Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 335 (2004) 629–643
PHYSICA A
www.elsevier.com/locate/physa

Random matrix theory for portfolio optimization: a stability approach

S. Sharifi, M. Crane, A. Shamaie, H. Ruskin*

Dublin City University, Glasnevin, Dublin 9, Ireland

Received 4 July 2003

Abstract

We apply random matrix theory (RMT) to an empirically measured financial correlation matrix, C, and show that this matrix contains a large amount of noise. In order to determine the sensitivity of the spectral properties of a random matrix to noise, we simulate a set of data and add different volumes of random noise. Having ascertained that the eigenspectrum is independent of the standard deviation of added noise, we use RMT to determine the noise percentage in a correlation matrix based on real data from S&P500. Eigenvalue and eigenvector analyses are applied and the experimental results for each of them are presented to identify qualitatively and quantitatively different spectral properties of the empirical correlation matrix to a random counterpart. Finally, we attempt to separate the noisy part from the non-noisy part of C. We apply an existing technique to cleaning C and then discuss its associated problems. We propose a technique of filtering C that has many advantages, from the stability point of view, over the existing method of cleaning.

Keywords: Random matrix theory; Portfolio optimization; Correlation matrix; Eigenvalues and eigenvectors

1. Introduction

Random matrix theory (RMT), originally developed for use in nuclear physics, has been described by, among others, Dyson in a series of papers beginning with Dyson [1] and subsequently in collaboration with Mehta (Mehta and Dyson [2], Dyson and Mehta [3], Mehta [4]) as the matrix representation of the average of all possible interactions in a nucleus. It can be used to identify non-random properties which are

* Corresponding author.
E-mail addresses: ssaba@computing.dcu.ie (S. Sharifi), mcrane@computing.dcu.ie (M. Crane), ashamaie@computing.dcu.ie (A. Shamaie), hruskin@computing.dcu.ie (H. Ruskin).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2003.12.016

deviations from the universal predictions of RMT; properties that are specific to the considered system. Close agreement between the distribution of the eigenvalues of a matrix M, with those from a matrix made up of random entries implies that M has entries that contain a considerable degree of randomness as has been shown in the literature [5–8]. This matrix consisting of random elements with unit variance and zero mean is called a random matrix [4]. In the case of a correlation matrix C, the level of agreement between its eigenvalue distribution and those of a random matrix, represents the amount of randomness (or noise) in C and thus, deviations from RMT represent genuine correlation (cf. [5–7]). This is precisely the problem that we wish to address, i.e., the identification of the true information (correlated assets) in a noisy financial correlation matrix. The method tests the null hypothesis that the distribution of eigenvalues of the correlation matrix is random. Since the correlation matrix is symmetric, the random matrix, with which it is compared, should also be symmetric [8].

Before applying the cleaning method to real correlation matrices, we need to determine the role of the *amount* of random noise on the spectral properties of a random matrix. This is done by examining the difference between the eigenspectrum of a correlation matrix made up of simulated data with different amounts of random noise and that of a random matrix. We can then proceed with confidence to examine the stability of real correlation matrices using real data. The empirical data set we use consists of 30-min intraday prices from the S&P500 Index from the beginning of April 1997 to the beginning of April 1999. This provides about 1500 data points for about 450 companies.

In this paper our initial objective, therefore, is to separate the noisy part from non-noisy part in C. Removal of the noise makes the optimization process more reliable, leaving the analyst in a better position to estimate the risk associated with the constructed portfolio. However, the techniques for removing noise from C should be considered carefully. A standard technique is initially applied to clean C but assessment of the results achieved reveal that it is not particularly satisfactory on the grounds of stability. We therefore, go on to discuss a filtering technique that takes account of the stability in a more precise way. Advantages of the new approach are validated by application to a financial data set from the S&P500.

2. Background and theory

2.1. Introduction: applications of RMT

Recently, a growing number of physicists have attempted to analyse and model financial markets. The history of this interest goes back to the work of Majorana [9] on analogies between statistical laws in physics and in the social sciences. In nuclear physics, the problem of understanding properties of matrices with random entries (which attempt to describe the interactions between sub-atomic particles) has a rich history (e.g. Wigner [10–12], Dyson [1], Mehta and Dyson [2], Dyson and Mehta [3] and Mehta [4]); in these, the assumption is made that the interactions between nuclear components are so complex that they can be taken as random. In

Ref. [10] the statistics of the eigenvalue properties of a random matrix were found to agree well with experimental data. In recent years authors such as Plerou et al. [6] have concentrated on comparing correlations between large movements of stocks and molecular motion in gases. RMT has also been applied to portfolio optimization (e.g. [13]), where co-movements among stocks are measured using the correlation (or, sometimes, covariance) matrix. Based on the results in Refs. [13,14], point to the inadequacies in Markowitz's Theory of Optimal Portfolios (cf. [15,16]). In Ref. [17], the authors introduce a technique (described below) to remove noise from the financial correlation matrix. For the resulting cleaned correlation matrix, Plerou et al. [7,8] and Mountfield and Ormerod [18] discuss the stability of the eigenvectors of the correlation matrix by examining their overlap over two consecutive temporal sub-periods. For those eigenvectors showing higher overlap over two sub-periods, stability is assumed to be higher and vice versa. The evidence suggests that the noisy part of C (as predicted by RMT) tends to show lower levels of stability.

