Seeking Gold in Sand: financial applications of Random Matrix Theory in stock market data

Mike Shengbo Wang*

Faculty of Mathematics & Department of Chemistry, University of Cambridge

November, 2016

Abstract

Covariance-correlation matrix estimation is central to modern portfolio theory; in this project, we take a Random Matrix Theory based approach to compare a proposed multi-layer structured correlation model, constructed using mode and clustering analyses, with the observed spectrum of the empirical correlation matrix of S&P 500 stock market data. We will analyse the dependence on layer depth of the constructed model, and obtain an accurate match between predicted and observed empirical correlation matrix spectral distributions.

1 Introduction

Covariance-correlation matrix is of fundamental importance wherever large data sets are involved and relations between many random variables are to be understood. In modern portfolio theory, accurate estimation of covariance-correlation is crucial to risk management and asset allocation [1], as correlations measure the tendency of collective movement of different stocks, and underpin the interactions between them.

In this project we will consider a Random Matrix Theory based approach to testing a proposed multi-layer structured correlation model, constructed with information extracted through mode and clustering analyses.

The following section gives an overview of the data analysed and its processing. In Section 3 we will introduce the Marčenko-Pastur law in Random Matrix Theory; in Sections 4 and 5 we will consider the features of our data using mode and hierarchical clustering analyses; then in Section 7 we propose a multi-layered structure in our correlation model, with the effect of layer depth analysed in Section 8; finally, in Section 10, we will briefly discuss possible developments to this project.

All computational work in this project has been carried out in MATLAB. The Live Scripts are published on the author's web-page [2].

*This project is part of the Summer Undergraduate Research Opportunities Programme (SUROP), and is supported by the Bridgwater Scheme.

2 Data Overview

The S&P 500 stock market data[1] studied are stored as a matrix whose rows represent the trading days and columns the different stocks. We first calculate the *logarithmic returns* for all consecutive trading days for each stock,

$$\text{log return} = \log \frac{p_i}{p_{i-1}} (\approx \frac{p_i - p_{i-1}}{p_{i-1}}), \quad i > 1$$

where p_i represent the price index of a stock on the i-th trading day.

This leaves us a matrix of $T = 1258$ rows of observations and $P = 452$ columns corresponding to each individual stock. We will demean each column by subtracting the column average and normalise the entries so that the total variance for any stock is one. The data matrix $X : T \times P$ is now standardised, and the *empirical* (*covariance-*)*correlation matrix* is simply

$$E = \frac{1}{T} X^T X. \tag{1}$$

We will denote the *underlying*, or *true*, *correlation matrix* by C – this is the object that we will attempt to model based on the spectrum of E.

3 Random Matrix Theory: The Marčenko-Pastur Law

Random Matrix Theory (RMT) was first introduced by John Wishart in 1928, who was the first director at the Statistical Laboratory at the University of Cambridge. It became a prominent field of study when the physicist Eugene Wigner applied it to spacing of energy levels in nuclear physics [3].

3.1 Statement of the Marčenko-Pastur law

The theory has a collection of universality laws, since it concerns the emergent behaviours of large classes of random matrices in the *asymptotic limit*, that is to say, when the dimensions of the matrix tend to infinity. One important instance of these is the Marčenko-Pastur law for a class of random matrices called the *Wishart ensemble*, which include all correlation matrices:

Theorem 1 (The Marčenko-Pastur law)**.** *If X is a $T \times P$ random matrix whose entries are independently identically distributed (i.i.d.) random variables (r.v.'s) with mean 0 and variance $\sigma^2 < \infty$, then the eigenvalue density function (e.d.f) of matrix (1) is*

$$f(\lambda) = \frac{1}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{r\lambda} \tag{2}$$

in the limit $P, T \to \infty$ and $P/T \to r \in (0, 1)$, where $\lambda_\pm = \sigma^2(1 \pm \sqrt{r})^2$.

[1] See the author's webpage [2].

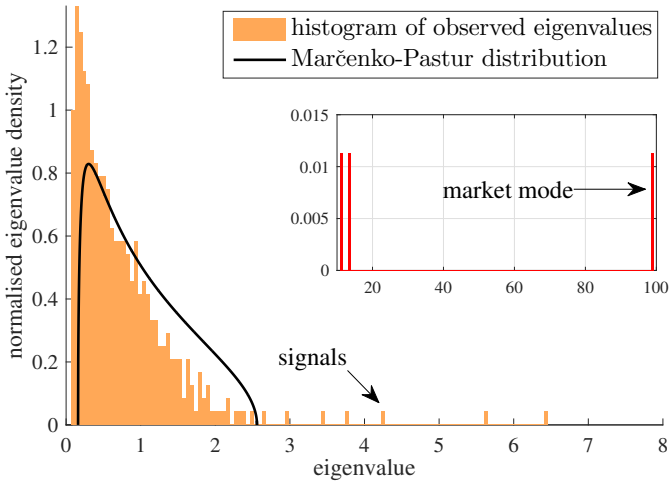


Figure 1: The Marčenko-Pastur distribution does not match the empirical eigenvalue distribution of the our S&P 500 stock market correlations. The eigenvalues lying outside the prediction range are regarded as signals that our stock market data are not purely random.

