalues.

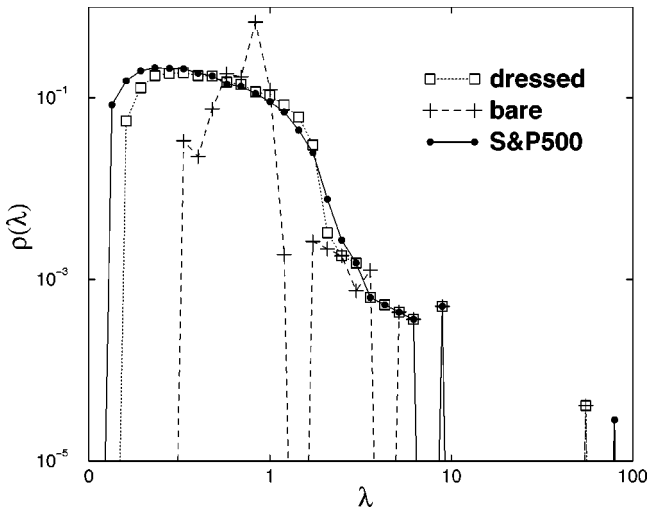

FIG. 5. Comparison of the spectrum of the S&P 500 correlation matrix (solid line ●), with noise-dressed (dotted □), and bare (dashed +) correlation matrices generated by Eq. (17).

VII. DISCUSSION AND CONCLUSIONS

The applicability of the method can be extended considerably to a generic data set $\{\vec{x}_i\}_{i=1}^N$. $\vec{x}_i$ need not be a time series. The distribution of $x_i(d)$ need not be Gaussian and it does not even need to be the same across i. For example, $x_i(d)$ may be the measure of the dth feature of the ith object or the concentration of species i in the dth sample of an experiment. The idea is to map the data set $\vec{x}_i$ into a Gaussian time series $\vec{\xi}_i$ to which we apply Eq. (2). The mapping results from requiring that nonparametric cross correlations $\tau^x_{i,j}=\tau^{\xi}_{i,j}$ are preserved. To do this in practice we compute Kendall's τ [12] for the $\vec{x}_i$ data sets: $\tau^x_{i,j}=\langle \mathrm{sgn}[x_i(d)-x_i(d')]\mathrm{sgn}[x_j(d)-x_j(d')]\rangle_{d<d'}$. For two Gaussian time series with correlation $C_{i,j}$ one can compute analytically $\tau^x_{i,j}$ in the limit $D\to\infty$. This leads to the relation

$$C_{i,j}=\sin\left(\frac{\pi\tau^x_{ij}}{2}\right) \tag{20}$$

between Gaussian and nonparametric correlations. This equation allows us to translate nonparametric correlations into Gaussian correlations. From these one can build H_c of Eq. (8) and study the clustering properties.

This procedure has been tested for the S&P 500 data set, for which it is known that $\xi_i(d)$ has non-Gaussian statistics [13]. We have found indistinguishable results that indicate that the deviations from Gaussian behavior have little or no effect on the results. We expect that this approach breaks down when the marginal distribution of $\xi_i(d)$ is such that the second moment is not defined. In that case $C_{i,j}$ computed from $\tau_{i,j}$ and Eq. (20) can even fail to be positive definite.

With respect to Ref. [2], our approach does not need any assumption on the form of the Hamiltonian. As input, the method only needs the correlation matrix $C_{i,j}$ (or $\tau_{i,j}$). The range of interactions is set by the correlations themselves. Indeed our method predicts a nontrivial ground state $\mathcal{S}_0$ that is not, in general, the ferromagnetic one.

For small D, the local interaction of Ref. [2] may well be more efficient in capturing the structure of data. Our method is most useful in cases where $D\sim N\gg 1$. These ideas can clearly be extended to models of correlations different from Eq. (2) as shown, for example, in the Appendix.

ACKNOWLEDGMENTS

We acknowledge R. Zecchina, R. Pastor-Satorras, and D. Vergni for interesting discussions and R. N. Mantegna for providing the S&P 500 data.