2.2. Financial underpinning

Normally, the price changes (or return) of stocks are employed to quantify the empirical correlation matrix [19]. Therefore, we need to calculate the price changes of assets $i = 1,\ldots,N$ over a time scale Δt. For the price $S_i(t)$ of the ith asset at time t, one can define its "price change" or "return" $G_i(t)$ as

$$G_i(t) = \ln S_i(t + \Delta t) - \ln S_i(t)\ . \tag{1}$$

It should be noted that the terms "return" and "price changes" are sometimes used interchangeably (as in Ref. [19] for instance) but, strictly speaking, they are different. It is usual to define a normalised return to standardise the different stock volatilities. Therefore, we normalise G_i with respect to its standard deviation σ_i as follows:

$$g_i(t) = \frac{G_i(t) - \widehat{G_i(t)}}{\sigma_i}\ , \tag{2}$$

where σ_i is the standard deviation of G_i for assets $i = 1,\ldots,N$ and $\widehat{G}_i$ is the time average of G_i over the period studied. It should be noted that the asset risk is taken to be σ_i^2 since we assume, as do Bouchaud and Potters [14], that the risk and variance are the same (as is implied by assuming Gaussian assets). It can be shown (cf. Refs. [7,8,19]), that the correlation matrix C can be defined by

$$C = \frac{1}{T}\, GG^{\mathrm{t}}\ , \tag{3}$$

where G is an $N \times T$ matrix with elements $(g_i(m)\colon\ i = 1,\ldots,m; m = 0,\ldots,T-1)$ and $^{\mathrm{t}}$ denotes matrix transpose. The so-called *efficient frontier* (the boundary between the possible and impossible portfolios in a risk-return graph) is given by

$$\frac{\partial}{\partial p_i}\,(D_p - \zeta G_p)\bigg|_{p_i = p_i^*} = 0\ , \tag{4}$$

where p_i, p_i^* denote the asset weights and those corresponding to the optimal portfolio (represented by the efficient frontier), D_p, G_p are the mean risk and return for the

portfolio respectively and ζ is some parameter. Note that for correlated assets, $D_p = \sum_{i,j=1}^{N} p_i p_j C_{ij}$. The spectral properties of C may be compared with those of a random correlation matrix [5,13]. Subsequently, Plerou et al. [8] specifically, define a random correlation matrix as one which is the product of N time series of T random elements with zero mean and unit variance. Statistical properties of random matrices have been known for many years in physical literature [20–22]. In particular, under the restriction of $T \to \infty, N \to \infty$, providing that $q = T/N \geqslant 1$ is fixed, it was shown by these authors that the distribution of eigenvalues λ of the random correlation matrix is given by

$$P_r(\lambda) = \begin{cases} \left(\dfrac{q}{2\pi\sigma^2} \dfrac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda} \right), & \lambda_{\min} \leqslant \lambda \leqslant \lambda_{\max}, \\ 0, & \text{elsewhere}, \end{cases} \tag{5}$$

where σ^2 is the variance of the elements of G; (in the case of a normalised matrix G, it is therefore equal to unity), and $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of the correlation matrix respectively, given by

$$\lambda_{\max/\min} = \sigma^2 \left(1 + \frac{1}{q} \pm 2\sqrt{\frac{1}{q}} \right) . \tag{6}$$

These are the theoretical maximum and minimum eigenvalues, that determine the bounds of the theoretical eigenvalue distribution. If the eigenvalues of C are beyond these bounds it is said that they deviate from the random (or theoretical) bound. To evaluate the practical benefit of RMT, we apply it first to generated data Section 3.1 and on a set of real data Section 3.2.

2.3. Noise removal from correlation matrix

In order to separate the noisy from the non-noisy parts of C, we divide it into two parts: that which conforms to the properties of a random correlation matrix ("noise") and that which contains deviations from RMT predictions (the "information" part). In the first approximation [13], the location of the theoretical (or random) bounds, determined by the theoretical maximum and minimum eigenvalues, allows us to distinguish "information" from "noise". To separate the noisy and non-noisy parts of C we use the method of Bouchaud and Potters [14]. The whole idea is to obtain a background measure for the noise element while retaining the information trace; based on the fact that the eigenvalues corresponding to the noise band are *not* expected to contain real information, they are all equally useless. In effect, they suggest flattening (see Fig. 1) the noise part by replacing it with the identity matrix, while keeping the trace the same. We assess this idea in practice, using Bouchaud and Potters's [14] suggestion where the prediction of risk obtained using C_{noisy} is compared with that of C_{clean}. We divide the total available time period into two equal sub-periods, so that the data set from the first sub-period generates an estimate of the future return which can be compared with the actual return in the second sub-period. In the return in the second sub-period, we assume that the investor has "perfect" prediction on the future average returns i.e., we ignore random error over the second sub-period. The data set we used for RMT