3.2 Interpretation of the Marčenko-Pastur law

For our standardised data, variance $\sigma^2 = 1$. The key parameter of the Marčenko-Pastur distribution is then the *concentration* $r = P/T$, which intuitively represents the abundance of observations compared to the number of variables. An interpretation of the Marčenko-Pastur law in our context is that if the logarithmic returns of our stocks are independently, identically distributed, i.e. totally random, then *regardless of the underlying distribution*[2], the observed eigenvalue distribution of E is governed by (2).

This result provides a natural test for the null hypothesis that the data are completely random: if we plot the observed eigenvalue density function against the Marčenko-Pastur distribution, any eigenvalues that lie far out from the Marčenko-Pastur prediction can be regarded as signals, suggesting that the data are not truly random. In Figure 1, we see that the Marčenko-Pastur distribution is far from a match to our observed empirical eigenvalue distribution. There are many signals above and below the edges of the Marčenko-Pastur prediction, all suggesting our S&P 500 stock market data are not purely random.

However, this is not at all surprising. We know that in reality many stocks are related and the market structure is entangled and complex. Our aim is to construct a better model for the underlying correlation matrix C, and then use the techniques in RMT to derive a new prediction for the limiting empirical eigenvalue distribution when $P, T \to \infty$, which improves the detection for any new signals.

4 Mode Analysis

The eigenvalue and eigenvector pairs of the empirical correlation matrix E are referred to as the *modes* of the market. They give insight into the interactions between individual stocks as well as market sectors.

One feature of these modes is the localisation of the eigenvector components, conveniently mea-

[2] As long as its first and second moments are bounded, which is a reasonable assumption.

sured by the *inverse participation ratio*

$$\mathrm{IPR}(\mathbf{v}) = \sum_{i=1}^{P} |\tilde{v}_i|^4$$

where $\tilde{\mathbf{v}}$ is the vector $\mathbf{v}$ demeaned and normalised. In Figure 2 we have shown the components of the market mode corresponding to the largest eigenvalue, the 9-th mode and the lowest mode corresponding to the least eigenvalue. We see that the market mode has a low IPR, which means in this mode all stocks move in a similar fashion, responding to the overall trend of the market (and hence its name). The 8th mode is more localised, and we can differentiate the edges of the market sectors. The lowest mode is highly localised, and the interaction between two companies in the energy and industrial sectors are clearly visible.

These different types of modes suggest there are correlation interactions at the stock, sector *and* market levels, so a layered model may be appropriate to capture such interactions.

4.1 Digression: uniformity of the market mode

It is observed that components of the market mode eigenvector are relatively uniform, and more crucially, have the same sign. The proposition below may explain this.

Proposition 1. *If A be a positive square matrix, then*

1) *it has a positive real eigenvalue λ_1 with multiplicity 1 that has the largest magnitude of all its eigenvalues;*
2) *it has a unique positive unit eigenvector and it corresponds to λ_1. All other eigenvectors must have at least one negative or non-real component.*

Proof. Let $\mathbf{e}_1$ be the unit eigenvector for the largest eigenvalue λ_1 of A. Then

$$\mathbf{e}_1 = \arg\max_{\|\mathbf{x}\|=1} \mathbf{x}^T A \mathbf{x}.$$

By reordering the basis we can without loss of generality (w.l.o.g.) assume that the first k components of $\mathbf{e}_1$ are negative, and the rest are positive, where $1 < k < n$ and n is the dimension of matrix A. Hence

$$\mathbf{e}_1^T A \mathbf{e}_1 = \sum_{i \le k} \sum_{j \le k} (\mathbf{e}_1)_i A_{ij} (\mathbf{e}_1)_j + \sum_{i > k} \sum_{j > k} (\mathbf{e}_1)_i A_{ij} (\mathbf{e}_1)_j.$$

