Noise Dressing of Financial Correlation Matrices

Laurent Laloux,[1,*] Pierre Cizeau,[1] Jean-Philippe Bouchaud,[1,2] and Marc Potters[1]
[1]*Science & Finance, 109-111 rue Victor Hugo, 92532 Levallois Cedex, France*
[2]*Service de Physique de l'État Condensé, Centre d'études de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette Cedex, France*
(Received 15 December 1998)

We show that results from the theory of random matrices are potentially of great interest to understand the statistical structure of the empirical correlation matrices appearing in the study of multivariate time series. The central result of the present study, which focuses on the case of financial price fluctuations, is the remarkable agreement between the theoretical prediction (based on the assumption that the correlation matrix is random) and empirical data concerning the density of eigenvalues associated to the time series of the different stocks of the S&P 500 (or other major markets). In particular, the present study raises serious doubts on the blind use of empirical correlation matrices for risk management.

PACS numbers: 05.45.Tp, 02.10.Sp, 05.40.Ca, 87.23.Ge

Empirical correlation matrices are of great importance in data analysis in order to extract the underlying information contained in "experimental" signals and time series (e.g., experimental data de-noising, pattern recognition, weather forecast, econometric data, multivariate analysis, etc.). In addition to the direct measure of correlations, various classes of statistical tools, such as principal component analysis, singular value decomposition, and factor analysis, strongly rely on the validity of the correlation matrix in order to obtain the meaningful part of the signal. Thus, it is important to understand quantitatively the effect of noise and of the finiteness of the time series in the determination of the empirical correlation. It is also advisable to develop "null-hypothesis" tests in order to check the statistical validity of the results obtained against totally random cases.

In the case of financial assets—on which we will focus in the following—the study of empirical correlation matrices is even more relevant, since an important aspect of risk management is the estimation of the correlations between the price movements of different assets. The probability of large losses for a certain portfolio or option book is dominated by correlated moves of its different constituents—for example, a position which is simultaneously long in stocks and short in bonds will be risky because stocks and bonds usually move in opposite directions in crisis periods. The study of correlation (or covariance) matrices thus has a long history in finance and is one of the cornerstone of Markowitz's theory of optimal portfolios [1,2]: given a set of financial assets characterized by their average return and risk, what is the optimal weight of each asset, such that the overall portfolio provides the best return for a fixed level of risk, or conversely, the smallest risk for a given overall return?

More precisely, the average return R_P of a portfolio P of N assets, is defined as $R_P = \sum_{i=1}^{N} p_i R_i$, where p_i ($i = 1, \ldots, N$) is the amount of capital invested in the asset i, and $\{R_i\}$ are the expected returns of the individual assets. Similarly, the risk on a portfolio can be associated to the total variance $\sigma_P^2 = \sum_{i,j=1}^{N} p_i C_{ij} p_j$, where $\mathbf{C}$ is the covariance matrix. The optimal portfolio, which minimizes the risk for a given value of R_P, can easily be found introducing a Lagrange multiplier and leads to a linear problem where the matrix $\mathbf{C}$ has to be inverted. In particular, the composition of the least risky portfolio has a large weight on the eigenvectors of $\mathbf{C}$ with the smallest eigenvalues [1,2].

However, a reliable empirical determination of a correlation matrix turns out to be difficult. For a set of N different assets, the correlation matrix contains $N(N-1)/2$ entries, which must be determined from N time series of length T; if T is not very large compared to N, one should expect that the determination of the covariances is noisy and, therefore, that the empirical correlation matrix is to a large extent random. In this case, the structure of the matrix is dominated by measurement noise; therefore, one should be very careful when using this correlation matrix in applications. In particular, as we show below, the smallest eigenvalues of this matrix are the most sensitive to this "noise," the corresponding eigenvectors being precisely the ones that determine the least risky portfolios. It is thus important to devise methods which allow one to distinguish "signal" from noise, i.e., eigenvectors and eigenvalues of the correlation matrix containing real information (which one would like to include for risk control) from those which are devoid of any useful information, and, as such, unstable in time. From this point of view, it is interesting to compare the properties of an empirical correlation matrix $\mathbf{C}$ to a null hypothesis purely *random* matrix as one could obtain from a finite time series of strictly independent assets. Deviations from the random matrix case might then suggest the presence of true information. The theory of random matrices has a long history since the 1950s [3], and many results are known [4]. As shown below, these results are also of genuine interest in a financial context (see also [5,6]).