APPENDIX

We define here the explicative factor model for stocks returns (also known as multi-index model, see e.g., [8]):

$$\xi_i(t)=\frac{\vec{v}_i\vec{\eta}(t)+\epsilon_i}{\sqrt{1+v_i^2}},\quad v_i^2=\vec{v}_i\vec{v}_i=\sum_{\alpha=1}^{L}(v_i^{\alpha})^2 \tag{A1}$$

where $\vec{v}_i$ are L dimensional vectors and $\vec{\eta}(t)$ is a L dimensional Gaussian random vector $\langle\eta^{\alpha}(t)\rangle=0$ and $\langle\eta^{\alpha}(t)\eta^{\beta}(t')\rangle=\delta_{\alpha,\beta}\delta_{t,t'}$.

The idea is that there are L explicative factors $\eta^1(t),\ldots,\eta^L(t)$ that describe the fluctuations of each stock price. This model is different from the one we considered in the text in that each time series is coupled to all other with a different strength. This can be easily understood by observing that the model (2) can be cast in the form of an explicative factor model with $v_i^{\alpha}=g_{s_i}\delta_{\alpha,s_i}$. This is a rather particular form of Eq. (A1). We observe, however, that L must be much smaller than N in order to avoid problems of overfitting with Eq. (A1), while Eq. (2) requires $L\approx N$.

As we did in Sec. II we look at the probability of observing the time series $\xi_i(t)$ given the model Eq. (A1) and the parameters $\vec{v}_i$:

$$P\{\xi_i(t)|\vec{v}_i\}=\prod_{t=1}^{D}\left\langle\prod_{i=1}^{N}\delta\left(\xi_i(t)-\frac{\vec{v}_i\vec{\eta}+\epsilon_i}{\sqrt{1+v_i^2}}\right)\right\rangle_{\vec{\eta},\epsilon}. \tag{A2}$$

After taking the average over the Gaussian variables one obtains the equivalent of Eqs. (6) and (5)

$$P\{\xi_i(t)|\vec{v}_i\}=e^{-DH\{\vec{v}\}},$$

$$H\{\vec{v}\}=\frac{1}{2}\sum_i\left[(1+v_i^2)-\ln(1+v_i^2)\right]-\frac{1}{2}\mathrm{Tr}\ln(1+V)$$

$$-\frac{1}{2}\mathrm{Tr}\frac{\chi}{1+V}, \tag{A3}$$

where we have defined the matrices

$$V_{\alpha,\beta}=\sum_{i=1}^{N}v_i^{\alpha}v_i^{\beta} \tag{A4}$$

$$\chi_{\alpha,\beta} = \sum_{i,j=1}^{N} C_{i,j} \sqrt{1+v_i^2} v_i^\alpha \sqrt{1+v_j^2} v_j^\beta , \qquad \text{(A5)}$$

and $C_{i,j}$ is defined in Eq. (1). We note that the second term in Eq. (A3) is subextensive, and could be neglected; nevertheless in the presence of the matrix χ a Monte Carlo simulation becomes excessively time consuming, since a change in $\vec{v}_k$ requires order N operations to compute the new matrix. This may considerably limit the practical applicability of this method.

[1] K. Rose, E. Gurewitz, and G. C. Fox, Phys. Rev. Lett. **65**, 945 (1990).

[2] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996); S. Wiseman, M. Blatt, and E. Domany, Phys. Rev. E **57**, 3767 (1997).

[3] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, *ibid.* **83**, 1471 (1999).

[4] R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).

[5] L. Kullmann, J. Kertesz, and R. N. Mantegna, Physica A **287**, 412 (2000).

[6] M. Marsili, A. Vergni, and L. Giada (to be published).

[7] J. D. Noh, Phys. Rev. E **61**, 5981 (2000).

[8] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (Wiley, New York, 1995).

[9] In principle, the low-temperature properties may be dominated by exponentially many metastable states, as in spin glasses. This would mean that the ansatz (2) is not appropriate. Indeed multiple solutions is an undesirable outcome of a statistical fit. In the cases discussed here, we found no glassy behavior.

[10] A. M. Sengupta and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).

[11] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953). Note that cluster algorithms, as those used in Ref. [2], cannot be used for H_c.

[12] See, e.g., W. H. Press *et al.*, *Numerical Recipes in Fortran* (Cambridge University Press, Cambridge, England, 1992), p. 633.

[13] R. N. Mantegna and H. E. Stanley, Nature (London) **376**, 46 (1995).