INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 3009–3032 PII: S0305-4470(03)37794-7

# A new method to estimate the noise in financial correlation matrices

**Thomas Guhr**[1] **and Bernd Kälber**[2,3]

[1] Matematisk Fysik, LTH, Lunds Universitet, Box 118, 22100 Lund, Sweden
[2] Max Planck Institut für Kernphysik, Postfach 103980, 69029 Heidelberg, Germany

E-mail: Thomas.Guhr@matfys.lth.se

Received 28 May 2002, in final form 5 November 2002
Published 12 March 2003
Online at stacks.iop.org/JPhysA/36/3009

**Abstract**

Companies belonging to the same industrial branch are subject to similar economical influences. Hence, the time series of their stocks can show similar trends implying a correlation. Financial correlation matrices measure the unsystematic correlations between time series of stocks. Such information is important for risk management. It has been found by Laloux *et al* that the correlation matrices are 'noise dressed', a major reason being the finiteness of the time series. We present a new and alternative method to estimate this noise. We introduce a power mapping of the elements in the correlation matrix which suppresses the noise and thereby effectively 'prolongs' the time series. Neither further data processing nor additional input is needed. To develop and test our method, we use a model suggested by Noh which can be viewed as a special case of a 'factor model' in economics. We perform numerical simulations for the time series and obtain correlation matrices. We support the numerics by a qualitative analytical discussion. With our approach, different correlation structures buried under this noise can be detected. Our method is general and can be applied to all systems in which time series are measured.

PACS numbers: 89.65.G, 02.50, 05.45.T

## 1. Introduction

There are different kinds of risk in a stock portfolio, for example, a systematic one due to the general trends in the economy which affect the entire market, and an unsystematic one due to events affecting only segments of the whole market under consideration, such as companies within the same industrial branch or in one country [1]. All this is borne out in the time series of the stocks. In a more general framework, the term 'stock of a company' should be

[3] Present address: Group Risk Control, Dresdner Bank AG, Jürgen-Ponto-Platz 1, D-60301 Frankfurt, Germany.

replaced by 'risk factor'. The corresponding portfolios of a bank or an investment company can contain hundreds or thousands of financial instruments, depending on many different risk factors [1]. However, to be explicit, we will use the term 'stocks' instead of 'risk factors'. Banks or portfolio managers are particularly interested in the unsystematic correlations [2, 3], which we refer to simply as correlations from now on. The higher the number of stocks belonging to the same branch in a portfolio, the more probable is the presence of sizable correlations in the portfolio and, hence, the higher is the risk that a bad performance of the corresponding branch drags down the value of the entire portfolio. Thus, one tries to control the risk by diversification, i.e. by a careful selection of the stocks and their relative weights in the portfolio. To this end, the best possible information about the correlations present in the portfolio is needed.

The elements of the correlation matrix are calculated in the following way: first, the time series of all stocks are labelled with a running index and normalized to zero mean and unit variance, second, the values of two such normalized time series, with indices $k$ and $l$, say, for equal times are multiplied with each other and, third, these products are time averaged, i.e. they are summed over time and divided by the lengths of the time series. This gives the $(k, l)$ element of the correlation matrix. The number of companies considered is then the size of this correlation matrix. We will assume that this number is large enough for a statistical analysis.

In recent years, it was shown that the true correlations one is interested in can be buried under noise in the commonly used correlation matrices: Laloux *et al* [4] studied an empirical correlation matrix and found that the bulk part of its spectral density can be modelled by a purely random matrix. Plerou *et al* [5] worked out spectral fluctuations and also found that they are compatible with random matrix statistics. Random matrices are often used as statistical models for quantum chaotic or related spectral problems, for reviews see [6–8]. Burda *et al* [9, 10] designed a proper model involving random Levy matrices. The presence of random matrix features was coined noise dressing of financial correlations. In practice, such noise can appear if the length of the time series employed to compute the correlation matrices is too short. Gopikrishnan *et al* [11] developed a method to reduce the noise by taking advantage of the fact that large eigenvalues outside the bulk of the spectral density can be associated with industrial branches. These authors [11] used this for portfolio optimization, see also [12]. Mantegna [13] and Bonanno *et al* [14] identified branches through stock correlations by constructing a minimal spanning tree. A maximum likelihood approach to reduce the noise was put forward by Marsili [15] and Giada and Marsili [16, 17]. In this context, we mention that financial correlation matrices have also been used to study the relation between the markets of different countries [18] or globalization effects [19].

Here, we present a new approach to identify and estimate the noise in such correlation matrices and, thereby, also the strength of the true correlations. As we do not use the large eigenvalues outside the bulk of the spectral density, our method is an alternative and a complement to the approach of [11]. It will be particularly useful if some of the large eigenvalues are not large enough to lie outside the bulk of the spectral density. Our method is based on a power mapping. It does not require any further data processing or any other type of input. To the best of our knowledge, it seems to be a new technique for the analysis of correlation matrices in time series problems. We have three goals: first, we develop an alternative method for noise identification, second, we introduce the power mapping as a new mathematical–statistical technique and discuss its features in some detail, third, we present our findings in a self-contained way to make possible a transfer to other time series problems and the corresponding correlation matrices.

The paper is organized as follows. In section 2, we outline the stochastic model, including a review of correlation matrices and a discussion of the noise dressing phenomenon. We study

the dependence of the spectral density as a function of the length of the time series in section 3. In section 4, we introduce the power mapping and use it to identify and estimate the noise. We summarize and conclude in section 5. Three more detailed discussions are given in the appendix.

## 2. Stochastic model

We formulate the model for financial correlations in section 2.1. Since we also aim at a heuristic analytical treatment later on, we do that in some detail. Correlation matrices and noise dressing are discussed in sections 2.2 and 2.3, respectively.

### *2.1. Normalized time series for correlated stocks*

Most models for time series $S(t)$ of stocks [20–23] have a random component, involving a random number $\varepsilon$ and a volatility constant $\sigma^2$, and a non-random component, the drift part, involving a drift constant $\mu$. Geometric Brownian motion

$$\frac{\mathrm{d}S}{S} = \mu\,\mathrm{d}t + \sigma\varepsilon\sqrt{\mathrm{d}t} \tag{1}$$

is particularly popular. The dimensionless quantity $\mathrm{d}S/S$ is referred to as return. Due to the central limit theorem, geometric Brownian motion leads, largely independently of the distribution for the random numbers $\varepsilon$, to a log–normal distribution of the stock prices. This is in fair agreement with the empirical distributions for price changes about one day and greater [24, 25]. The tails of the empirical distributions are, in general, much fatter [20, 21]. To describe this, one uses, for example, autoregressive processes which take into account that the volatility is also a fluctuating quantity, see the discussion in [20]. In the present context of correlations, however, it suffices to model the time series in the spirit of geometric Brownian motion. In economics, one wishes to measure the correlations independent of the drift. Thus, one usually takes the logarithms or the logarithmic differences of the time series to remove the exponential trend in the data due to the drift. Moreover, one normalizes the data to zero mean value and unit variance.

In view of these requirements, Noh [26] suggested a proper model to study financial correlations. From an economics viewpoint, it is of the one-factor type and can be interpreted as an application of the arbitrage pricing model due to Ross [27], see also the discussion in appendix A. From a physics viewpoint, as pointed out in [15, 28], Noh's model has much in common with certain models involving interacting Potts spins [29]. Consider a market involving $K$ companies, labelled $k = 1, \ldots, K$, and $B$ industrial branches, labelled $b = 1, \ldots, B$. The companies within the same industrial branch are assumed to be correlated. The companies are ordered in such a way that the indices $k$ of companies within the same branch follow each other. For example, we have $K = 50$ companies and $B = 3$ branches and we assume that the first branch with $b = 1$ consists of the first $\kappa_1 = 5$ companies, the second branch with $b = 2$ consists of the next $\kappa_2 = 12$ companies, the third branch with $b = 3$ consists of the next $\kappa_3 = 8$ companies, and, finally, we assume that the remaining $\kappa = 25$ companies are in no branch. The branch index $b$ is viewed as a function of the company index $k$, i.e. we have $b = b(k)$. For the $\kappa$ companies which are not in any branch, we set $b = b(k) = 0$. We refer to $\kappa_b$ as the size of the industrial branch $b$. Obviously we have

$$\sum_{b=1}^{B} \kappa_b + \kappa = K. \tag{2}$$

Of course, we assume that $\kappa_b > 1, b = 1, \ldots, B$. The number $\kappa$ can be any non-negative integer, including zero.

The normalized time series $M_k(t), k = 1, \ldots, K$ of the returns for the $K$ companies are modelled as the sum of two purely random contributions: the first one models the correlations within a given branch and is thus common to this branch, involving random numbers $\eta_b(t)$, the second one is specific for the company and involves random numbers $\varepsilon_k(t)$,

$$M_k(t) = \sqrt{\frac{p_{b(k)}}{1 + p_{b(k)}}}\eta_{b(k)}(t) + \frac{1}{\sqrt{1 + p_{b(k)}}}\varepsilon_k(t). \tag{3}$$

The two contributions are weighted with a parameter $p_{b(k)}$, common to all companies in the branch $b$. We assume that the $\eta_{b(k)}(t)$ and $\varepsilon_k(t)$ are uncorrelated and standard normal distributed with zero mean value. The weights are assumed to be positive with $p_{b(k)} \geqslant 0$. Since the distributions are symmetric, this is the most general form of the weights. In the case that $k$ is not in any branch, i.e. for $b = 0$, we set $p_{b(k)} = 0$. Here, we use discrete time steps and normalize the time units such that $\mathrm{d}t = 1$. The time series $M_k(t)$ consist of $T$ time values at $t = 1, \ldots, T$.

A comment regarding the branches in the model is in order. An international portfolio is likely to contain globally operating, large companies. Their activities directly affect various branches, but they are also under different economical influences in different countries. In such cases, it is common to add a further label 'country' to the label 'branch', denoted by $b$ in our case. Generally speaking, it is often necessary to go from a one-factor model of the type discussed in the present paper to a two-factor or multi-factor model (see [27]). From the viewpoint of a simulation, this refined correlation structure implies the need for precise information about the weights which generalize our $p_b$. On the other hand, not much changes from the viewpoint of data analysis, which is the main motivation for the present work.

### 2.2. Correlation matrices

The time average of a function $F(t)$ over the time series is defined as

$$\langle F(t)\rangle_T = \frac{1}{T}\sum_{t=1}^{T} F(t) \tag{4}$$

which depends on the length $T$ of the time series. If the time series are infinitely long, $T \to \infty$, we have

$$\langle M_k(t)\rangle_\infty = 0 \qquad \text{and} \qquad \left\langle M_k^2(t)\right\rangle_\infty = 1. \tag{5}$$

Here, we simply used $\langle \eta_{b(k)}(t)\rangle_\infty = 0$, $\langle \varepsilon_k(t)\rangle_\infty = 0$ and $\langle \eta_{b(k)}(t)\eta_{b(l)}(t)\rangle_\infty = \delta_{b(k)b(l)}$, $\langle \eta_{b(k)}(t)\varepsilon_l(t)\rangle_\infty = 0$, $\langle \varepsilon_k(t)\varepsilon_l(t)\rangle_\infty = \delta_{kl}$.