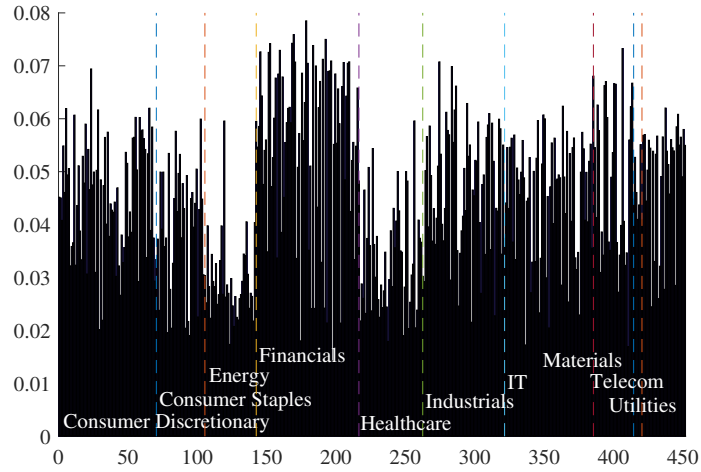
However, switching the signs of $(\mathbf{e}_1)_{i \le k}$ preserves the norm while increasing the sum above, so by *reductio ad absurdum*, the non-zero components of $\mathbf{e}_1$ must be of the same sign. In fact, $\mathbf{e}_1$ cannot have a zero component by the consistency condition $A\mathbf{e}_1 = \lambda_1 \mathbf{e}_1$.

By orthogonality of eigenvectors belonging to different eigen-subspaces, the other eigenvectors must have components of mixed signs. □

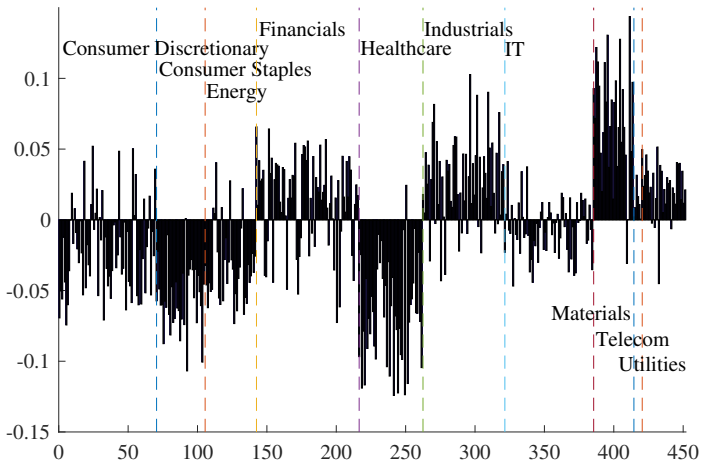
5 Hierarchical Clustering Analysis

If we remove the market mode from the correlation matrix E, i.e. compute the 'modified' correlation matrix

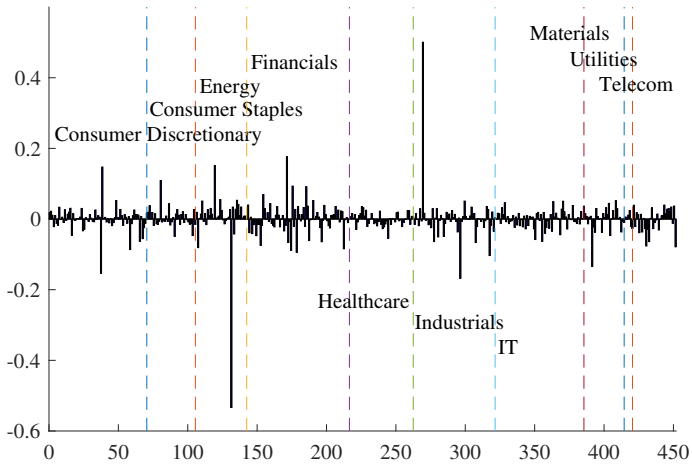
$$E' = E - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T \tag{3}$$



(a) The market mode with eigenvalue 99.1 and IPR 3.98×10^{-5}.