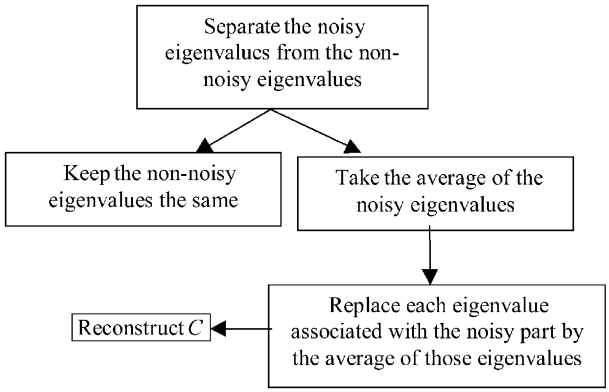

Fig. 1. Bouchaud and Potters's [14] procedure for cleaning and removing the noise from C.

prediction tests is also used for the rest of the experiments, but with the restriction that a smaller window of the set is considered. The reason for this restriction is the heavy computational overhead associated with the tests. First of all, we construct the correlation matrix using the first 600 data points for 200 stocks from S&P500 Intraday data. Next, we clean the matrix by following the procedure above (Fig. 1). Subsequently, we need to extract the optimal portfolios and efficient frontiers of both C_{noisy} (original C) and C_{clean}. Here, we restate the expressions *predicted risk* and *realised risk* [14], which will be used frequently in the remainder of this paper. The efficient frontier calculated using the return on the second sub-period and the correlation matrix for the first sub-period, is called *prediction* of the portfolio and the associated risk is called the *predicted* risk. Using this return and the correlation matrix, calculated using the second sub-period combined with the weights of the same family of portfolios as the predicted ones, we design another set of portfolios; known as the *realisation* of the portfolio [17]. The associated risk is denoted the *realised* risk. Bouchaud and Potters [14] argued that the predicted and realised risks are closer when the cleaned matrix is used in delineating the efficient frontier. They attribute the closeness of the mentioned curves to the power of C_{clean} in predicting the future risk and conclude that the stability of C_{clean} is higher than the stability of the original C. However (as we go on to show), we believe that, not only does this suggested technique not improve the stability of C, it actually reduces it.

2.4. Stability of the correlation matrix

The stability of the correlation matrix is an important aspect, e.g. Ref. [23], who has stated that eigenvectors and principal components can only be confidently interpreted if they are stable. The issue is thus how to determine the stability of C after cleaning it? In particular, we need to know if the stability of C_{clean} higher or lower. The aim clearly is to remove noisy elements from C in such a way that maximum stability is conserved. The majority of work in this area (e.g. [17,24]) indicates that the overlap

of the eigenvectors of two consecutive time sub-periods determines the consistency (or convergence) of the eigenvectors (the overlap of two vectors gives the amount of rotation of the second vector with respect to the first one). In the case where the vectors are normalised, the dot product of the vectors represents the cosine of the angle between them and gives a measure of the overlap. If the directions of the eigenvectors remain similar over the two sub-periods, then the cosine value should be large. Otherwise, it will be small.

The first eigenvectors, as argued earlier, are those which deviate from the random bound and provide most of the information. As expected, these eigenvectors have the highest stability and the degree of overlap is significant. To measure the stability of the matrix as a whole and its eigenvectors, we employ a principal component technique [25]. This examines the effect on v_k (the kth eigenvector) of small changes in the associated eigenvalue λ_k and argues that this is important because it gives information on the stability of the principal components. The principal components can only be securely interpreted if they are stable with respect to small changes in the values of the λ_k's. Specifically, the technique permits investigation of the perturbation of an eigenvector derived for a small reduction/increase, ε, in the corresponding eigenvalues. The component, $v(i)$ is determined, where this is the one that diverges the most from the ith eigenvector, v_i, but has an eigenvalue which is at most ε greater/less than that of v_i, such that the angle θ between $v(i)$ and v_i can be calculated by

$$\cos\theta = \begin{cases} \left(1 + \dfrac{\varepsilon}{\lambda_i - \lambda_{i+1}}\right)^{-1/2}, & \varepsilon \text{ decreased from } \lambda_i\,, \\ \left(1 + \dfrac{\varepsilon}{\lambda_{i-1} - \lambda_i}\right)^{-1/2}, & \varepsilon \text{ increased to } \lambda_i\,, \end{cases} \tag{7}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots \geqslant \lambda_n$. This equation demonstrates that the effect on v_i of an ε change in λ_i is an inverse function of $\lambda_i - \lambda_{i+1}$. Thus it is not the absolute size of the eigenvalue which determines whether that component is stable or not but rather its separation in terms of eigenvalue size from the next component. Relatively isolated (early) components with large eigenvalues should therefore be fairly stable, but later components, all of which have similar non-zero variances, will not be stable. So the largest non-zero eigenvalue and corresponding eigenvector can be used to find the smallest perturbation in v_i which leads to a change ε in λ_i.