The correlation coefficient between two companies labelled $k$ and $l$ is the average over the product of the two normalized time series,

$$C_{kl}(T) = \frac{1}{T}\sum_{t=1}^{T} M_k(t)M_l(t) = \langle M_k(t)M_l(t)\rangle_T. \tag{6}$$

If one views the numbers $M_k(t)$ as the entries of a $K \times T$ rectangular matrix $M$, one has

$$C(T) = \frac{1}{T}MM^\dagger = \langle M(t)M^\dagger(t)\rangle_T \tag{7}$$

for the $K \times K$ correlation matrix $C$. As these averages depend on the length $T$ of the time series, we add the argument $T$ to the correlation matrix. Within our model outlined above, one finds for infinitely long time series [15, 16], $T \to \infty$,

$$C_{kl}(\infty) = \lim_{T\to\infty} C_{kl}(T) = \frac{1}{1 + p_{b(k)}}(p_{b(k)}\delta_{b(k)b(l)} + \delta_{kl}). \tag{8}$$

Thus, the matrix $C(\infty)$ consists of $B$ square blocks on the diagonal of dimensions $\kappa_b \times \kappa_b$ with off-diagonal entries $p_b/(1+p_b)$ for branch $b$, and a $\kappa \times \kappa$ unit matrix for the companies which are in no branch. The diagonal entries are all unity. All other entries are zero: the correlation coefficients between companies which are not in any branch, those between companies belonging to different branches and those between a company which is in a branch and another one which is not. Some issues related to the normalization of correlation matrices are discussed in appendix A.

### *2.3. Noise dressing in the model*

If the time series have only a finite length, $T < \infty$, it is obvious that the averages $\langle \eta_{b(k)}(t)\eta_{b(l)}(t)\rangle_T$, $\langle \eta_{b(k)}(t)\varepsilon_l(t)\rangle_T$ and $\langle \varepsilon_k(t)\varepsilon_l(t)\rangle_T$ give neither zero nor unity, but some finite numbers. This describes the noise dressing of financial correlation matrices which was found in [4] in the framework of Noh's model. Due to the finiteness of the time series, there is a purely random offset to every correlation coefficient, burying the true correlation coefficient which would be found if the time series were sufficiently long.

As we aim at an analytical discussion later on, we formulate the noise dressing more quantitatively. We employ a standard result of mathematical statistics [30]: consider two time series of uncorrelated random numbers $\alpha_k(t)$ and $\alpha_l(t)$, $t = 1, \ldots, T$ with standard normal distributions and zero mean value. The average $\langle \alpha_k(t)\alpha_l(t)\rangle_T$ is a random number, following a Gauss distribution centred at mean value unity with variance $2/T$, if $k = l$, and following a Gauss distribution centred at mean value zero with variance $1/T$, if $k \neq l$. Hence, we can write this average to leading order in $T$ as

$$\langle \alpha_k(t)\alpha_l(t)\rangle_T = \delta_{kl} + \sqrt{\frac{1+\delta_{kl}}{T}}\, a_{kl} \tag{9}$$

where the $a_{kl}$ are uncorrelated random numbers, independent of $T$, with standard normal distribution and zero mean value. This yields for the correlation coefficients the expression

$$\begin{aligned} C_{kl}(T) = {} & \sqrt{\frac{p_{b(k)}}{1+p_{b(k)}}}\sqrt{\frac{p_{b(l)}}{1+p_{b(l)}}}\left(\delta_{b(k)b(l)} + \sqrt{\frac{1+\delta_{b(k)b(l)}}{T}}\, a_{b(k)b(l)}\right) \\ & + \frac{1}{\sqrt{1+p_{b(k)}}}\frac{1}{\sqrt{1+p_{b(l)}}}\left(\delta_{kl} + \sqrt{\frac{1+\delta_{kl}}{T}}\, a_{kl}\right) + \sqrt{\frac{p_{b(k)}}{1+p_{b(k)}}}\frac{1}{\sqrt{1+p_{b(l)}}}\frac{1}{\sqrt{T}}\, a_{b(k)l} \\ & + \sqrt{\frac{p_{b(l)}}{1+p_{b(l)}}}\frac{1}{\sqrt{1+p_{b(k)}}}\frac{1}{\sqrt{T}}\, a_{kb(l)} \end{aligned} \tag{10}$$

to leading order in $T$. The first two terms stem from the averages $\langle \eta_{b(k)}(t)\eta_{b(l)}(t)\rangle_T$ and $\langle \varepsilon_k(t)\varepsilon_l(t)\rangle_T$, the last two from the interference averages $\langle \eta_{b(k)}(t)\varepsilon_l(t)\rangle_T$. There are four sets of random numbers, $a_{b(k)b(l)}$, $a_{b(k)l}$, $a_{kb(l)}$ and $a_{kl}$, respectively. These sets are understood to be different from each other. To avoid cumbersome notation, we do not add further indices to specify that. For example, the number $a_{b(1)b(1)}$ is different from $a_{11}$. As required, the limit $T \to \infty$ of equation (10) correctly yields equation (8).

## 3. Spectral density and length of the time series

As shown in [4], the spectral density is a well-suited observable to study the noise dressing of empirical correlation matrices. In section 3.1, we numerically investigate the spectral density of our model as a function of the length $T$ of the time series. We support our findings by an analytical discussion in section 3.2.

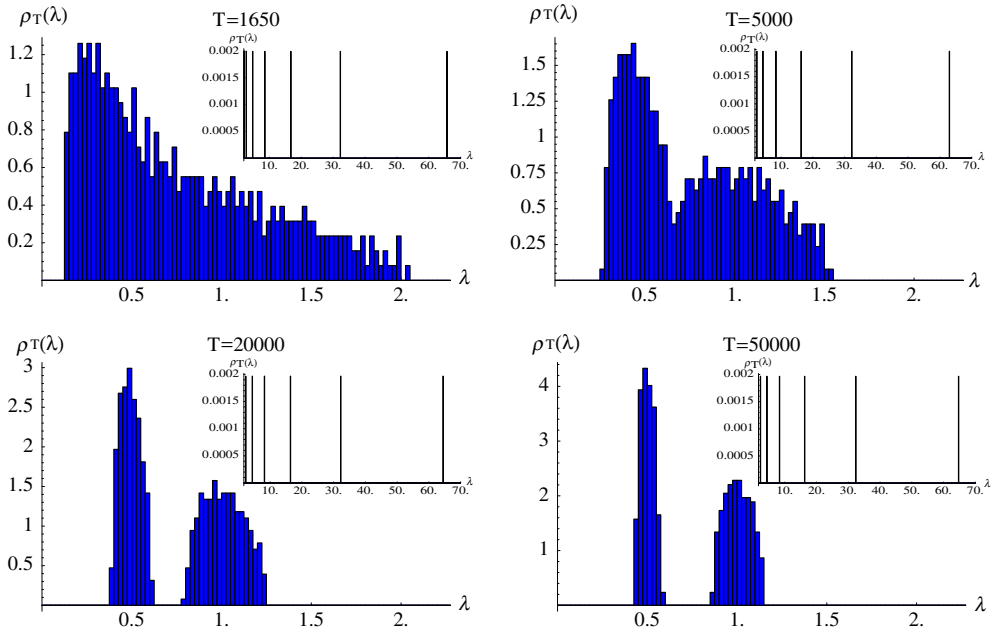


**Figure 1.** Spectral densities $\rho_T(\lambda)$ of simulated correlation matrices. The length $T$ of the time series increases from top to bottom according to $T = 1650$, 5000, 20 000, 50 000. The presentation of every density is split into the regions $0 \leqslant \lambda \leqslant 2.2$ and $2.2 \leqslant \lambda \leqslant 70$. The densities are given in units of $K$.

**Table 1.** The sizes $\kappa_b$ and the weights $p_b = 1 - 1/\kappa_b$ for the $B = 6$ industrial branches used in the numerical simulation.

| $b$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\kappa_b$ | 4 | 8 | 16 | 32 | 64 | 128 |
| $p_b$ | 0.75 | 0.88 | 0.94 | 0.97 | 0.98 | 0.99 |

### *3.1. Numerical simulation*

As the number of companies in the analysis [4, 5] of real market data is several hundreds, we choose this number as $K = 508$ for our simulation. We assume that there are $B = 6$ industrial branches whose sizes $\kappa_b$ and weights $p_b$ are listed in table 1. Since Noh [26] showed that, for example, the choice

$$p_b = 1 - \frac{1}{\kappa_b} \tag{11}$$

leads to spectral densities having a shape similar to the empirical data [4, 5], we use the same parametrization here. In addition, we have $\kappa = 256$ companies which are not in any branch.

We simulate the correlation matrices $C(T)$ for four different lengths $T = 1650$, 5000, 20 000, 50 000 of the time series, calculate the eigenvalues $\lambda_k$, $k = 1, \ldots, K$ and work out the spectral densities $\rho_T(\lambda)$. The results are shown in figure 1. The first density $\rho_{1650}(\lambda)$ for the shortest time series with $T = 1650$ resembles very much the densities of the empirical correlation matrices found in [4, 5]. There is a bulk spectrum in the interval $0 < \lambda < 2$. Moreover, a number of isolated peaks in the interval $2 < \lambda < 70$ is seen. Each of these large eigenvalues corresponds to an individual industrial branch. For increasing lengths $T$,

the isolated peaks are stable, but the bulk spectrum separates into two groups of eigenvalues. The left group is roughly centred around $\lambda = 1/2$, the right one around $\lambda = 1$. We will see in the analytical discussion to be performed in section 3.2 that the left group can clearly be associated with the true correlations due to the industrial branches, while the right group only represents noise around the trivial self-correlation of the companies for $k = l$.