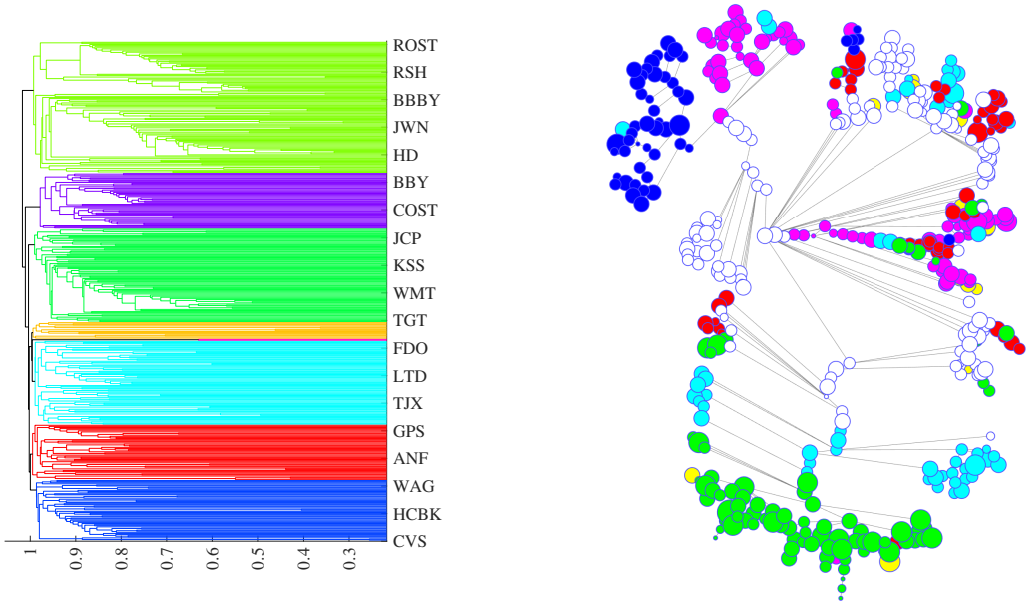


(b) The 8th mode with eigenvalue 3.46 and IPR 6.51×10^{-3}.



(c) The lowest mode with eigenvalue 0.0596 and IPR 0.149.

Figure 2: Plots of the eigenvector components of three different modes with their IPRs calculated.



(a) Dendrogram

(b) Minimum spanning tree

Figure 3: Two visualisations of the stock market structure and relations without the market mode. The minimum spanning tree stock nodes are coloured by market sector.

where $\lambda_1, \mathbf{v}_1$ are the largest eigenvalue and eigenvector, then hierarchical clustering analysis may reveal hidden market structure and relations under the overall market movement.

To perform clustering we must specify a distance measure; a natural choice is the *dissimilarity distance*, defined by

$$d_{ij} = 1 - \text{corr}(i, j) \tag{4}$$

where $\text{corr}(i, j)$ is the correlation between stocks i and j. Since correlations are reflexive and always between -1 and 1, this meets the criteria of a *metric*[3]. It is also a convenient choice as d_{ij} is linear in the correlations.

We will here adopt the *average linkage method*, which means the distance between two clusters I, J is the average distance between all pairs of stocks from the two sector

$$D_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} d_{ij} \tag{5}$$

where $|\cdot|$ is the order of a cluster set.

Performing the clustering the analysis in MATLAB generates a dendrogram shown in Figure 3, alongside a minimal spanning tree. These two visualisations of the stock market structure complement each other. The MATLAB output provides information about clusters which we will use to construct our multi-layer structured correlation model in the following section.

6 Re-classification of Market Sectors

Our clustering analysis has given us a new way of defining different market sectors based on the average dissimilarity distances. We have found that if we do not remove the market mode, the average linkage method produces a large number of singleton clusters. To avoid this, we have removed

[3]The triangle inequality can be easily checked.

the market mode before performing clustering. We then redo the mode analysis in Section 4 to find whether this new classification of market sectors is satisfactory. The results are presented in Figure 4.

We see that the new classification of the market sectors unfortunately does not have clear boundaries as we have seen in Figure 2. This is due to the removal of the market mode: despite being reasonably uniform, the market mode still contains a substantial amount of structural information about the market, as we could already differentiate the sectors in the original market mode plot in Figure 2.

7 A Multi-layer Structured Correlation Model and Its Predictions

There is a major caveat here: although we have removed the market mode in our clustering analysis, to construct the multi-layer structured correlation model we will restore it. This is because, as could be seen in Figure 4a, the market mode is not perfectly uniform and the boundaries of one particular sector could already be distinguished.

This means that in this particular case the market mode contains structural information about the stock market, and we need to keep it if we are to build an accurate correlation model. In fact, if we did not, as computations have shown, the constructed correlation matrix might not be positive semi-definite and arbitrary control of the negative entries would have to be implemented.