The empirical correlation matrix $\mathbf{C}$ is constructed from the time series of price changes $\delta x_i(t)$ (where i labels the asset and t the time) through the following equation (In the

following we assume that the average value of the δx's has been subtracted off, and that the δx's are rescaled to have a constant unit volatility $\sigma^2 = \langle \delta x_i^2 \rangle = 1$.):

$$\mathbf{C}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \delta x_i(t) \delta x_j(t) . \tag{1}$$

We can symbolically write Eq. (1) as $\mathbf{C} = 1/\mathrm{T}\,\mathbf{M}\,\mathbf{M}^{\mathrm{T}}$, where $\mathbf{M}$ is a $N \times T$ rectangular matrix, and $^{\mathrm{T}}$ denotes matrix transposition. The null hypothesis of independent assets, which we consider now, translates itself in the assumption that the coefficients $M_{it} = \delta x_i(t)$ are independent, identically distributed, random variables, the so-called random Wishart matrices or Laguerre ensemble of the random matrix theory [7,8]. (Note that even if the "true" correlation matrix $\mathbf{C}_{\mathrm{true}}$ is the identity matrix, its empirical determination from a finite time series will generate nontrivial eigenvectors and eigenvalues; see [7,9].) We will note $\rho_C(\lambda)$ the density of eigenvalues of $\mathbf{C}$, defined as

$$p_C(\lambda) = \frac{1}{N} \frac{dn(\lambda)}{d\lambda} , \tag{2}$$

where $n(\lambda)$ is the number of eigenvalues of $\mathbf{C}$ less than λ. Interestingly, if $\mathbf{M}$ is a $T \times N$ random matrix, $\rho_C(\lambda)$ is self-averaging and exactly known in the limit $N \to \infty$, $T \to \infty$ and $Q = T/N \geq 1$ fixed [7,9] and reads

$$\begin{aligned} \rho_C(\lambda) &= \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda} , \\ \lambda_{\min}^{\max} &= \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q}) , \end{aligned} \tag{3}$$

with $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, and where σ^2 is equal to the variance of the elements of $\mathbf{M}$ [9], equal to 1 with our normalization. In the limit $Q = 1$ the normalized eigenvalue density of the matrix $\mathbf{M}$ is the well-known Wigner semicircle law, and the corresponding distribution of the *square* of these eigenvalues (that is, the eigenvalues of $\mathbf{C}$) is then indeed given by (3) for $Q = 1$. The most important features predicted by Eq. (3) are as follows:

(i) The fact that the lower "edge" of the spectrum is strictly positive (except for $Q = 1$); there is therefore no eigenvalues between 0 and $\lambda_{\min}$. Near this edge, the density of eigenvalues exhibits a sharp maximum, except in the limit $Q = 1$ ($\lambda_{\min} = 0$), where it diverges as $\sim 1/\sqrt{\lambda}$.

(ii) The density of eigenvalues also vanishes above a certain upper edge $\lambda_{\max}$.

Note that the above results are valid only in the limit $N \to \infty$. For finite N, the singularities present at both edges are smoothed: the edges become somewhat blurred, with a small probability of finding eigenvalues above $\lambda_{\max}$ and below $\lambda_{\min}$, which goes to zero when N becomes large. The precise way in which these edges become sharp in the large N limit is actually known [10].

Now, we want to compare the empirical distribution of the eigenvalues of the correlation matrix of stocks corresponding to different markets with the theoretical prediction given by Eq. (3), based on the assumption that the correlation matrix is purely random. We have studied numerically the density of eigenvalues of the correlation matrix of $N = 406$ assets of the S&P 500, based on daily variations during the years 1991–1996, for a total of $T = 1309$ days (the corresponding value of Q is 3.22).

An immediate observation is that the highest eigenvalue λ_1 is 25 times larger than the predicted $\lambda_{\max}$—see Fig. 1 inset. The corresponding eigenvector is, as expected, the "market" itself; i.e., it has roughly equal components on all of the N stocks. The simplest "pure noise" hypothesis is therefore clearly inconsistent with the value of λ_1. A more reasonable idea is that the components of the correlation matrix which are orthogonal to the market is pure noise. This amounts to subtracting the contribution of $\lambda_{\max}$ from the nominal value $\sigma^2 = 1$, leading to $\sigma^2 = 1 - \lambda_{\max}/N = 0.85$. The corresponding fit of the empirical distribution is shown as a dotted line in Fig. 1. Several eigenvalues are still above $\lambda_{\max}$ and might contain some information, thereby reducing the variance of the effectively random part of the correlation matrix. One can therefore treat σ^2 as an adjustable parameter. The best fit is obtained for $\sigma^2 = 0.74$ and corresponds to the dark line in Fig. 1, which accounts quite satisfactorily for 94% of the spectrum, while the 6% highest eigenvalues still exceed the theoretical upper edge by a substantial amount. Note that still a better fit could be obtained by allowing for a slightly