### *3.2. Analytical discussion*

For infinitely long time series, $T \to \infty$, the correlation matrix $C(\infty)$ is block diagonal according to equation (8), and the spectrum of the eigenvalues $\lambda_k(\infty)$, $k = 1, \ldots, K$ is easily calculated. For every branch $b$, we have $\kappa_b - 1$ degenerate eigenvalues $1/(1 + p_b)$ and one numerically much larger eigenvalue $(1 + \kappa_b p_b)/(1 + p_b)$ which is for large size $\kappa_b$ roughly proportional to $\kappa_b$. Moreover, there are $\kappa$ degenerate eigenvalues which are unity for the companies which are in no branch. Thus, the spectral density is given by

$$\rho_\infty(\lambda) = \sum_{b=1}^{B}(\kappa_b - 1)\delta\left(\lambda - \frac{1}{1 + p_b}\right) + \sum_{b=1}^{B}\delta\left(\lambda - \frac{1 + \kappa_b p_b}{1 + p_b}\right) + \kappa\delta(\lambda - 1). \tag{12}$$

This formula helps in understanding the last density for the longest time series with $T = 50\,000$ in figure 1. The first term corresponds to the left group, representing the true correlations. The second term of the $B$ large eigenvalues yields the isolated peaks of the spectra. The last term in the formula represents the noise peaked at unity. The eigenvectors $u_k(\infty)$, $k = 1, \ldots, K$ can also be calculated in a straightforward manner for $T \to \infty$.

The density (12) is smeared out for finite time series, $T < \infty$. To estimate the resulting noisy density $\rho_T(\lambda)$ to leading order in $T$, we write the correlation matrix in the form $C(T) = C(\infty) + C_1/\sqrt{T}$ where $C_1$ can be read off from the expansion (10). We also write $\lambda_k(T) = \lambda_k(\infty) + \lambda_{1,k}/\sqrt{T}$ for the eigenvalues and $u_k(T) = u_k(\infty) + u_{1,k}/\sqrt{T}$ for the eigenvectors. One quickly finds $\lambda_{1,k} = u_k^\dagger(\infty)C_1u_k(\infty)$. Since the elements of every $u_k(\infty)$ are numbers depending on the weights $p_b$ and the sizes $\kappa_b$, every coefficient $\lambda_{1,k}$ is, according to equation (10), a linear combination of the standard normal distributed random numbers $a_{b(k)b(l)}$, $a_{b(k)l}$, $a_{kb(l)}$ and $a_{kl}$. Hence, it is itself a Gaussian distributed random number $a_k$ with zero mean value and a variance $v_k^2$ which is some function of the $p_b$ and the $\kappa_b$. We arrive at

$$\lambda_k(T) = \lambda_k(\infty) + \frac{v_k}{\sqrt{T}}a_k \tag{13}$$

for $k = 1, \ldots, K$ to leading order in $T$. We note that in this expansion, to order $1/\sqrt{T}$, the Gaussian random variables $a_k$ are uncorrelated. The effect of the smearing out, i.e. of the noise in the model, is marginal for the $B$ numerically large eigenvalues $(1 + \kappa_b p_b)/(1 + p_b)$, $b = 1, \ldots, B$. Their positions change the less, the larger the size $\kappa_b$. The effect of the noise is much stronger on the numerically smaller eigenvalues. Moreover, their degeneracies are lifted and distributions develop around their mean values. For the eigenvalues $1/(1 + p_b)$, $b = 1, \ldots, B$ which belong to the $B$ industrial branches and, hence, describe true correlations, the mean value is

$$\mu_B = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{1 + p_b} \tag{14}$$

while it simply reads $\mu_0 = 1$ for the eigenvalues which are unity and were generated by the time series not belonging to any branch. As the weights satisfy $p_b > 0$ if $b = 1, \ldots, B$, we always have

$$\mu_B < 1 = \mu_0. \tag{15}$$

This important relation implies that the centres of the distributions generated by the noise from the two sets of degenerate eigenvalues are always different from each other. The distribution due to the true correlations will always be left of that which is due to trivial self-correlation for $k = l$. In the numerical simulation, we chose $p_b = 1 - 1/\kappa_b$. Assuming that the sizes $\kappa_b$ are large, we find to leading order

$$\mu_B = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{2-1/\kappa_b} = \frac{1}{2} + \frac{1}{4B}\sum_{b=1}^{B}\frac{1}{\kappa_b} \tag{16}$$

which explains why the distribution due to the true correlations is roughly centred around $1/2$ in figure 1.

To find an estimate for the noisy density $\rho_T(\lambda)$, we argue phenomenologically. Some further analytical properties are compiled in appendix B. We first replace the first term of equation (12) with a Gaussian distribution centred at $\mu_B$,

$$G\left(\lambda - \mu_B, v_B^2/T\right) = \sqrt{\frac{T}{2\pi v_B^2}}\exp\left(-\frac{(\lambda-\mu_B)^2}{2v_B^2/T}\right). \tag{17}$$

For its variance, we write $v_B^2/T$ where $v_B$ is a proper geometric average of those numbers $v_k$ in equation (13) which stem from the industrial branches. As a weight factor, we sum over the multiplicities in the degenerate case (12),

$$\sum_{b=1}^{B}(\kappa_b - 1) = K - \kappa - B. \tag{18}$$

Similarly, we could replace the third term of equation (12) with a Gaussian centred at $\mu_0 = 1$. However, it is better to employ the fact that the $\kappa \times \kappa$ block of the companies which are in no branch belongs to a chiral random matrix ensemble [31, 12] whose density is known [32, 33] to have the algebraic form

$$\rho_{\mathrm{ch}}(\lambda, \kappa/T) = \frac{T}{2\pi\kappa}\,\mathrm{Re}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{19}$$

where $\lambda_\pm$ are the largest and the smallest eigenvalues, respectively, given by

$$\lambda_\pm = 1 + \frac{\kappa}{T} \pm 2\sqrt{\frac{\kappa}{T}}. \tag{20}$$

Both of them converge to $\mu_0 = 1$ for $T \to \infty$. Outside the interval $\lambda_- < \lambda < \lambda_+$, the square root in equation (19) is strictly imaginary. Thus, the real part ensures that the function $\rho_{\mathrm{ch}}(\lambda, \kappa/T)$ is zero outside the supporting interval $\lambda_- < \lambda < \lambda_+$. As a weight factor, we take the multiplicity $\kappa$ in the degenerate case (12). We note that, analogous to $v_B$, there would be in principle a scale $v_0$ entering the density $\rho_{\mathrm{ch}}(\lambda, \kappa/T)$. It would result from a proper average of those numbers $v_k$ in equation (13) which do not stem from the industrial branches. However, since we start out from normalized time series and since the weights $p_b$ do not contribute here, we expect $v_0 = 1$. The second term of equation (12) we leave unchanged, because we can ignore the shift in the positions of these largest eigenvalues. Thus, collecting everything, we find

$$\rho_T(\lambda) = (K - \kappa - B)G\left(\lambda - \mu_B, v_B^2/T\right) + \sum_{b=1}^{B}\delta\left(\lambda - \frac{1+\kappa_b p_b}{1+p_b}\right) + \kappa\rho_{\mathrm{ch}}(\lambda, \kappa/T). \tag{21}$$

As the functions $G\left(\lambda - \mu_B, v_B^2/T\right)$ and $\rho_{\mathrm{ch}}(\lambda, \kappa/T)$ are normalized to unity, the noisy density $\rho_T(\lambda)$ is correctly normalized to the total number $K$ of eigenvalues. In figure 2 we fit formula

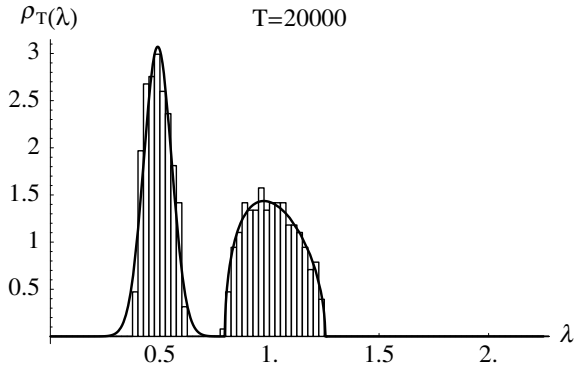


**Figure 2.** Bulk part of the spectral density $\rho_{20000}(\lambda)$ for $T = 20\,000$. The phenomenological formula (21) is fitted to it. The density is given in units of $K$.

(21) to the bulk part of the spectral density $\rho_{20000}(\lambda)$ for $T = 20\,000$ which was presented in figure 1. The agreement is very good, the fit yields $v_B = 9$.

In appendix B, we show that the first term in equation (21) is an envelope for several peaks, each one corresponding to one industrial branch. To leading order $1/\sqrt{T}$, one has

$$\rho_T(\lambda) = \sum_{b=1}^{B}(\kappa_b - 1)\bar{G}_T\left(\lambda - \frac{1}{1+p_b}, v^2\right) + \sum_{b=1}^{B}\delta\left(\lambda - \frac{1+\kappa_b p_b}{1+p_b}\right) + \kappa\bar{G}_T(\lambda - 1, v^2) \quad (22)$$

where we defined the average

$$\bar{G}_T(z, v^2) = \frac{1}{K}\sum_{k=1}^{K} G\left(z, v_k^2/T\right) \quad (23)$$

over the $K$ Gaussians peaks resulting from the smearing out of every eigenvalue. The parameters $v_k^2$ are functions of the sizes $\kappa_b$ and the weights $p_b$. If these variances $v_k^2$ are small enough, only those eigenvalues which are close to $1/(1 + p_b)$ will give a non-negligible contribution to the averages $\bar{G}_T$ in the first term of equation (22). Thus, in this case, each industrial branch yields one peak entering the sum over the branches in the first term of equation (22). If the parameters are such that these peaks cannot be resolved, we arrive at the estimate (21), where additionally the last term has been replaced by equation (19).

## 4. Noise identification by power mapping

In section 4.1, we introduce the power mapping and discuss its properties numerically. We give a qualitative analytical explanation in section 4.2. In section 4.3, we demonstrate how the power mapping can detect different correlation structures. Finally, we define a measure for the noise in section 4.4.

### *4.1. Power mapping*

The true correlations buried under the noise become visible in the spectral density if the time series are long enough. Thus, if we found a procedure that is equivalent, in some sense, to a prolongation of the time series, we would be able to identify and quantify the noise in a given correlation matrix. In the following, we develop such a procedure, the power mapping.

We map the correlation matrix $C(T)$ to the matrix $C^{(q)}(T)$. Here, $q$ is a positive number and the elements of $C^{(q)}(T)$ are calculated according to the definition

$$C_{kl}^{(q)}(T) = \mathrm{sign}(C_{kl}(T))|C_{kl}(T)|^q. \tag{24}$$

Thus, the mapping preserves the sign of the matrix element $C_{kl}(T)$ and raises the modulus of it to the $q$th power. We note that $C^{(q)}(T)$ is the matrix of the powers of the elements of $C(T)$, but not the power of the matrix $C(T)$. This is crucial, because the spectra of $C(T)$ and $(C(T))^q$ are, for $q \neq 0$, related in a simple way: if $\lambda_k, k = 1, \ldots, K$ are the eigenvalues of $C(T)$, $\lambda_k^q, k = 1, \ldots, K$ are the eigenvalues of $(C(T))^q$. The spectrum of $C^{(q)}(T)$ is much more complicated and depends on the eigenvalues and the eigenvectors of $C(T)$. However, depending on the numerical value chosen for $q$, it allows one to suppress, in $C^{(q)}(T)$, certain elements of $C(T)$. This will now be demonstrated through a numerical study of the spectral densities.