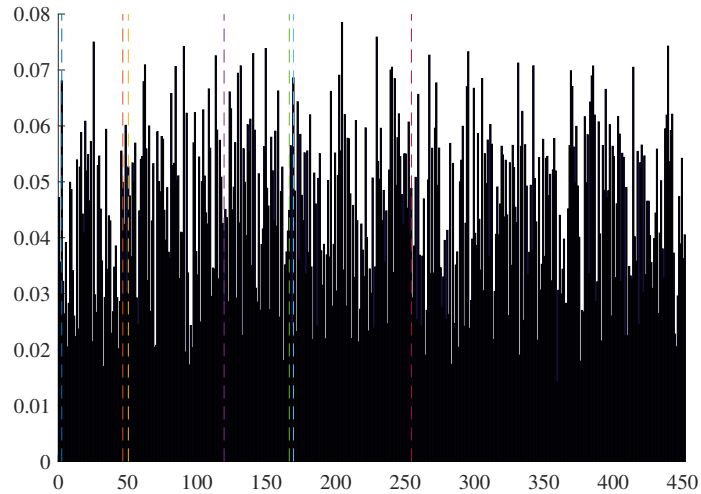
7.1 The construction of the correlation model

In the multi-layered model, the diagonal blocks of the correlation matrix model C represent the correlations inside the sub-clusters in the lowest layer from the top of the hierarchy. These diagonal blocks make up large diagonal blocks at a higher layer, and at each layer the off-diagonal blocks represent the correlation interactions between intermediate clusters in that layer. All background entries will be filled in with average correlations in that part of the layer, and the diagonal entries will be set to unit.

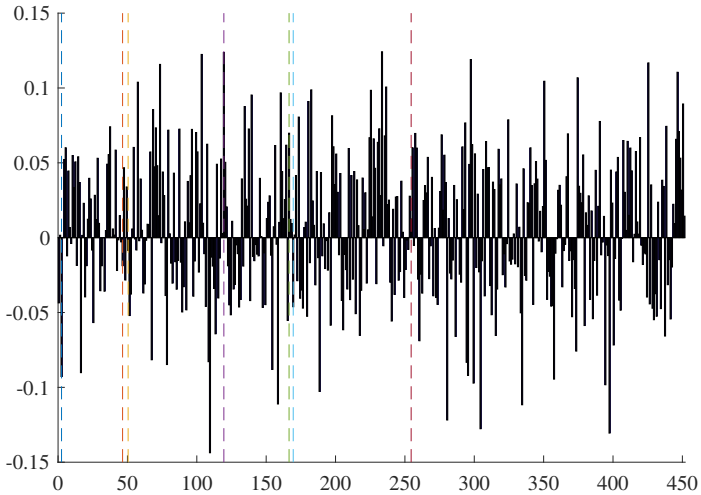
For example, for a two-sector market toy model,

$$C = \left(\begin{array}{cccc|cccc} 1 & \alpha_1 & \cdots & \alpha_1 & & & & \\ \alpha_1 & 1 & \ddots & \vdots & & \beta & & \\ \vdots & \ddots & \ddots & \alpha_1 & & & & \\ \alpha_1 & \cdots & \alpha_1 & 1 & & & & \\ \hline & & & & 1 & \alpha_2 & \cdots & \alpha_2 \\ & \beta & & & \alpha_2 & 1 & \ddots & \vdots \\ & & & & \vdots & \ddots & \ddots & \alpha_2 \\ & & & & \alpha_2 & \cdots & \alpha_2 & 1 \end{array}\right) \tag{6}$$

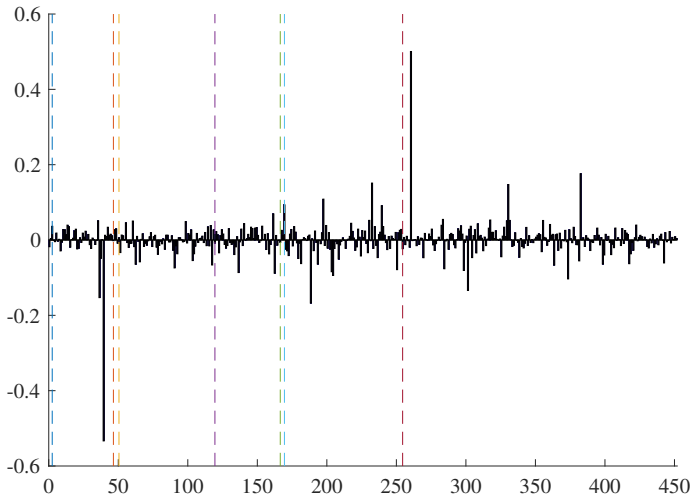
where diagonal blocks represent two clusters with respective average internal correlations $\alpha_{1,2}$, and the constant off-diagonal block entry β represents the average interaction correlation between the two clusters.



(a) The market mode with eigenvalue 99.1 and IPR 3.98×10^{-5}.