3. Data: simulated and real

3.1. Simulated data

In order to see the results of RMT (and in particular the influence on the spectral properties of a random matrix of the amount of added random noise) on a correlation matrix made up of simulated data we proceeded as follows. A set of 450 sinusoidal time series with 1500 observations (i.e., the same size as our real data set) is generated with random amplitude and random phase. Next, we add random noise normally distributed

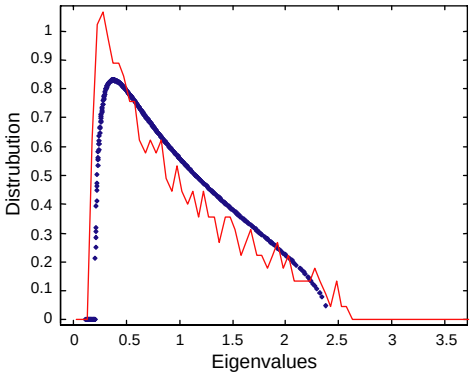

Fig. 2. Eigenspectrum of C from generated data (dotted line) and theoretical distribution (diamonds).

with zero mean and a given standard deviation to the time series $Z \approx N(0, \sigma^2)$. The standard deviation σ is chosen to be some value fixed for a given simulation. We control the amount of noise added to the generated data by varying σ. We study the behaviour of the correlation matrix constructed from the noisy time series by applying RMT in order to see the effect of different volumes of noise in the time series on the noise content in the correlation matrix. First by using Eq. (3), the correlation matrix C is constructed. Since the number of observations is $T = 1500$ and the number of time series is $N = 450$ the inequality $q = T/N \geqslant 1$ is satisfied. Therefore, we can apply the RMT to our generated data and plot the distribution of the eigenvalues of C. Using Eq. (5), the theoretical distribution of the eigenvalues of the correlation matrix is calculated for a given value of added noise. The actual (empirical) distribution of the eigenvalues of C is also calculated and together with the theoretical one is shown in Fig. 2 for a single value of added noise. Some eigenvalues in the empirical set deviate from the theoretical graph; these are non-noisy eigenvalues corresponding to that part of the correlation matrix which has real information. Interestingly, we have observed that the variance of added noise has little or no effect on the distribution of the eigenvalues. This shows that the amount of meaningful information to be gained from C is independent of the standard deviation of the added noise with the implication that RMT can be used with confidence with the real data.

3.2. Real data

The data relate to more than 450 companies (N is sufficiently large), and over 1500 observations ($T \to \infty$). Firstly, we construct the empirically measured correlation matrix, C, by using Eq. (3) as in the previous section, and then compute the eigenvalues λ_k where $k = 1, \ldots, N$ is in ascending order.

3.2.1. Eigenvalue analysis

The distribution of the eigenvalues of the corresponding random correlation matrix is also calculated using Eq. (5). Fig. 3 shows the results of our experiments on the

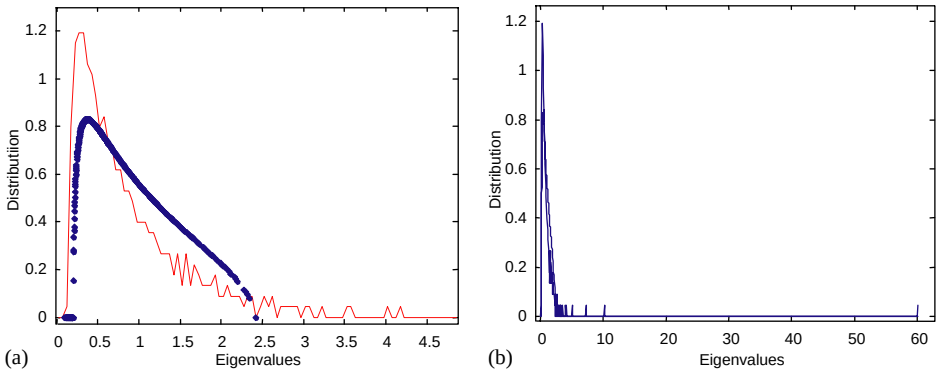

Fig. 3. Eigenvalue distribution for the correlation matrix C from S&P500 data from 4/97 to 4/99 (diamonds indicate eigenvalues from RMT): (a) partial spectral distribution, (b) full spectral distribution (a).

30-min data for 452 stocks and 1500 records. As with the results of RMT on the simulated data, we can observe two things from Fig. 3.

(1) The bulk of the eigenvalues of C conform to those of the random matrix Fig. 2. This means that there is a measure of randomness in the bulk majority of the eigenvalues. Therefore, in agreement with Laloux et al. [13], we conclude that the corresponding eigenvalues are random and we take this part of C as the noisy band.
(2) Fig. 3 (inset) illustrates the same quantity but on different scales, with deviations from RMT for a small number of the largest eigenvalues visible. Our experiments indicate that 22 eigenvalues are outside the noise band and the rest can be seen to be consistent with RMT results. In other words, just 4.7% of the eigenvalues deviate from the RMT prediction.

Again, this is in agreement with Laloux et al. [13] who argue that at most 6% of the eigenvalues are non-noisy. Thus, less than 5% of the eigenvalues appear to carry most of the information, as found in Refs. [7,8,13]. In the next section, the noise and information content of the correlation matrix is examined qualitatively using eigenvector analysis.