We simulate correlation matrices for $K = 508$ companies from time series of length $T_0 = 1650$ with the sizes $\kappa_b$ and the weights $p_b$ given in table 1. We do so to make possible a comparison with figure 1, where the impact of increasing length for the time series was studied, starting from a correlation matrix with the same sizes and weights and for the same length $T = 1650$. To improve the statistical significance and to better illustrate the effect of the power mapping, we simulate an ensemble of 25 such correlation matrices from time series of length $T_0 = 1650$. We choose powers $q = 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5$ and calculate, for a fixed $q$, the 25 power mapped matrices $C^{(q)}(T)$ from the 25 matrices $C(T)$ of the ensemble. Then, the individual spectra are evaluated and the density $\rho_{T_0}^{(q)}(\lambda)$ results as the ensemble average. All densities are shown in figure 3.

The power mapping transforms the original density $\rho_{T_0}^{(1)}(\lambda) = \rho_{T_0}(\lambda)$ for $q = 1$ into densities $\rho_{T_0}^{(q)}(\lambda)$ which show for intermediate values of $q$ two clearly separated peaks. The best separation is obtained near $q = 1.5$. For values beyond $q = 2$, the two peaks glue together again, and the separation is lost. What do these two peaks in the densities $\rho_{T_0}^{(q)}(\lambda)$ for $1.25 \leqslant q \leqslant 2.0$ represent? Comparison with figure 1 shows that, indeed, the left peak in figure 3 corresponds to the one for the true correlations in figure 1, while the right peak in figure 3 corresponds to the one for the noise around the trivial self-correlation of the companies for $k = l$ in figure 1. Hence, we have found the desired procedure which roughly amounts to a prolongation of the time series.

### *4.2. Qualitative analytical discussion*

To leading order in the length $T$ of the time series, the $C_{kl}(T)$ are given by equation (10). To understand the effect of the power mapping, we distinguish three different cases in considering $|C_{kl}(T)|^q$. First, we power map the diagonal elements $C_{kk}(T)$. As equation (10) shows, the vast majority of the matrix elements will be positive if $T$ is sufficiently large. Thus, to simplify the discussion, we ignore the absolute value sign. To leading order in $T$, we have

$$(C_{kk}(T))^q = 1 + \frac{q}{1 + p_{b(k)}} \left( \sqrt{2} p_{b(k)} a_{b(k)b(k)} + \sqrt{2} a_{kk} + \sqrt{p_{b(k)}} (a_{b(k)k} + a_{kb(k)}) \right) \frac{1}{\sqrt{T}}. \tag{25}$$

Second, we power map the off-diagonal elements $C_{kl}(T)$ in the blocks of the industrial branches where $k \neq l$ but $b(k) = b(l)$. For the same reason as in the previous case, we ignore the absolute value sign, and find

$$(C_{kl}(T))^q = \left( \frac{p_{b(k)}}{1 + p_{b(k)}} \right)^q + \frac{q (p_{b(k)})^{q-1}}{(1 + p_{b(k)})^q} \Big( \sqrt{2} p_{b(k)} a_{b(k)b(k)} + \sqrt{2} a_{kl} + \sqrt{p_{b(k)}} (a_{b(k)l} + a_{kb(l)}) \Big) \frac{1}{\sqrt{T}} \tag{26}$$

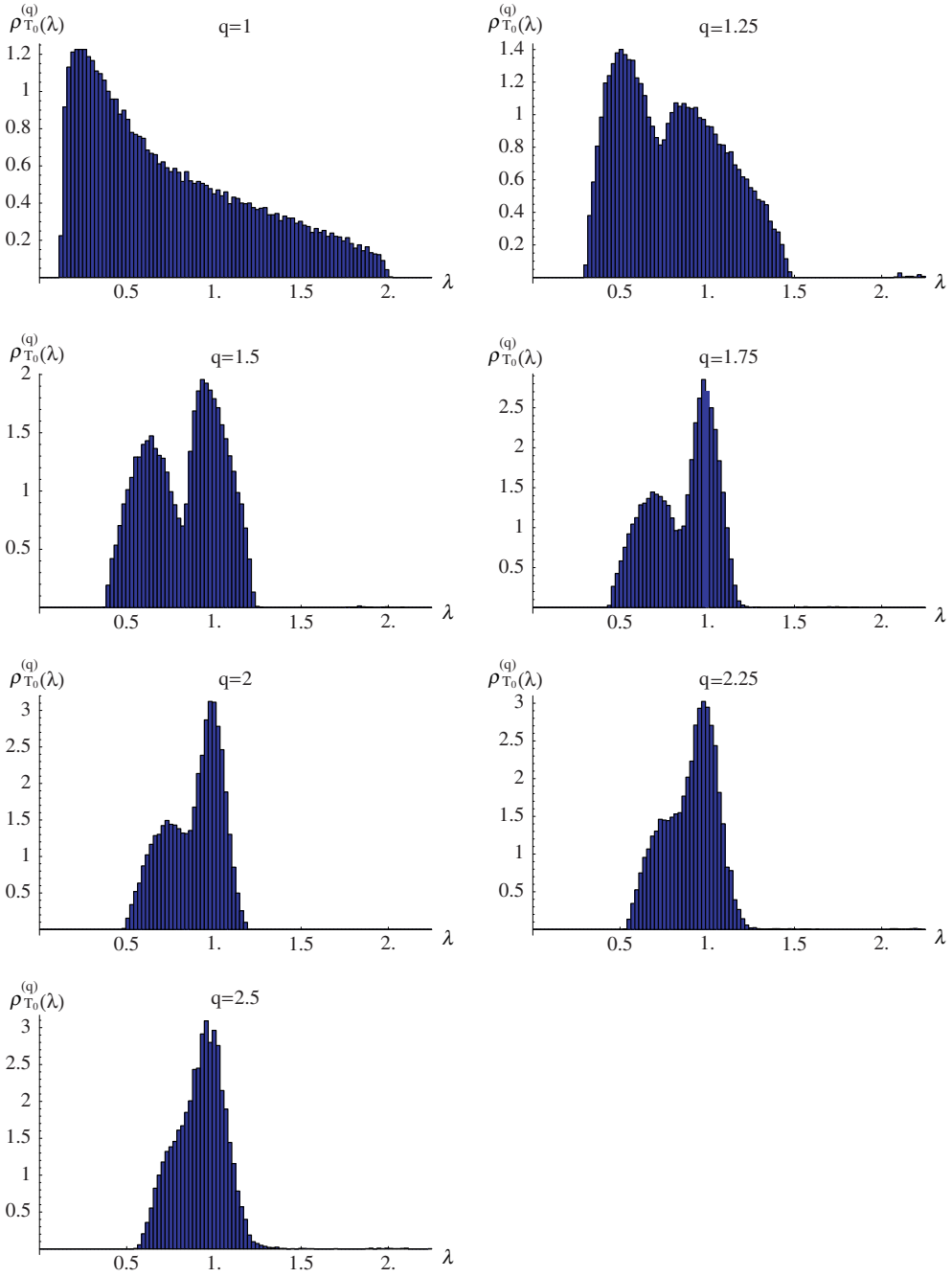


**Figure 3.** Spectral densities $\rho_{T_0}^{(q)}(\lambda)$ of the power mapped correlation matrices $C^{(q)}$. The length of the time series is always $T_0 = 1650$. The values of the powers used are $q = 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5$. An ensemble of 25 matrices $C$ was simulated which were power mapped onto $C^{(q)}$. The densities are given in units of $K$.

to leading order in $T$. Third, we power map the elements $C_{kl}(T)$ outside the blocks, where $k \neq l$ and $b(k) \neq b(l)$. Since all Kronecker $\delta$ in equation (10) are zero in this case, we obtain

$$|C_{kl}(T)|^q = \frac{1}{((1+p_{b(k)})(1+p_{b(l)}))^{q/2}} |\sqrt{p_{b(k)}}\sqrt{p_{b(l)}}a_{b(k)b(l)} + a_{kl} + \sqrt{p_{b(k)}}a_{b(k)l} + \sqrt{p_{b(l)}}a_{kb(l)}|^q \frac{1}{T^{q/2}} \tag{27}$$

as the leading order in $T$. We note that the matrix elements $C_{kl}(T)$ in this third case will be positive or negative with equal probability.

In the first two cases (25) and (26), the powers $(C_{kl}(T))^q$ contain a $T$ independent term plus a term which is of order $1/\sqrt{T}$. In the third case, however, there is no $T$ independent term, and the leading order of the whole expression is $1/T^{q/2}$. Thus, for $q > 1$, $|C_{kl}(T)|^q$ vanishes faster in the third case than in the first two terms. As the case (27) comprises all elements outside the blocks, where $k \neq l$ and $b(k) \neq b(l)$, we find that, for $q > 1$, the power mapped matrix $C^{(q)}(T)$ is block diagonal to leading order $1/\sqrt{T}$. There are two consequences: first, a separation between the industrial branches on one hand and the companies which are in no branch on the other hand takes place and, second, the individual industrial branches are also separated from each other. This explains why the power mapping is comparable to a prolongation of the time series. At first sight, one would expect that the effect is the stronger, the larger the value of $q$. However, this is not so, because the $T$ independent terms are different for the matrix elements in the blocks: in the first case (25) of the diagonal elements, it is unity, but a number smaller than unity in the second case (26). The larger the value of $q$, the smaller the latter term becomes. Hence, the unity in the diagonal elements dominates more and more, driving the eigenvalues towards unity. The effect discussed above which is comparable to a prolongation of the time series, leads to the separation of the spectral densities into two peaks. If $q$ becomes too large, this is counteracted, and the two peaks merge again. Consequently there must be an optimal value for $q$. In the numerical simulation we found that it is roughly $q = 1.5$.