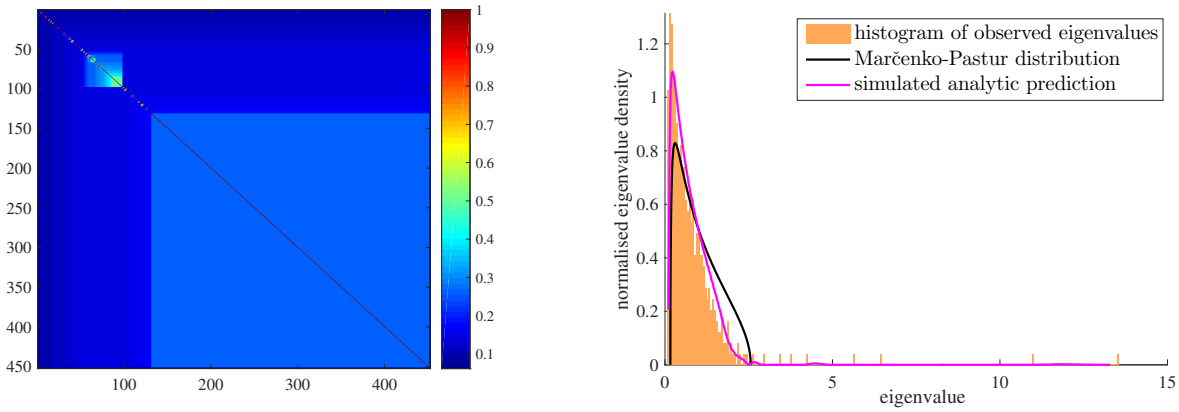


(b) The 8th mode with eigenvalue 3.46 and IPR 6.51×10^{-3}.



(c) The lowest mode with eigenvalue 0.0596 and IPR 0.149.

Figure 4: Plots of the eigenvector components of the same three modes. The vertical coloured lines define the boundaries of newly classified market sectors.



(a) Heat-map

(b) Simulated analytic prediction

Figure 5: The heat-map of a 50-layer correlation matrix model and its simulated analytic prediction for the empirical correlation matrix spectrum.

7.2 New predictions based on the model

Now that we have a model from the underlying correlation matrix C with eigenvalues denoted $\lambda_1, \ldots, \lambda_P$, we will used techniques in RMT to derive a predicted limiting eigenvalue distribution for the empirical correlation matrix E.

Let the limiting empirical eigenvalue density function (e.d.f.) of E be $f(\lambda)$, then Marčenko and Pastur have shown [4] that its Stieltjes transform, related by the transform pair

$$G(z) = \int_{-\infty}^{\infty} \mathrm{d}\lambda \frac{f(\lambda)}{\lambda - z}, \qquad f(\lambda) = \lim_{\epsilon \to 0} \operatorname{Im} G(\lambda + i\epsilon), \tag{7}$$

must satisfy the integral equation

$$-\frac{1}{G(z)} = z - r \int_{-\infty}^{\infty} \mathrm{d}\lambda \frac{\lambda \nu(\lambda)}{1 + \lambda G(z)}, \tag{8}$$

where the eigenvalue density function of the underlying correlation matrix is $\nu(\lambda) = \sum_{j=1}^{Q} p_j \delta(\lambda - \lambda_j)$ with $p_j \equiv n_j/P$, where n_j are the multiplicity of the Q distinct eigenvalues λ_j.

This results in a polynomial equation of degree Q:

$$[1 + zG(z)] \prod_{i=1}^{Q} [1 + \lambda_i G(z)] = rG(z) \sum_{i=1}^{Q} p_i \lambda_i \prod_{j \neq i}^{Q} [1 + \lambda_j G(z)] \,. \tag{9}$$

Solving this polynomial equation would give the new predicted distribution for the empirical correlation matrix spectrum, but in practice it is computationally costly and may suffer numerical instability. Instead, simulations of the empirical correlation matrix by randomly generated Gaussian data subject to correlation matrix C would suffice.

In Figure 5, we have shown the heat-map of a 30-layer correlation model constructed with information from clustering analysis along with the simulated analytic prediction for the empirical correlation matrix spectrum.

8 Dependence on Layer Depth of the Proposed Model

The key parameter of the proposed model that we could control is the layer depth, i.e. the number of layers constructed. To increase the layer depth, we essentially divide the lowest layer: in this process the original diagonal sub-blocks are split into two smaller diagonal sub-blocks, and new off-diagonal blocks are created.