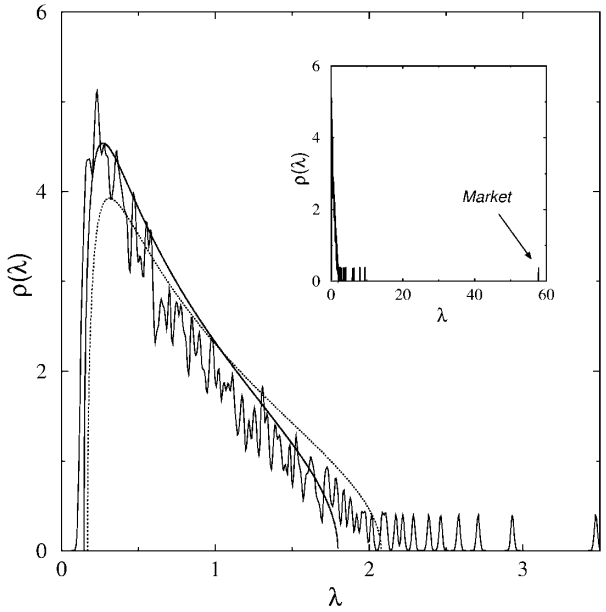

FIG. 1. Smoothed density of the eigenvalues of $\mathbf{C}$, where the correlation matrix $\mathbf{C}$ is extracted from $N = 406$ assets of the S&P 500 during the years 1991–1996. For comparison we have plotted the density Eq. (3) for $Q = 3.22$ and $\sigma^2 = 0.85$: this is the theoretical value obtained assuming that the matrix is purely random except for its highest eigenvalue (dotted line). A better fit can be obtained with a smaller value of $\sigma^2 = 0.74$ (solid line), corresponding to 74% of the total variance. Inset: Same plot, but including the highest eigenvalue corresponding to the market, which is found to be 25 times greater than $\lambda_{\max}$.

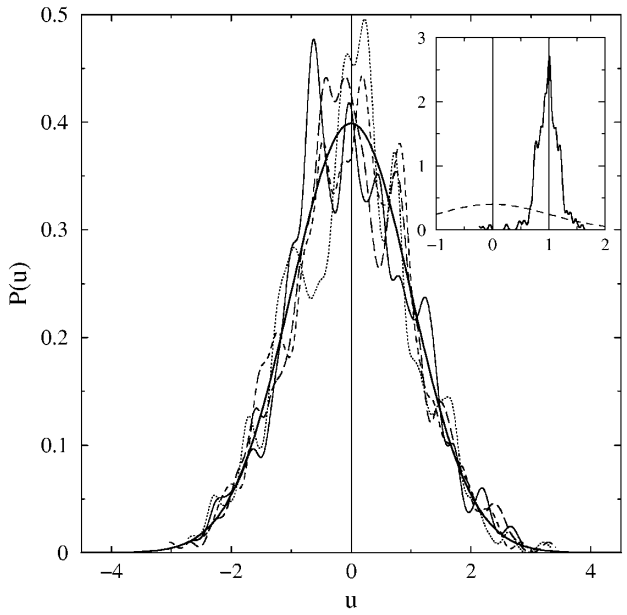

FIG. 2. Distribution of the eigenvector components $u = v_{\alpha,i}$, for five different eigenvectors well inside the interval $[\lambda_{\min}, \lambda_{\max}]$, and comparison with the no information assumption, Eq. (4). Note that there are *no* adjustable parameters. Inset: Plot of the same quantity for the highest eigenvalue, showing marked differences with the theoretical prediction (dashed line), which is indeed expected.

smaller effective value of Q, which could account for the existence of volatility correlations [11].

We have repeated the above analysis on different stock markets (e.g., Paris) and found very similar results. In a first approximation, the location of the theoretical edge, determined by fitting the part of the density which contains most of the eigenvalues, allows one to distinguish "information" from noise. However, a more precise procedure can be applied, where the finite N effects are adequately treated, using the results of [10], and where the effect of variability in σ^2 for the different assets can be addressed [9,11].