For infinitely long time series $T \to \infty$, the power mapped correlation matrix $C^{(q)}(\infty)$ is trivially block diagonal. The eigenvalues $\lambda_k^{(q)}(\infty)$ and the spectral density $\rho_\infty^{(q)}(\lambda)$ are found along lines similar to those leading to equation (12). We arrive at

$$\rho_\infty^{(q)}(\lambda) = \sum_{b=1}^{B}(\kappa_b - 1)\delta\left(\lambda - \left(1 - \left(\frac{p_b}{1+p_b}\right)^q\right)\right) + \sum_{b=1}^{B}\delta\left(\lambda - \left(1 + (\kappa_b - 1)\left(\frac{p_b}{1+p_b}\right)^q\right)\right) + \kappa\delta(\lambda - 1). \tag{28}$$

This correctly reduces to equation (12) for $q = 1$. As already argued, the peaks due to the first term move towards the peak due to the third term if $q$ is made large. Again, the $\delta$ peaks are smeared out if the length of the time series is finite. We show in appendix C that the noisy spectral density $\rho_T^{(q)}(\lambda)$ of the power mapped correlation matrix is to leading order in $T$ given by

$$\rho_T^{(q)}(\lambda) = \sum_{b=1}^{B}(\kappa_b - 1)\bar{G}_T\left(\lambda - \left(1 - \left(\frac{p_b}{1+p_b}\right)^q\right), (v^{(q)})^2\right) + \sum_{b=1}^{B}\delta\left(\lambda - \left(1 + (\kappa_b - 1)\left(\frac{p_b}{1+p_b}\right)^q\right)\right) + \kappa\bar{G}_{T^q}(\lambda - 1, (\tilde{v}^{(q)})^2) \tag{29}$$

where we use definition (23). The parameters $(v^{(q)})^2$ and $(\tilde{v}^{(q)})^2$ are functions of $q$ and of the weights $p_b$ and the sizes $\kappa_b$. Due to the recovered block structure, the companies in the industrial branches are separated from those which are in no branch and, importantly, also

**Table 2.** The sizes $\kappa_b$ and the weights $p_b$ for the industrial branches used in the numerical simulation for the three ensembles of correlation matrices with different structures. The total dimension of the matrices is always $K = 508$.

| | Top structure[a] | | Middle structure[b] | | Bottom structure[c] | |
|---|---|---|---|---|---|---|
| $b$ | $\kappa_b$ | $p_b$ | $\kappa_b$ | $p_b$ | $\kappa_b$ | $p_b$ |
| 1 | 2 | 0.5 | 4 | 0.008 | 4 | 0.75 |
| 2 | 4 | 0.75 | 8 | 0.01 | 104 | 0.99 |
| 3 | 7 | 0.85 | 16 | 0.03 | | |
| 4 | 10 | 0.9 | 32 | 0.07 | | |
| 5 | 15 | 0.93 | 64 | 0.2 | | |
| 6 | 20 | 0.95 | 128 | 0.99 | | |
| 7 | 30 | 0.96 | | | | |
| 8 | 42 | 0.97 | | | | |
| 9 | 64 | 0.984 | | | | |
| 10 | 98 | 0.989 | | | | |
| 11 | 140 | 0.99 | | | | |

[a] The number of industrial branches is $B = 11$, $\kappa = 76$ companies are in no branch.
[b] The number of industrial branches is $B = 6$, $\kappa = 256$ companies are in no branch.
[c] The number of industrial branches is $B = 2$, $\kappa = 400$ companies are in no branch.

the industrial branches are separated from each other: each branch contributes one peak to the first term in equation (29). If these peaks cannot be resolved, we can replace them with one peak and arrive at the estimate

$$\rho_T^{(q)}(\lambda) = (K - \kappa - B)G\left(\lambda - \mu_B^{(q)}, \left(v_B^{(q)}\right)^2/T\right) + \sum_{b=1}^{B} \delta\left(\lambda - \left(1 + (\kappa_b - 1)\left(\frac{p_b}{1+p_b}\right)^q\right)\right) + \kappa G\left(\lambda - 1, \left(v_0^{(q)}\right)^2/T^q\right). \quad (30)$$

Here, the parameters $\left(v_B^{(q)}\right)^2$ and $\left(v_0^{(q)}\right)^2$ collect and combine the previous parameters $(v^{(q)})^2$ and $(\tilde{v}^{(q)})^2$. Moreover,

$$\mu_B^{(q)} = \frac{1}{B}\sum_{b=1}^{B}\left(1 - \left(\frac{p_b}{1+p_b}\right)^q\right) = 1 - \frac{1}{B}\sum_{b=1}^{B}\left(\frac{p_b}{1+p_b}\right)^q \quad (31)$$

approximates the centre of the Gaussian due to the true correlations in the first term of equation (30). The third term estimates the peaks due to the noise. As discussed in appendix C, these two contributions are affected differently when finite lengths of the time series are concerned. Thus, it is useful to scale the variance with $1/T$ in the first term, but with $1/T^q$ in the third one.

### *4.3. Detecting different correlation structures*

We consider three examples for different correlation structures. The parameters are listed in table 2, and the corresponding correlation matrices $C$ are displayed in figure 4. We refer to the three examples as top, middle and bottom structures, respectively. For all three structures, the total number of companies is $K = 508$ and the length of the time series is $T_0 = 1650$. The top structure has many branches ($B = 11$) with relatively strong correlations and little noise, the middle structure has fewer branches ($B = 6$) with weaker correlations and more noise and, finally, the bottom structure has only $B = 2$ branches with stronger correlations, and a considerable amount of noise. We note that the weights $p_b$ are not chosen according to equation (11). Rather, we adjusted them in such a way that the spectral densities $\rho_{T_0}(\lambda)$

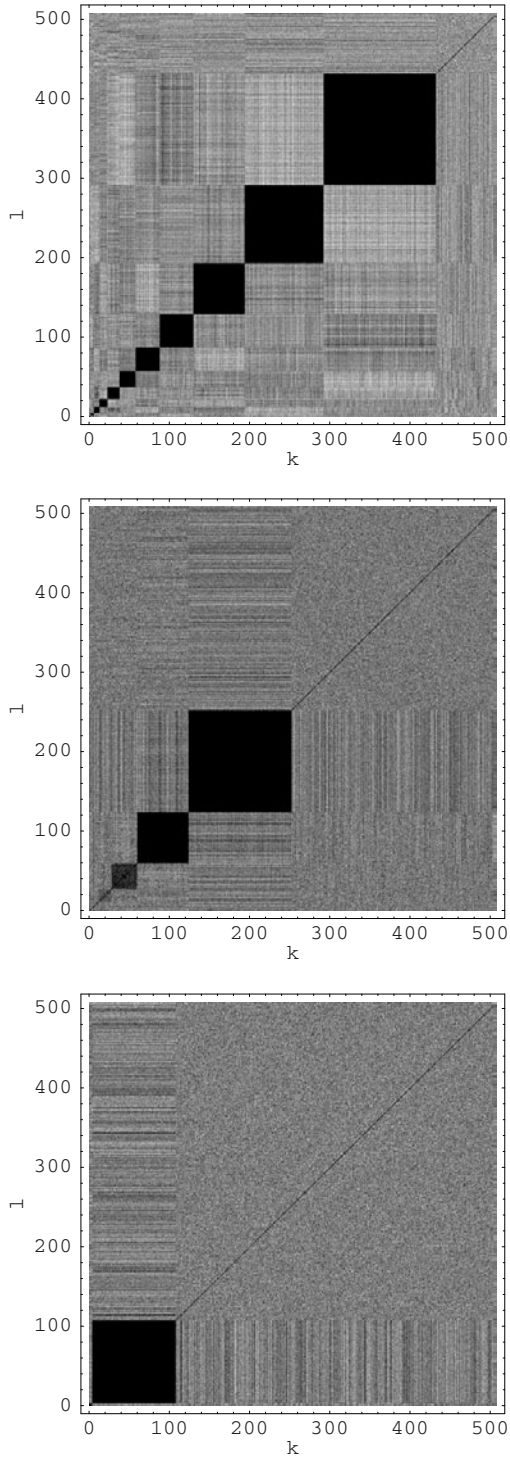


**Figure 4.** Three correlation matrices $C$ with different correlation structures. The parameters for the top, middle and bottom structures, respectively, are given in table 2. The amount of noise in the middle and the bottom structures is much higher than in the top structure.

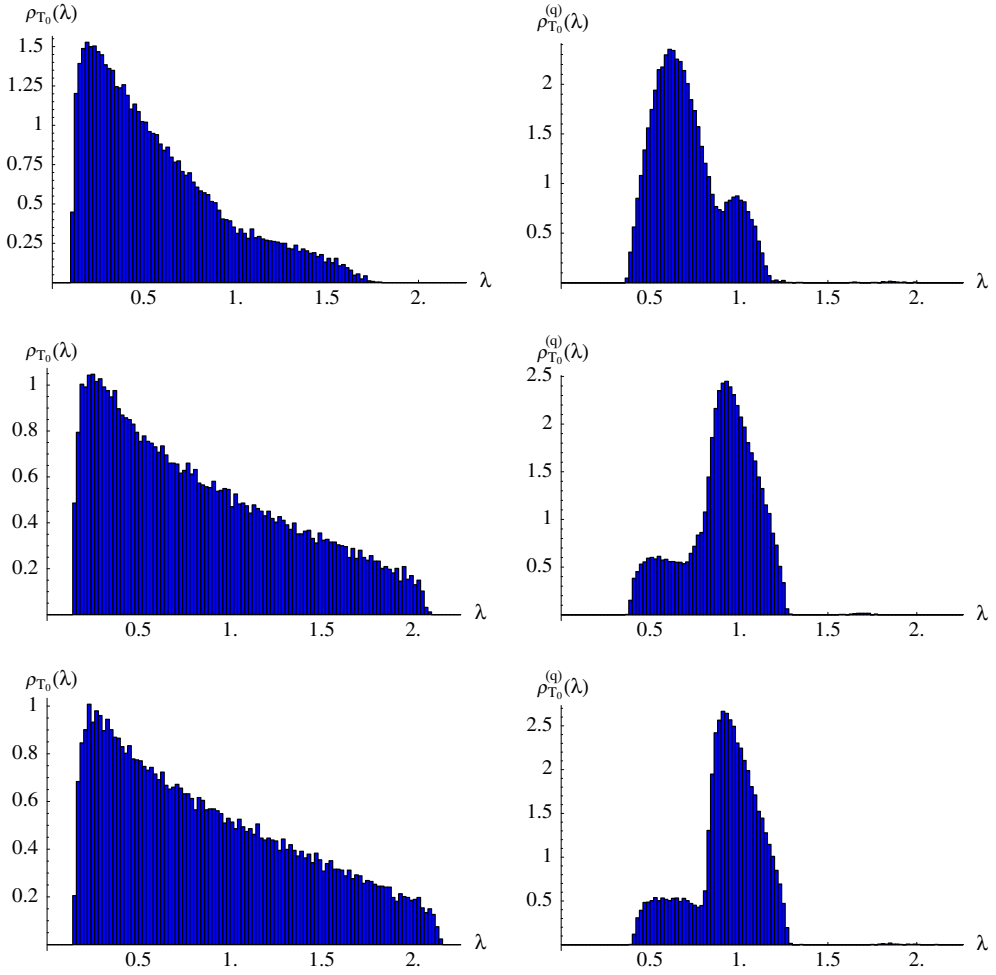


**Figure 5.** Left column: spectral densities $\rho_{T_0}(\lambda)$ of the original correlation matrices $C$. Right column: spectral densities $\rho^{(q)}_{T_0}(\lambda)$ of the same, but power mapped correlation matrices $C^{(q)}$. The length of the time series is always $T_0 = 1650$. An ensemble of 25 matrices $C$ was simulated for each structure. The value of the power used is always $q = 1.5$. The 25 matrices $C$ which were simulated for each structure were individually power mapped onto matrices $C^{(q)}$, forming a new ensemble for each structure. The densities are given in units of $K$.

for the middle and the bottom structures look as similar as possible. This can be seen in the left column of figure 5. The top structure serves as a reference example. As the figure shows, the true correlations are so dominating that the spectral density is already almost separated in two substructures which are separated by a kink around $\lambda = 1$. Thus, the top structure is an idealizing example.