3.2.2. Eigenvector analysis

The eigenvector analysis looks at the structure of the eigenvectors and compares the eigenvector component distribution with those of the corresponding random matrix. The eigenvector components of the random matrix are normally distributed with zero mean [13], so the expectation is that the eigenvectors corresponding to the noise band of the correlation matrix follow a similar distribution. Fig. 4 represents the distribution of the eigenvector components corresponding to our empirical correlation matrix. The eigenvectors associated with the largest eigenvalue, and some of the smaller ones are shown. It is seen that the distribution of the corresponding "market" eigenvector, in

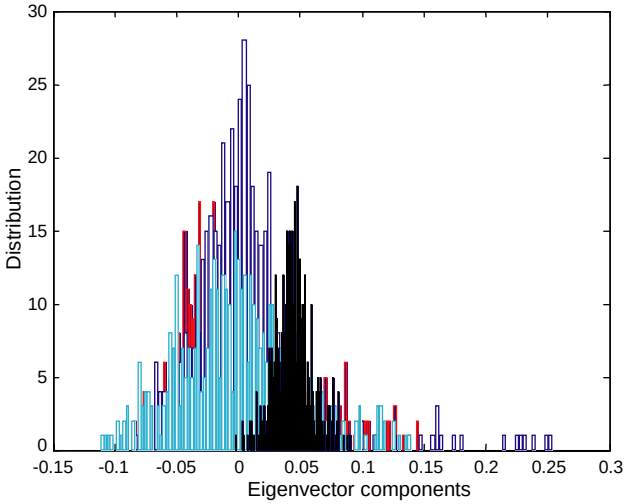

Fig. 4. Eigenvector components of C (largest in black).

black, does not follow the same structure as the others. The components of the market eigenvector are distributed around a mean of 0.045 and a variance of 0.05, whereas the other eigenvector components are distributed with zero mean and a much wider variance. In fact, the dispersion of the components around the mean increases as one examines eigenvectors associated with progressively smaller eigenvalues. Although this eigenvector analysis is not as precise as the eigenvalue analysis above, it suggests that the market eigenvector does not behave similarly to the eigenvectors of the random matrix, and therefore implicitly, it represents the most information-rich as well as reliable part of the correlation matrix.

4. Optimal portfolios: a stability approach

4.1. Stability of the correlation matrix

In this section we study the empirical correlation matrix from a stability point of view. As mentioned above, various authors (e.g. [17,23]) have studied eigenvector convergence in terms of their overlap. We compare the overlap of eigenvectors over two consecutive sub-periods. The first sub-period is the first 600 records of our S&P500 data for 200 stocks and the second is the next 600 records. Fig. 5 shows that after the initial few eigenvectors (corresponding to the largest eigenvalues), the overlap (measured by $\cos\theta$) falls into the noise level [17,26] indicated by the line $1/\sqrt{N}$, for N eigenvectors. This indicates again the presence of a large quantity of noise in the correlation matrix. In this section we study the stability of C using the Krzanowski [25] model, Eq. (7). The angle between the eigenvector i of the original correlation matrix C and $v(i)$ is calculated; where $v(i)$ is the perturbation of the ith eigenvector

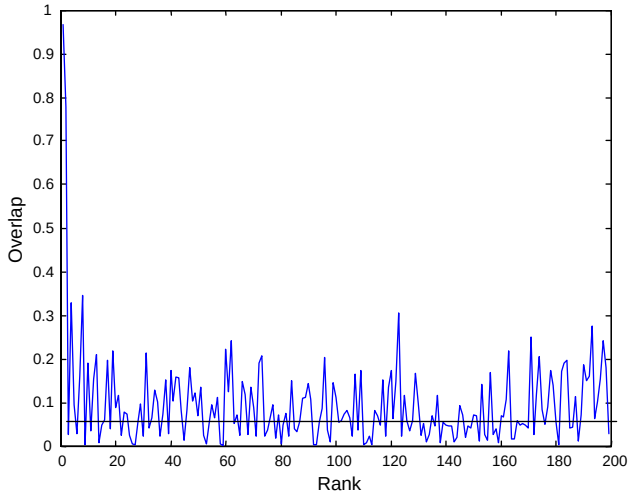

Fig. 5. Eigenvector overlap between the two sub-periods. Horizontal line is the noise level.

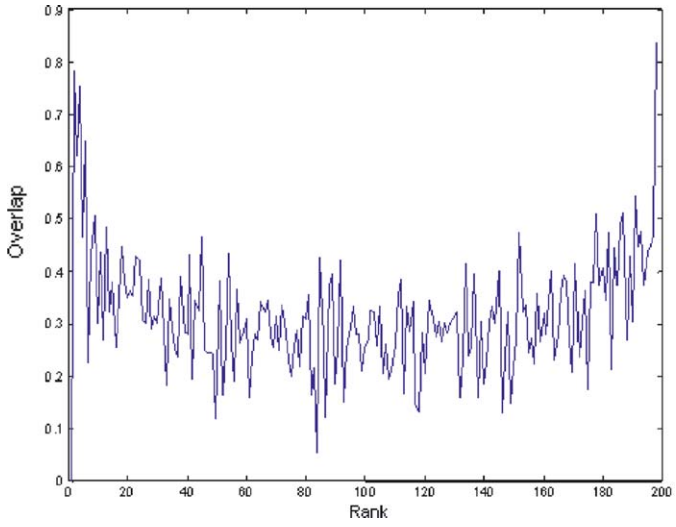

Fig. 6. Overlap ($\cos\theta$) versus rank for C, θ represents the perturbation of v_i for a 0.2% change in λ_i.

derived for a small change ε in λ_i (Fig. 6). Thus ε is determined by the empirical changes in the average value of λ_i from the first sub-period to the second sub-period and approximates to 0.2% in our experiments.