The idea that the low lying eigenvalues are essentially random can also be tested by studying the statistical structure of the corresponding *eigenvectors*. The ith component of the eigenvector corresponding to the eigenvalue λ_α will be denoted as $v_{\alpha,i}$. We can normalize it such that $\sum_{i=1}^{N} v_{\alpha,i}^2 = N$. If there is no information contained in the eigenvector $v_{\alpha,i}$, one expects that for a fixed α, the distribution of $u = v_{\alpha,i}$ (as i is varied) is a maximum entropy distribution, such that $\overline{u^2} = 1$. This leads to the so-called Porter-Thomas distribution in the theory of random matrices:

$$P(u) = \frac{1}{\sqrt{2\pi}} \exp -\frac{u^2}{2} . \qquad (4)$$

As shown in Fig. 2, this distribution fits extremely well the empirical histogram of the eigenvector components, except for those corresponding to the highest eigenvalues, which lie beyond the theoretical edge $\lambda_{\max}$. We show in the inset the distribution of u's for the highest eigenvalue, which is markedly different from the "no information" assumption, Eq. (4). This eigenvector is nearly uniform, which reflects that all assets are most affected by a common factor: the market itself. This is the tenet of the simple one-factor β model, much used in financial applications [1].

We have finally studied correlation matrices corresponding not to price variations but to the (time dependent) volatilities of the different stocks, determined from the study of intraday fluctuations. These matrices should contain some relevant information for option trading and hedging [12]. The obtained results are again very similar to those shown in Figs. 1 and 2.

To summarize, we have shown that results from the theory of random matrices are of great interest in understanding the statistical structure of the empirical correlation matrices. The central result of the present study is the remarkable agreement between the theoretical prediction and empirical data concerning both the density of eigenvalues and the structure of eigenvectors of the empirical correlation matrices corresponding to several major stock markets. Indeed, in the case of the S&P 500, 94% of the total number of eigenvalues fall in the region where the theoretical formula (3) applies. Hence, less than 6% of the eigenvectors, which are responsible for 26% of the total volatility, appear to carry some information. This method might be very useful to extract the relevant correlations between financial assets of various types, with interesting potential applications to risk management and portfolio optimization. It is clear from the present study that Markowitz's portfolio optimization scheme based on a purely historical determination of the correlation matrix is not adequate, since its lowest eigenvalues (dominating the smallest risk portfolio) are dominated by noise. In order to remove this bias, a better procedure would be to associate to each eigenvector with the "noise band" a constant eigenvalue, chosen such that the sum of the eigenvalues coincides with the trace of the original correlation matrix.

We want to thank J. P. Aguilar, M. Meyer, J. M. Lasry, and S. Galluccio for discussions.

*To whom correspondence should be sent.
Electronic address: laurent.laloux@science-finance.fr

[1] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (J. Wiley and Sons, New York, 1995); H. Markowitz, *Portfolio Selection: Efficient Diversification of Investments* (J. Wiley and Sons, New York, 1959).
[2] J. P. Bouchaud and M. Potters, *Theory of Financial Risk* (Aléa-Saclay, Eyrolles, Paris, 1997) (in French, English translation to be published).
[3] For a review, see O. Bohigas and M. J. Giannoni, *Mathematical and Computational Methods in Nuclear Physics*, Lecture Notes in Physics Vol. 209 (Springer-Verlag, Berlin, 1983).
[4] M. Mehta, *Random Matrices* (Academic Press, New York, 1995).
[5] S. Galluccio, J. P. Bouchaud, and M. Potters, Physica (Amsterdam) **259A**, 449 (1998).

[6] L. Laloux, P. Cizeau, J. P. Bouchaud, and M. Potters, *Random Matrix Theory*, Risk Magazine **12**, 69 (1999).
[7] A. Edelman, SIAM J. Matrix Anal. Appl. **9**, 543 (1988), and references therein.
[8] T. H. Baker, P. J. Forrester, and P. A. Pearce, J. Phys. A **31**, 6087 (1998).
[9] A. M. Sengupta and P. P. Mitra, e-print cond-mat/9709283, where the authors generalize the results of random Wishart matrices in the presence of multivariate correlations both in space and time.
[10] M. J. Bowick and E. Brézin, Phys. Lett. B **268**, 21 (1991); see also M. K. Sener and J. J. M. Verbaarschot, Phys. Rev. Lett. **81**, 248 (1998); J. Feinberg and A. Zee, J. Stat. Phys. **87**, 473 (1997).
[11] L. Laloux, P. Cizeau, J. P. Bouchaud, and M. Potters (to be published).
[12] For a review, see J. Hull, *Options, Futures and Other Derivative Securities* (Prentice-Hall, Englewood Cliffs, NJ, 1997). For a physicist approach, see [2].