We now apply the power mapping and calculate matrices $C^{(q)}$ according to equation (24). We choose the value $q = 1.5$ which we identified as optimal in section 4.1. The resulting spectral densities $\rho^{(q)}_{T_0}(\lambda)$ are shown in the right column of figure 5. For each of the three structures, we see the separation into two peaks, the left one corresponding to the true correlations, the right one to the noise. As expected, the spectral density for the top structure consisting of a big left peak and a small right peak differs considerably

from the ones for the middle and the bottom structures where the left peaks are small and the right peaks big. This nicely confirms our expectation that the power mapping is capable of efficiently distinguishing the gross structures in the correlation matrices.

It is now interesting to compare the spectral densities for the middle and bottom structures: although the shape of densities $\rho_{T_0}(\lambda)$ in the left column of figure 5 can hardly be distinguished, the densities $\rho_{T_0}^{(q)}(\lambda)$ for the power mapped correlation matrices, shown in the right column, have similar, but distinguishable shapes. For both structures, the left peak is small, the right one big. However, for the bottom structure, the right peak is narrower and its left shoulder is steeper. Hence, the power mapping also gives useful information about the fine structure in the correlation matrices.

### *4.4. Measuring correlations and noise*

Using the estimate (30), we can gain some further understanding of why $q = 1.5$ is a good value for the power mapping. In doing so, we will also obtain a measure for the ratio of correlations and noise. This measure, however, will only be sensitive to, on the one side, the industrial branches seen as one big contribution and, on the other side, the companies which do not belong to a branch. Thus, this particular measure does not say much about the noise between the industrial branches. The two Gaussian peaks which emerge from the power mapping, i.e. the one due to the true correlations and the one due to the noise, are separated if there is some space between the left side of the former and the right side of the latter. According to equation (30), the peaks should be separated if we have

$$1 - \frac{v_0^{(q)}}{T^{q/2}} > \mu_B^{(q)} + \frac{v_B^{(q)}}{\sqrt{T}} \tag{32}$$

or, equivalently, if the function

$$H(q,T) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{p_b}{1+p_b}\right)^q - \frac{v_0^{(q)}}{T^{q/2}} - \frac{v_B^{(q)}}{\sqrt{T}} \tag{33}$$

is positive. Here, we used equation (31). In figure 6, we display the function $H(q,T)$ for different power mapped correlation matrices starting from the original one used in section 3.1 with $T = T_0 = 1650$. Obviously, we may define the optimal $q$ value as the point where $H(q,T_0)$ reaches its maximum which is given by the equation $\partial H(q,T_0)/\partial q = 0$. Indeed, this value is close to $q = 1.5$. The parameters $v_0^{(q)}$ and $v_B^{(q)}$ used in figure 6 were obtained from the fitting procedure to be described in the following. As the dependence of $H(q,T)$ on $q$ is very complicated, we do not go into a further analytical discussion.

To fit the spectral density of the power mapped correlation matrices, we use an ansatz motivated by the estimate (30). As we only want to fit the bulk part, we ignore the second term in the estimate (30) due to the large eigenvalues and take

$$\rho_{\text{bulk}}^{(q)}(\lambda) = R_B^{(q)} G\big(\lambda - \mu_B^{(q)}, \big(\sigma_B^{(q)}\big)^2\big) + R_0^{(q)} G\big(\lambda - 1, \big(\sigma_0^{(q)}\big)^2\big). \tag{34}$$

The prefactors $R_B^{(q)}$, $R_0^{(q)}$, the standard deviations $\sigma_B^{(q)}$, $\sigma_0^{(q)}$ and the mean value $\mu_B^{(q)}$ are fit parameters. This relatively high number of fit parameters is acceptable, because the data to be fitted have a clear structure. Moreover, it should be emphasized that we are only interested in obtaining proper estimates by these fits. From the previous discussion, we expect that the fit should approximately give, first, $R_B^{(q)} = K - \kappa - B$ and $R_0^{(q)} = \kappa$ for the prefactors and, second, the scaling behaviour $\big(\sigma_B^{(q)}\big)^2 = \big(v_B^{(q)}\big)^2/T$ and $\big(\sigma_0^{(q)}\big)^2 = \big(v_0^{(q)}\big)^2/T^q$ for the variances. We employ the parameters $\sigma_B^{(q)}$ and $\sigma_0^{(q)}$ instead of the parameters $v_B^{(q)}$ and $v_0^{(q)}$,

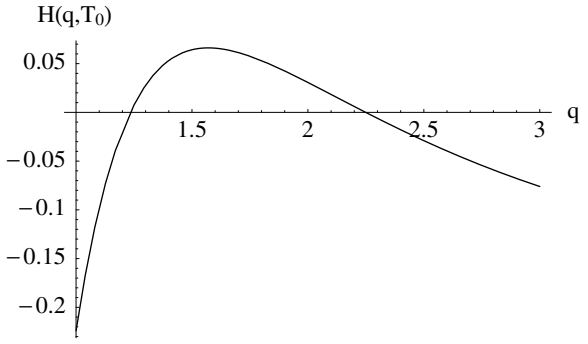


**Figure 6.** A typical example for the function $H(q, T)$ versus the power $q$. The parameters for the original correlation matrix in section 3.1 are used with $T = T_0 = 1650$. The optimal $q$ value is defined by the maximum value of $H(q, T_0)$ which is seen to be reached near $q = 1.5$.

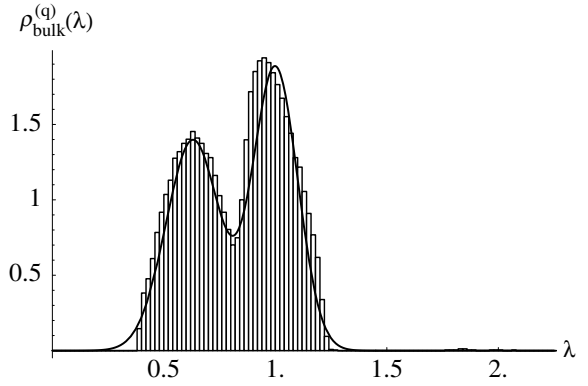


**Figure 7.** Fit of the ansatz (34) to the bulk part of the spectral density for the power mapped correlation matrix used in section 4.1 with $q = 1.5$. The density is given in units of $K$.

because, in practice, one will mostly have to deal with correlation matrices for one fixed length $T$ of the time series. As an example, we fit the ansatz (34) to the spectral density of the power mapped correlation matrix discussed in section 4.1 for $q = 1.5$. The result is shown in figure 7. We obtain for the prefactors $R_B^{(q)} = 0.42K = 213$, $R_0^{(q)} = 0.47K = 239$, for the standard deviations $\sigma_B^{(q)} = 0.12$, $\sigma_0^{(q)} = 0.10$ and for the mean value $\mu_B^{(q)} = 0.63$. Thus, we obtain the estimate $\kappa = R_0^{(q)} = 239$ for the number of companies which are not in any branch. This is in fair agreement with the true value $\kappa = 256$. The sum $R_B^{(q)} + R_0^{(q)} = 0.89K = 452$ deviates by about 10% from the theoretical value $R_B^{(q)} + R_0^{(q)} = K - B = 502$. This is not so surprising, considering the fact that our ansatz (34) can only be a rough approximation. Hence, one can use the ratio

$$r^{(q)} = \frac{R_0^{(q)}}{R_B^{(q)}} \tag{35}$$

to characterize a correlation matrix. The larger $r^{(q)}$, the more noise is present, the smaller $r^{(q)}$, the stronger are the true correlations. In our example, we have $r^{(q)} = 1.12$, implying that noise and true correlations are more or less equally strong. Again, we point out that this measure is largely independent of the noise between the industrial branches. Thus, the noise

we are talking about in the present context is only that between all branches and the companies which do not belong to any branch. Advantageously, the ratio (35) is not so sensitive to the total number $K$ of companies, implying that correlation matrices of different sizes $K$ can have the same $r^{(q)}$. Of course, the number $K$, the dimension of the correlation matrix, is always known in an analysis. In the previous subsection, we discussed a correlation matrix, labelled 'middle structure', also involving $\kappa = 256$ companies which are in no branch. Most of the correlations within the branches are so weak that the left peak in the middle of the right column in figure 5 is considerably smaller than the one in figure 7. The ratio $r^{(q)}$ makes the desired distinction: the overall strength of the true correlations is much weaker than in the present case. In addition, the standard deviation $\sigma_B^{(q)}$ and the mean value $\mu_B^{(q)}$ yield information about the spreading of the weights $p_b$ and about their average.

Finally, we test the quality of the scaling behaviour $\sigma_B^{(q)} \propto 1/\sqrt{T}$ and $\sigma_0^{(q)} \propto 1/T^{q/2}$ for the standard deviations. To this end, we generate correlation matrices for various lengths $T$ of the time series and power map them using $q = 1.0$, $q = 1.25$ and $q = 1.5$. We fit the resulting spectral densities using the ansatz (34), extract the standard deviations $\sigma_B^{(q)}$ and $\sigma_0^{(q)}$, and fit the latter to the expected scaling behaviour $1/\sqrt{T}$ and $1/T^{q/2}$, respectively. The expectation is well confirmed for $\sigma_B^{(q)}$. In the case of $\sigma_0^{(q)}$, the general trend is reproduced for the three $q$ values. For $q = 1.5$, however, the most interesting value, the agreement is good. This is encouraging, because the steps that led to the estimate (30) involved various approximations. In any case, as already argued, the practical applicability of our ansatz (34) is not affected by these scaling questions.

As figure 8 shows, both standard deviations $\sigma_B^{(q)}$ and $\sigma_0^{(q)}$ become smaller as $q$ is made larger while $T$ is kept fixed. This is why the power mapping can be viewed as effectively prolonging the time series: to obtain given values for the standard deviations, one can either keep $q = 1$ and make the time series longer or make $q$ larger and leave $T$ unchanged. For example, compared to the correlation matrix with $T = 1000$ and $q = 1$, the noise is reduced by 60% when one goes to $T = 10\,000$ without changing $q = 1$ and it is reduced by 75% when one applies the power mapping with $q = 1.5$ while keeping $T = 1000$. This is most important for applications, because time series for stocks with $T = 10\,000$, say, are seldom available, while time series with $T = 1000$ are easily accessible. Thus, power mapping can lead to better estimates for the risk and for related quantities such as the value-at-risk [2, 3].