As expected, Fig. 6 demonstrates that the largest eigenvectors are the most stable ones (i.e., $\cos\theta$ large). Further, the last (smallest) eigenvectors show higher stability than those in the middle since the former approach zero and consequently, $\cos\theta$ in

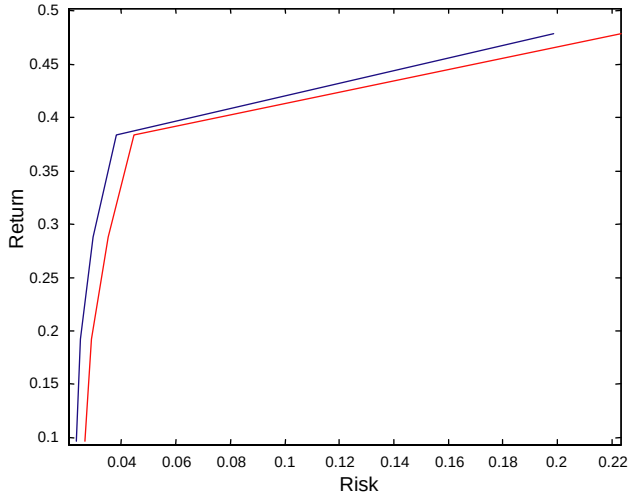

Fig. 7. Efficient frontiers optimal portfolios from the original matrix C: top curve gives predicted risk and bottom curve shows realised risk.

Eq. (7) is larger. Examining Fig. 7 (which shows the predicted and realised risk of the family of optimal portfolios calculated using 30-min returns from 01/04/1997 to 01/04/99), it can be seen that the top end of the predicted and realised curves are further apart whereas in the lower and middle areas are closer. Since the area of the efficient frontier associated with the highest risk (Fig. 7, top area) corresponds to the largest eigenvalues (the most stable ones), then we can conclude that, as stability decreases, the curves get progressively closer. This contradicts the conclusion of Bouchaud and Potters [14] who attribute the closeness of these curves to higher stability. We discuss this in greater detail in the next section which deals with the stability of the *cleaned* correlation matrix.

4.2. Stability of the cleaned correlation matrix C_{clean}

We examine the stability of C_{clean} using the method of Bouchaud and Potters [14] for cleaning. It can be seen from Fig. 8(a) that the stability (measured by their overlap) of the eigenvectors of C_{clean} has declined noticeably to its lowest position after the 11th eigenvector (the edge of noisy and non-noisy components as determined by RMT prediction). This reflects the fact that the noisy band of eigenvalues is replaced by their average, which means no separations between eigenvalues at all. Fig. 8 also illustrates the negative relationship between stability of eigenvectors (and therefore C) and the distance between the predicted and realised curves. As the stability increases, the distance between the two curves decreases. The implication is important: the cleaning method of Bouchaud and Potters [14] seems to impact adversely on the stability of the eigenvectors after the initial few; a new method described below not only maintains the stability for low rank, but also increases that in the middle and high rank as well.

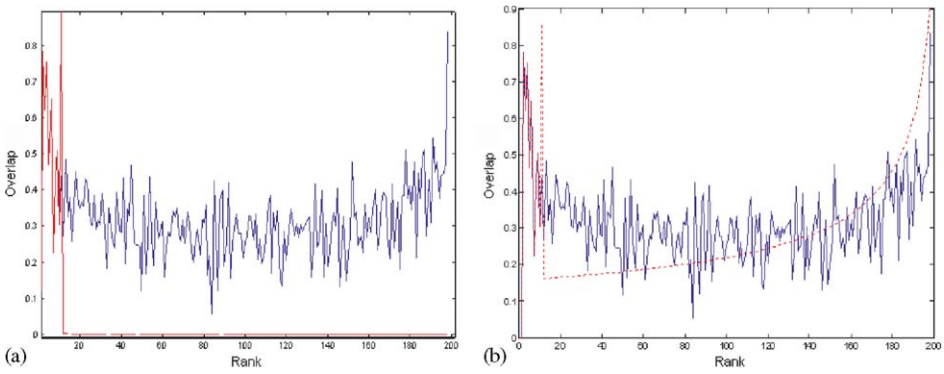

Fig. 8. Overlap ($\cos\theta$) versus rank for the original S&P500 data correlation matrix, C, and C_{clean} using (a) Bouchaud and Potters [14] and (b) Krzanowski [25] techniques.

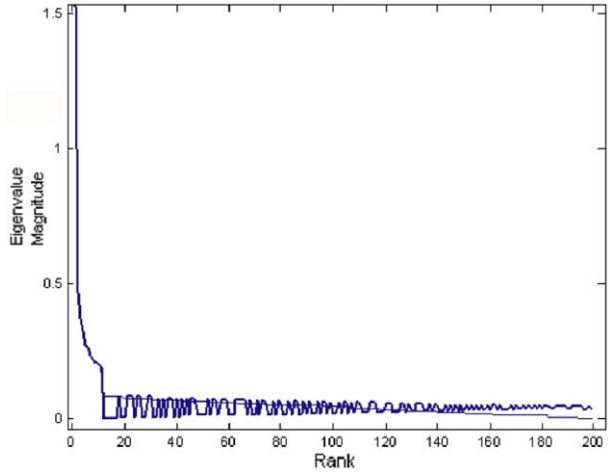

Fig. 9. Eigenvalue magnitude versus rank for C_{clean} (solid line) and C (dashed line).