8.1 A fundamental structure of the proposed model

To understand this process in detail as well as its effect on the eigenvalues of the correlation model, we consider the following matrices:

$$M := M_m(1,\alpha), \quad M_{1,2} := M_{m_{1,2}}(1,\alpha_{1,2}) \quad \text{and} \quad M' := \begin{pmatrix} M_1 & B \\ B^T & M_2 \end{pmatrix} \tag{10}$$

where $m = m_1 + m_2$, and

$$M_n(x,y) \equiv \underbrace{\begin{pmatrix} x & y & \cdots & y \\ y & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & y \\ y & \cdots & y & x \end{pmatrix}}_{n}, \quad B = \beta \underbrace{(1,\ldots,1)}_{m_1}^T \underbrace{(1,1,\ldots,1)}_{m_2}.$$

The interpretation of these matrices is that $M, M_{1,2}$ are all diagonal sub-blocks in the correlation model C and they have the same general form of $M_n(x,y) : n \times n$ with diagonal entries $x = 1$ and off-diagonal $y = \alpha, \alpha_{1,2}$. When layer division happens, the sub-block M becomes M', and we can view this change as a perturbation to entries α to $\alpha_{1,2}, \beta$ depending on its location, whether in $M_{1,2}$ or in $B^{(T)}$.

8.2 Determining the characteristic equations of matrices (10)

The matrix $M_n(x,y)$ can be reduced to a lower-triangular form by elementary operations:

$$\begin{aligned}
\det\left[M_n(x,y) - \lambda I\right] &= \begin{vmatrix} x-\lambda-y & 0 & \ldots & 0 & y-x+\lambda \\ 0 & x-\lambda-y & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & y-x+\lambda \\ 0 & \ldots & 0 & x-\lambda-y & y-x+\lambda \\ y & \ldots & y & y & x-\lambda \end{vmatrix} \\
&= \begin{vmatrix} x-\lambda-y & 0 & \ldots & 0 & 0 \\ 0 & x-\lambda-y & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \ldots & 0 & x-\lambda-y & 0 \\ y & \ldots & y & y & x-\lambda+(n-1)y \end{vmatrix} \\
&= (x-\lambda-y)^{n-1}\left[x-\lambda+(n-1)y\right].
\end{aligned}$$

Hence the original diagonal sub-block M has an eigenvalue $1-\alpha$ of multiplicity $m-1$ and a non-degenerate eigenvalue $1+(m-1)\alpha$.

Using the identity for invertible matrix block V

$$\begin{pmatrix} S & T \\ U & V \end{pmatrix} \begin{pmatrix} I & 0 \\ -V^{-1}U & I \end{pmatrix} = \begin{pmatrix} S - TV^{-1}U & T \\ 0 & V \end{pmatrix},$$

we have

$$\det(M' - \lambda I) = \det\left[(M_1 - \lambda I) - B(M_2 - \lambda I)^{-1}B^T\right]\det(M_2 - \lambda I)$$

for $\lambda \neq 1 - \alpha_2, 1 + (m_2 - 1)\alpha_2$ not an eigenvalue of M_2. Therefore

$$\begin{aligned}\det(M' - \lambda I) &= \det\left[(M_1 - \lambda I) - \beta^2 \sum_i \sum_j \left\{(M_2 - \lambda I)^{-1}\right\}_{ij} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}\right] \det(M_2 - \lambda I) \\ &= \det\left[M_{m_1}(1 - \gamma, \alpha_1 - \gamma) - \lambda I\right]\det(M_2 - \lambda I) \\ &= (1 - \lambda - \alpha_1)^{m_1 - 1}\left[1 - \lambda - \alpha_1 + m_1(\alpha_1 - \gamma)\right]\det(M_2 - \lambda I) \end{aligned} \tag{11}$$

where $\gamma(\lambda) = \beta^2 \sum_{i,j} \left\{(M_2 - \lambda I)^{-1}\right\}_{ij}$.

Hence we see that for $\alpha_1 \neq \alpha_2$, by symmetry $1 \leftrightarrow 2$, the characteristic equation of M' must be of the form

$$0 = \det(M' - \lambda I) = (1 - \lambda - \alpha_1)^{m_1 - 1}(1 - \lambda - \alpha_2)^{m_2 - 1}p(\lambda) \tag{12}$$

where $p(\lambda) = 0$ is a quadratic equation related to

$$1 - \lambda - \alpha_1 + m_1(\alpha_1 - \gamma) = 0. \tag{13}$$

The eigenvalues $1 - \alpha_{1,2}$ of $M_{1,2}$ are still eigenvalues of M' with respective multiplicities $m_{1,2} - 1$, and the remaining two eigenvalues of M' are roots of $p(\lambda) = 0$. To solve this we need to find γ, so we must be able to invert $M_2 - \lambda I$ to find $\gamma(\lambda)$.