## 5. Summary and conclusion

We developed a new method to identify and estimate the noise and, hence, also the strength of the true correlations in a financial correlation matrix. The essence of our approach is the power mapping. It suppresses those matrix elements which can be associated with the noise, first, between the industrial branches and those companies which are in no branch and, second, between different industrial branches. Our approach could be seen as a generalization and extension of shrinking techniques [34] used in contexts different from the present one. Importantly, the spectral density changes drastically due to the suppression. The method itself yields a criterion to choose the optimal power. Loosely speaking, our method is an effective prolongation of the time series. This feature makes it particularly suited for problems in which, for whatever reason, only relatively short time series are available.

We developed the method using Noh's model which is known to reproduce the crucial features of empirical correlation matrices. Of course, the model is schematic. Generalizations, such as a multi-factor extension [27], could also describe some finer details of correlation matrices. This would not affect the applicability of our method, because it is completely

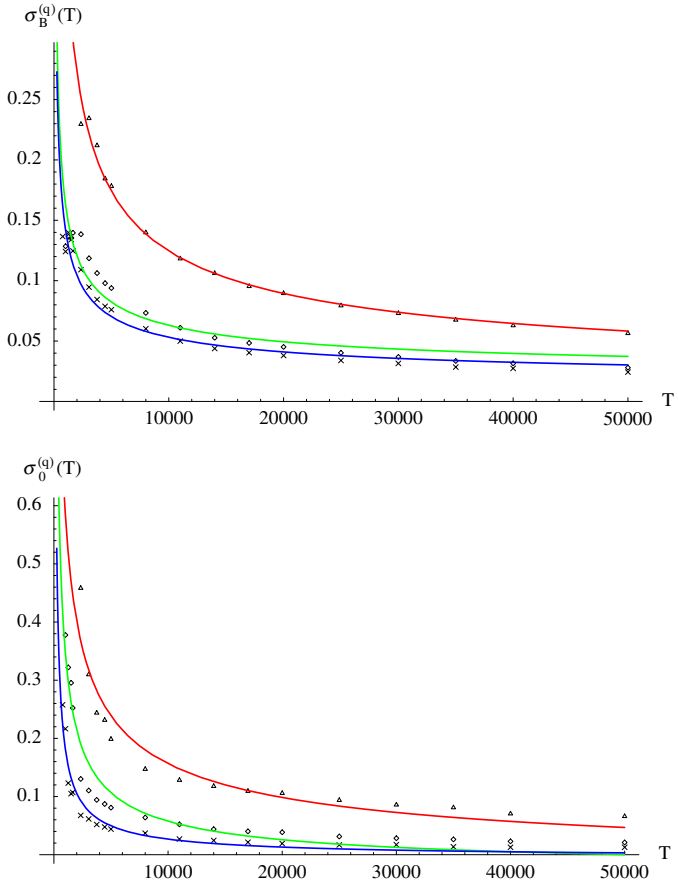


**Figure 8.** Standard deviations $\sigma_B^{(q)}$ (top) and $\sigma_0^{(q)}$ (bottom) versus the length $T$ of the time series. The dots are the data points obtained from the fits to the spectral densities. The solid lines are the fits of the expected scaling behavior $1/\sqrt{T}$ and $1/T^{q/2}$ to these data points. The top, middle and bottom curves correspond to $q = 1.0$, $q = 1.25$ and $q = 1.5$, respectively.

input-free and model-independent. Nevertheless, we plan to apply it to the more complex correlation structure of multi-factor models.

The maximum likelihood approach of Marsili [15] and Giada and Marsili [16, 17] also provides a method to suppress the noise in the entire spectrum and to obtain the true or bare correlations. As we understand it, this method relies on a matching between an original and a 'synthetic' correlation matrix and a fitting procedure based upon that. Thereby, the assumption is made that the correlation structure is compatible with a generalization of Noh's model, i.e. with a multi-factor model. Although this input is not too restrictive, we find it worthwhile to mention that our method is free of any assumption of this kind. There is no matching or any similar intermediate step in our approach. We developed and tested it for Noh's model, but none of its features went into the power mapping method, it is model-free.

Moreover, our method is a fundamentally different alternative to the technique of Gopikrishnan *et al* [11] (see also [12]). These authors use the fact that the large eigenvalues outside the bulk can be associated with the industrial branches. The large eigenvalues are then interpreted as the most relevant ones and the eigenvalues in the bulk are simply set to zero. The resulting reduced diagonal matrix of the eigenvalues is now rotated back into the

basis of the original correlation matrix. Thus, a filtered correlation matrix is obtained. It is shown that it yields a considerably improved risk estimation. The power mapping reduces the noise in an entirely different way, using all information in the correlation matrix. Imagine, for example, that many stocks go into the correlation matrix which are only correlated with a few others. Thus, the correlation matrix contains various small blocks. As the large eigenvalues are roughly proportional to the size of these blocks, it can easily happen that the eigenvalues due to these small blocks slip under the bulk of the spectrum. While one would, by construction, disregard them in the filtering approach described above, the information due to those eigenvalues is present in our method and cannot be missed. This is why it states an alternative and a complement. In this context, we underline that the power mapping allows one to distinguish different correlation structures. The application of our method to correlation matrices obtained from real market data is in progress and will be published elsewhere.

We presented our method in the framework of stocks, but it can be used for every correlation matrix of time series, independent of what these time series describe: in the context of finance, these can be stocks, interest rates, exchange rates or other risk factors, or, in other fields of science, completely different observables. Here, it is important that our approach does not involve further data processing or any other additional input.

**Acknowledgments**

We thank L A N Amaral, P Gopikrishnan, V Plerou, B Rosenow, H E Stanley (Boston University), T Rupp (Yale University), P Neu (Dresdner Bank, Frankfurt) and R Dahlhaus (Universität Heidelberg) for fruitful discussions. TG acknowledges support from the Heisenberg Foundation and thanks H A Weidenmüller and the Max Planck Institut für Kernphysik (Heidelberg) for hospitality during the initial stages of the work.

## Appendix A. Normalizations and features of correlation matrices

Empirical correlation matrices for the time series $S_k(t), k = 1, \ldots, K$ of stocks, say, are often constructed in the following way: to remove the drift in the data, first, one takes the logarithms and, second, normalizes the resulting time series $G_k(t), k = 1, \ldots, K$ to zero mean and unit variance,

$$M_k(t) = \frac{G_k(t) - \langle G_k(t)\rangle_T}{\sqrt{\left\langle G_k^2(t)\right\rangle_T - \langle G_k(t)\rangle_T^2}}. \tag{A.1}$$

By construction, we have $\langle M_k(t)\rangle_T = 1$, independently of how the averaging procedure is defined and for all $T$. We note that this is not the case for our model in equation (3). Using this normalization, the correlation matrix is given by

$$C_{kl}(T) = \langle M_k(t)M_l(t)\rangle_T = \frac{\langle G_k(t)G_l(t)\rangle_T - \langle G_k(t)\rangle_T\langle G_l(t)\rangle_T}{\sqrt{\left\langle G_k^2(t)\right\rangle_T - \langle G_k(t)\rangle_T^2}\sqrt{\left\langle G_l^2(t)\right\rangle_T - \langle G_l(t)\rangle_T^2}}. \tag{A.2}$$

Again, by construction, the diagonal elements are unity, $C_{kk}(T) = 1$, independent of the averaging procedure chosen and also for all $T$. Once more, this differs from the diagonal elements of the correlation matrix in our model, as can be seen from equation (10) for $k = l$.

We now investigate the off-diagonal matrix elements for $k \neq l$. To leading order in the length $T$ of the time series, we assume

$$\langle G_k(t)\rangle_T = g_k + \frac{f_k}{\sqrt{T}} \qquad \langle G_k(t)G_l(t)\rangle_T = \beta_{kl} + \frac{\gamma_{kl}}{\sqrt{T}} \tag{A.3}$$

where $g_k$, $f_k$ and $\beta_{kl}$, $\gamma_{kl}$ are constants. A straightforward calculation then yields

$$C_{kl}(T) = \frac{1}{\sqrt{\beta_{kk} - g_k^2}\sqrt{\beta_{ll} - g_l^2}} \left( c_{kl}^{(0)} + \frac{c_{kl}^{(1)}}{\sqrt{T}} \right)$$

$$c_{kl}^{(0)} = \beta_{kl} - g_k g_l \tag{A.4}$$

$$c_{kl}^{(1)} = \gamma_{kl} - g_k f_l - g_l f_k - \frac{1}{2}(\beta_{kl} - g_k g_l) \left( \frac{\gamma_{kk} - 2g_k f_k}{\beta_{kk} - g_k^2} + \frac{\gamma_{ll} - 2g_l f_l}{\beta_{ll} - g_l^2} \right)$$

to leading order in $T$. We assume that the variances for infinite $T$ are non-zero, i.e. $\beta_{kk} - g_k^2 > 0$. As required, we have $C_{kk}(T) = 1$. For $k \neq l$, the forms of the expansions to leading order in equations (10) and (A.4) coincide. In the present discussion, we have not explicitly incorporated a structure due to branches. Such information would appear, for example, in the parameters $\beta_{kl}$.

As compared to the correlation matrices resulting from the model in equation (3), the difference in the normalization of the diagonal elements in equation (A.2) has only a marginal and negligible influence on the eigenvalues and the spectral density, because there are many more off-diagonal matrix elements than diagonal ones. This is also clearly borne out in Noh's [26] and in our numerical results. Moreover, there is no structural difference for the off-diagonal matrix elements. Hence, we are convinced that the results of the present work carry over to models involving a normalization of the kind (A.1). Our main reason for working with the normalization in Noh's model is the fact that the limiting case for infinitely long time series $T \to \infty$ can conveniently be handled analytically.

We mention in passing that practitioners in banks and investment companies commonly work with time series having a typical length of 250 days. These time series are often exponentially weighted before the analysis. We do not employ such a weighting procedure, because we are interested in the question, how information about correlations of longer time series can, to some extent and under proper conditions, be estimated from shorter time series.