4.3. A new approach to matrix filtering

We propose a new method of filtering C based on the work of Krzanowski [25] to preserve the stability of the matrix as much as possible. The principle is to replace the noisy eigenvalues with components that have maximal separation from each other while maintaining a fixed sum. In Fig. 9, the noisy part of the graph is changed to an oblique line. These eigenvalues are indicated by a dotted line. The slope is determined so that on one hand the most separation between components is attained and on the other hand none of the eigenvalues is replaced by negative values (as all the eigenvalues of the correlation matrix are positive). The matrix C_{clean} is reconstructed

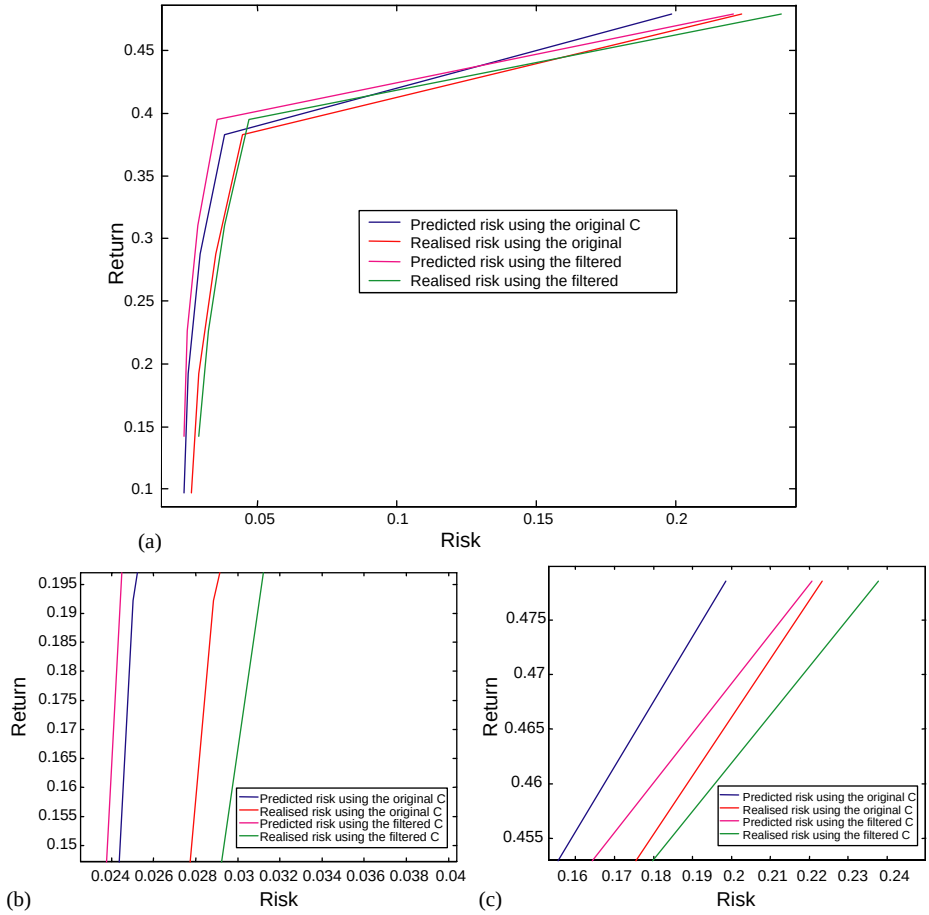

Fig. 10. (a) Efficient frontier for optimal portfolios from original C (red and blue) and filtered C (green and magenta); (b) Low risk/low return area of (a); (c) High risk/high return area of (a).

using $C_{clean}=VD_{new}V^{t}$ (V is the matrix of eigenvectors, D_{new} the reconstructed diagonal matrix of eigenvalues given by the oblique line in Fig. 9) and the resulting eigenvalue distribution of C_{clean} is shown by a solid line in Fig. 9.

To observe the stability of this new C_{clean} we compute $\cos\theta$ (in Eq. (7)) again (see Fig. 8(b)). As can be seen, the stability of noisy eigenvectors of the original matrix is higher than those of the filtered matrix up to approximately 150th eigenvector and is lower thereafter. Clearly, this method gives higher stability in comparison with flattening the eigenvalues [14]. In this context, we expect that the predicted and realised risk-return curves are closer in the interval between 11th and 150th in comparison to those of the original C. Again the largest eigenvalues correspond to the *riskiest* portfolios exposed in the top area of the efficient frontier in Fig. 10(c). Equally the smallest eigenvalues correspond to the least risky portfolios (lower area of the efficient

frontier Fig. 10(b)). As can be seen, distance between the upper ends of the curves decreases whereas that at the central part is greater than that for the original C matrix. This supports our assertion that stability is inversely related to the distance between the predicted and realised risks. As the stability increases, the distance between curves decreases and vice versa.

We conclude from this and the results in Fig. 8 that our method is superior to that of Bouchaud and Potters [14] from the point of view of impact of cleaning a correlation matrix on the stability of the eigenvectors. We also note that the closeness of the predicted and realised curves does not necessarily represent the strength of future risk prediction. Indeed, when the correlation matrix is less stable the predicted and realised curves are closer than the case with more stability.