## Appendix B. Noisy spectral density before power mapping

In general, the spectral density $\rho_T(\lambda)$ for a given length $T$ of the time series can be written as the spectral average of the joint probability density function $P_T(\lambda(T))$ for the eigenvalues $\lambda_k(T), k = 1, \ldots, K$ of the correlation matrix $C(T)$,

$$\rho_T(\lambda) = \int \mathrm{d}[\lambda(T)] P_T(\lambda(T)) \sum_{k=1}^{K} \delta(\lambda - \lambda_k(T)) \tag{B.1}$$

where $\mathrm{d}[\lambda(T)]$ stands for the product of all differentials $\mathrm{d}\lambda_k(T)$. In our notation, we distinguish the argument $\lambda$ of the density and the eigenvalues $\lambda_k(T)$ of the correlation matrix which are always written with their argument $T$. In particular, equation (B.1) is also valid for $T \to \infty$ and we have

$$\rho_\infty(\lambda) = \int \mathrm{d}[\lambda(\infty)] P_\infty(\lambda(\infty)) \sum_{k=1}^{K} \delta(\lambda - \lambda_k(\infty)). \tag{B.2}$$

As a consequence of equation (13), the joint probability densities for finite and infinite $T$ can, to leading order $1/\sqrt{T}$, be related according to

$$P_T(\lambda(T)) = \int \mathrm{d}[\lambda(\infty)] P_\infty(\lambda(\infty)) \int \mathrm{d}[a] \prod_{l=1}^{K} G(a_l, 1) \prod_{k=1}^{K} \delta \left( \lambda_k(T) - \lambda_k(\infty) - \frac{v_k}{\sqrt{T}} a_k \right)$$

$$= \int \mathrm{d}[\lambda(\infty)] P_\infty(\lambda(\infty)) \prod_{k=1}^{K} G\left(\lambda_k(T) - \lambda_k(\infty), v_k^2/T\right). \tag{B.3}$$

Here, $G(z, w^2)$ is the Gaussian depending on the variable $z$, centred at zero, with variance $w^2$. We plug equation (B.3) into equation (B.1), do the integrals over $\lambda(T)$ and find

$$\rho_T(\lambda) = \int \mathrm{d}[\lambda(\infty)] P_\infty(\lambda(\infty)) \sum_{k=1}^{K} G\left(\lambda - \lambda_k(\infty), v_k^2/T\right)$$
$$= \sum_{k=1}^{K} \int_{-\infty}^{+\infty} \mathrm{d}\lambda' G\left(\lambda' - \lambda, v_k^2/T\right) \int \mathrm{d}[\lambda(\infty)] P_\infty(\lambda(\infty)) \delta(\lambda' - \lambda_k(\infty)). \quad \text{(B.4)}$$

In the second step, we inserted the integration over $\lambda'$ using a $\delta$ function. Since we may assume that the joint probability density $P_T(\lambda(\infty))$ is invariant under the exchange of its arguments $\lambda_k(\infty)$, we can employ equation (B.2) and integrate over the eigenvalues. We arrive at

$$\rho_T(\lambda) = \int_{-\infty}^{+\infty} \bar{G}_T(\lambda' - \lambda, v^2) \rho_\infty(\lambda') \, \mathrm{d}\lambda' \quad \text{(B.5)}$$

where we defined the average

$$\bar{G}_T(z, v^2) = \frac{1}{K} \sum_{k=1}^{K} G\left(z, v_k^2/T\right) \quad \text{(B.6)}$$

over the $K$ Gaussians. The parameters $v_k^2$ are functions of the sizes $\kappa_b$ and the weights $p_b$. Thus, to leading order $1/\sqrt{T}$, we can obtain the spectral density for finite $T$ from that for infinite $T$ by convoluting the latter with a superposition of Gaussians. Formula (B.5) is valid for every spectral density $\rho_\infty(\lambda)$. We now apply it to the spectral density (12) and obtain equation (22). As the second term contains the large, widely spaced eigenvalues, we neglect the smearing out there. With this additional approximation, the result (22) is valid to leading order $1/\sqrt{T}$. It gives an analytical motivation for the estimate (21).

## Appendix C. Noisy spectral density after power mapping

The crucial effect of the power mapping is the preservation of the block structure in $C^{(q)}(T)$ up to order $1/\sqrt{T}$. This is evident from equations (25) and (26). Only the inclusion of the order $1/T^{q/2}$ destroys this block structure as seen in equation (27). Hence, to understand the noisy spectral density up to order $1/\sqrt{T}$, we can apply the methods of appendix B individually to those $\kappa_b \times \kappa_b$ blocks which contain the companies of one branch. A modification occurs for the $\kappa \times \kappa$ block collecting the companies which do not belong to a branch. According to equation (25), this block is up to order $1/\sqrt{T}$ still a diagonal matrix, implying that the eigenvalues equal the diagonal elements. Since we have $p_b = 0$, they are given by $\lambda_k^{(q)}(T) = (C_{kk}(T))^q = 1 + q a_{kk}\sqrt{2/T}$. On the other hand, employing a line of reasoning similar to that in section 3.2, the eigenvalues of the blocks corresponding to a branch will have the form

$$\lambda_k^{(q)}(T) = \lambda_k^{(q)}(\infty) + \frac{v_k^{(q)}}{\sqrt{T}} a_k^{(q)} \quad \text{(C.1)}$$

where $a_k^{(q)}$ are independent Gaussian distributed variables with zero mean and unit variance. The parameters $v_k^{(q)}$ result from a superposition of $\kappa_b$ terms of order unity. Thus, the Gaussian smearing out will be stronger, roughly by a factor $\kappa_b$, for the eigenvalues $\lambda_k^{(q)}(\infty)$ of the blocks corresponding to a branch than for the eigenvalues $\lambda_k^{(q)}(\infty) = 1$ which do not belong to a branch. If we make the assumption that the sizes $\kappa_b$ are large, we can neglect the smearing out of the latter eigenvalues. We emphasize that this is an additional approximation which

is not motivated by the asymptotic expansion in $T$. Under this assumption, we obtain from equation (28) to leading order $1/\sqrt{T}$ the estimate

$$\rho_T^{(q)}(\lambda) = \sum_{b=1}^{B}(\kappa_b - 1)\bar{G}_T\left(\lambda - \left(1 - \left(\frac{p_b}{1+p_b}\right)^q\right), (v^{(q)})^2\right) + \sum_{b=1}^{B}\delta\left(\lambda - \left(1 + (\kappa_b - 1)\left(\frac{p_b}{1+p_b}\right)^q\right)\right) + \kappa\delta(\lambda - 1). \tag{C.2}$$

Since the second term involves the large eigenvalues outside the bulk whose spacing is large compared to the smearing out, we may assume that the corresponding $\delta$ functions are not affected, either. To find an estimate to leading order $1/T^{q/2}$, we must apply the methods of appendix B to the entire matrix, because the block structure is destroyed. We replace equation (C.1) with

$$\lambda_k^{(q)}(T) = \lambda_k^{(q)}(\infty) + \frac{v_k^{(q)}}{\sqrt{T}}a_k^{(q)} + \frac{\tilde{v}_k^{(q)}}{T^{q/2}}\tilde{a}_k^{(q)}. \tag{C.3}$$

As we are only interested in a qualitative discussion, we make the further assumption that the $\tilde{a}_k^{(q)}$ are independent Gaussian variables with zero mean and unit variance. The smearing out to order $1/T^{q/2}$ will not affect the first term of equation (C.2), because it is already of order $1/\sqrt{T}$. However, we cannot neglect it in the third term, because the parameters $\tilde{v}_k^{(q)}$ are now of order $K$. We obtain equation (29). Replacing each of the sums over the functions $\bar{G}_T$ and $\bar{G}_{T^q}$ with one Gaussian, we arrive at the estimate (30).

## References


[1] Elton E J and Gruber M J 1995 *Modern Portfolio Theory and Investment Analysis* (New York: Wiley)
[2] Eller R and Deutsch H P 1998 *Derivate und Interne Modelle, Modernes Risikomanagement* (Stuttgart: Schäffer-Pöschel)
[3] Longerstaey J, Zangari A and Howard S 1996 *Risk Metrics™—Technical Document* (New York: J P Morgan)
[4] Laloux L, Cizeau P, Bouchaud J P and Potters M 1999 *Phys. Rev. Lett.* **83** 1467
[5] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 *Phys. Rev. Lett.* **83** 1471
[6] Mehta M L 1990 *Random Matrices* 2nd edn (San Diego, CA: Academic)
[7] Haake F 2001 *Quantum Signatures of Chaos* 2nd edn (Berlin: Springer)
[8] Guhr T, Müller-Groeling A and Weidenmüller H A 1998 *Phys. Rep.* **299** 189
[9] Burda Z, Jurkiewicz J, Nowak M A, Papp G and Zahed I 2001 *Physica* A **299** 181
[10] Burda Z, Jurkiewicz J, Nowak M A, Papp G and Zahed I 2001 *Preprint* cond-mat/0103108, cond-mat/0103109
[11] Gopikrishnan P, Rosenow B, Plerou V, Amaral L A N and Stanley H E 2001 *Phys. Rev.* E **64** 035106(R)
[12] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N, Guhr T and Stanley H E 2002 *Phys. Rev.* E **65** 066126
[13] Mantegna R N 1999 *Eur. Phys. J.* B **11** 193
[14] Bonanno G, Vandewalle N and Mantegna R N 2000 *Phys. Rev.* E **62** 7615(R)
[15] Marsili M 2000 *Preprint* cond-mat/0003241
[16] Giada L and Marsili M 2001 *Phys. Rev.* E **63** 061101
[17] Giada L and Marsili M 2002 *Preprint* cond-mat/0204202
[18] Drozdz S, Grummer F, Ruf F and Speth J 2001 *Physica* A **294** 226
[19] Maslov S 2001 *Physica* A **301** 397
[20] Mantegna R N and Stanley H E 2000 *An Introduction to Econophysics* (Cambridge: Cambridge University Press)
[21] Bouchaud J P and Potters M 2000 *Theory of Financial Risks* (Cambridge: Cambridge University Press)
[22] Voit J 2001 *The Statistical Mechanics of Financial Markets* (Heidelberg: Springer)
[23] Paul W and Baschnagel J 1999 *Stochastic Processes: From Physics to Finance* (Berlin: Springer)
[24] Gopikrishnan P, Plerou V, Amaral L A, Meyer M and Stanley H E 1999 *Phys. Rev.* E **60** 5305
[25] Gopikrishnan P, Plerou V, Amaral L A, Meyer M and Stanley H E 1999 *Phys. Rev.* E **60** 6519
[26] Noh J D 2000 *Phys. Rev.* E **61** 5981

[27] Ross S 1976 *J. Econ. Theory* **13** 341
[28] Kullmann L, Kertész J and Mantegna R N 2000 *Physica* A **287** 412
[29] Wu F Y 1982 *Rev. Mod. Phys.* **54** 235
[30] Krengel U 1991 *Einführung in die Wahrscheinlichkeitstheorie und Statistik* (Braunschweig: Vieweg)
[31] Verbaarschot J J M and Wettig T 2000 *Ann. Rev. Nucl. Part. Sci.* **50** 343
[32] Dyson F J 1971 *Rev. Mex. Fís.* **20** 231
[33] Sengupta A M and Mitra P P 1999 *Phys. Rev.* E **60** 3389
[34] Dahlhaus R 2002 Private communication, Oberwolfach