5. Conclusions

We have applied RMT to determine the noise in an empirically measured correlation matrix, C. As a preliminary, we examined RMT results on simulated data with varying volumes of noise. The independence of the number of deviated eigenvalues from the volume of added noise implies that results of RMT are also independent of the amount of noise in the data. For a set of actual data from S&P500 we deduced that less than 5% of the eigenvalues carry useful information with the rest reflecting noise. This is in agreement with previous work of Bouchaud and Potters [14] which indicates that at most 6% of C is information. These results were obtained principally by eigenvalue analysis and confirmed in outline by complementary eigenvector analysis which indicated that the market eigenvector (the eigenvector corresponding to the largest eigenvalue) has a different construction to the other eigenvectors, implying that most information in C is measured by this quantity. Finally we have examined the well-known and commonly used technique for noise removal from a correlation matrix [14]. But, however, find that it decreases the level of stability of C. We have alternatively applied the Krzanowski [25] model to study the stability of the financial correlation matrix after removal of noise and conclude that this offers real improvement on the usual method. The improvement was tested by comparing the realised and the predicted optimal portfolios [14], with expectation of a shorter distance between the realised and the predicted risk for C_{clean} than that of the original C (attributed by the authors to the higher stability of C_{clean}). We show that this is not the case and in fact there is a negative relationship between the stability of C and the closeness of the predicted and realised risks. This assertion is also demonstrated by the experiments of filtering C, based on the Krzanowski [25] model. The commonly used technique of noise removal not only fails to assure stability, but can actually lead to a considerable deterioration. This finding offers valuable insight for portfolio optimisation.

References

[1] F. Dyson, Statistical theory of the energy levels of complex systems. i–iii, J. Math. Phys. 3 (1962) 140–175.

[2] M. Mehta, F. Dyson, Statistical theory of the energy levels of complex systems iv, J. Math. Phys. 4 (5) (1963) 713–719.
[3] F. Dyson, M. Mehta, Statistical theory of the energy levels of complex systems. v, J. Math. Phys. 4 (5) (1963) 701–712.
[4] M. Mehta, Random Matrices, Academic Press, New York, 1991.
[5] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, H.E. Stanley, Universal and non-universal properties of cross-correlations in financial time series, Phys. Rev. Lett. 83 (7) (1999) 1471–1474.
[6] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, H.E. Stanley, Econophysics: financial time series from a statistical point of view, Physica A 279 (2000) 443–456.
[7] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, H.E. Stanley, A random matrix theory approach to financial cross-correlations, Physica A 287 (2000) 374–382.
[8] V. Plerou, P. Gopikrishnan, B. Rosenow, L. Amaral, T. Guhr, H.E. Stanley, Collective behaviour of stock price movements; a random matrix approach, Physica A 299 (2001) 175–180.
[9] E. Majorana, Il valore elle leggi statistiche nella fisica e nelle scienze sociali, Scientia 36 (1942) 58–66.
[10] E. Wigner, On a class of analytic functions from the quantum theory of collisions, Ann. Math. 53 (1951) 36–67.
[11] E. Wigner, On the statistical distribution of the widths and spacings of nuclear resonance levels, Proc. Cambridge Philos. Soc. 47 (1951) 790–798.
[12] E. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. Math. 62 (1955) 548–564.
[13] L. Laloux, P. Cizeau, J. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. 83 (7) (1999) 1467–1470.
[14] J. Bouchaud, M. Potters, Theory of Financial Risks—From Statistical Physics to Risk Management, Cambridge University Press, UK, 2000.
[15] H. Markowitz, Portfolio Selection: Efficient Diversification of Investments, Wiley, New York, 1959.
[16] E. Elton, M. Gruber, Modern Portfolio Theory and Investment Analysis, Wiley, New York, 1981.
[17] L. Laloux, P. Cizeau, M. Potters, J. Bouchaud, Random matrix theory and financial correlations, Int. J. Theor. Appl. Finance 3 (3) (2000) 391–397.
[18] C. Mountfield, P. Ormerod, Market correlation and market volatility in US Blue Chip stocks, Volterra Consulting Ltd., Technical Report, UK, 2001.
[19] V. Plerou, P. Gopikrishnan, L. Amaral, H.E. Stanley, Collective behaviour of stock prices—a random matrix theory approach, Physica A 299 (2001) 175–180.
[20] F. Dyson, Distribution of eigenvalues for a class of real symmetric matrices, Rev. Mex. Fis. 20 (1971) 231–237.
[21] A. Edelman, Eigenvalues and condition numbers of random matrices, SIAM J. Matrix. Anal. App. 9 (1988) 543–560.
[22] A. Sengupta, P. Mitra, Distributions of singular values for some random matrices, Phys. Rev. E 60 (3) (1999) 3389–3392.
[23] I. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[24] Y. Lee, Noise detection from financial correlation matrices, Academic Report, Massachusetts Institute of Technology, US, 2001.
[25] W. Krzanowski, Sensitivity of principal components, J. Royal Stats. Soc. B 46 (3) (1984) 558–563.
[26] S. Strongin, M. Petsch, G. Sharenow, Beating benchmarks, a stockpicker's reality, J. Portfol. Manage. 26 (2000) 11–27.