To this end, we turn to the *Sherman-Morrison formula* for help:

Theorem 2 (Sherman–Morrison). *For an invertible matrix A and column vectors $\mathbf{u}, \mathbf{v}$ of compatible dimensions such that $1 + \mathbf{v}^T A^{-1}\mathbf{u} \neq 0$, the following formula holds:*

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

Proof. By direct verification. □

By setting $A = (1 - \lambda - \alpha_2)I$, $\mathbf{u} = \mathbf{v} = \sqrt{\alpha_2}\underbrace{(1, \ldots, 1)}_{m_2}^T$, we have

$$(M_2 - \lambda I)^{-1} = \frac{1}{1 - \lambda - \alpha_2}\left[I - \frac{\alpha_2}{1 - \lambda - \alpha_2 + m_2\alpha_2}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}\right].$$

Therefore

$$\gamma(\lambda) = \frac{\beta^2}{1 - \lambda - \alpha_2}\left(m_2 - \frac{m_2^2\alpha_2}{1 - \lambda - \alpha_2 + m_2\alpha_2}\right) = \frac{\beta^2 m_2}{1 - \lambda - \alpha_2 + m_2\alpha_2}$$

and equation (13) can be reduced to a more symmetric form in variable $\mu \equiv 1 - \lambda$ after rearranging,

$$q(\mu) = \mu^2 + (m_1\alpha_1 + m_2\alpha_2 - \alpha_1 - \alpha_2)\mu + \left[m_1 m_2 \alpha_1\alpha_2 - (m_1 + m_2 - 1)\alpha_1\alpha_2 - \beta^2 m_1 m_2\right] = 0. \tag{14}$$

Although in this derivation we have assumed $\lambda \neq 1 - \alpha_2, 1 + (m_2 - 1)\alpha_2$, the characteristic equation (12) with $p(\lambda) \equiv q(\mu)$ is valid $\forall\lambda$ because any divergence is offset by the $\det(M_2 - \lambda I)$ factor in equation (11).

8.3 Interpretation of the solutions of the polynomial equation (14)

In the case $\alpha_1 = \alpha_2 = \beta \equiv \alpha$, equation (14) has two roots $\lambda_1 = 1 + (m_1 + m_2 - 1)\alpha$ and $\lambda_2 = 1 - \alpha$, just as expected for eigenvalues of M since now $M' = M$. Here we note that λ_2 coincides with the other eigenvalues arising from the factor $(1 - \lambda - \alpha_1)^{m_1 - 1}(1 - \lambda - \alpha_2)^{m_2 - 1}$ in the characteristic polynomial (11).

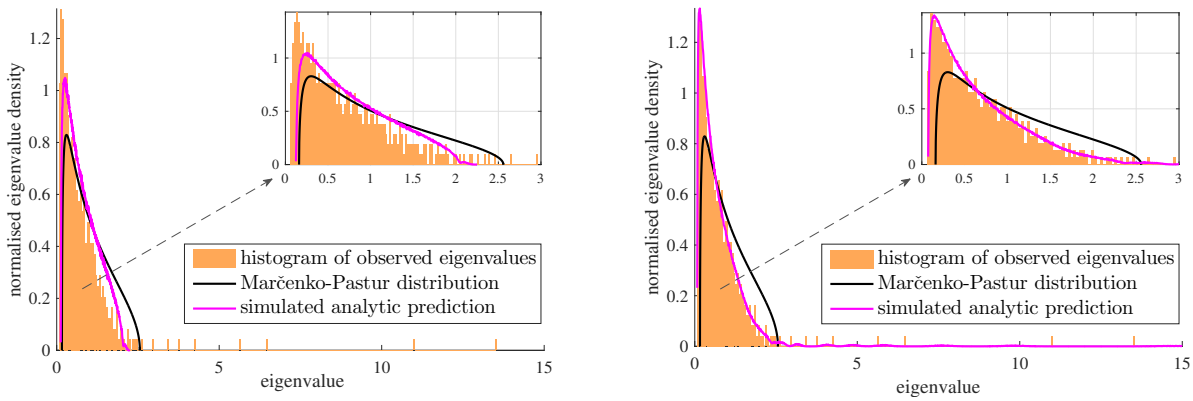
When layer division takes place to increase the layer number, we may have $\alpha_1 = \alpha_2 \equiv \alpha \neq \beta$ so that equation (14) is perturbed to

$$\mu^2 + (m-2)\alpha\mu + \left[(m_1 - 1)(m_2 - 1)\alpha^2 - \beta^2 m_1 m_2\right] = 0 \tag{15}$$

with roots denoted by $\lambda'_{1,2}$, where we recall $m = m_1 + m_2$. By trace consideration of M' or the properties of quadratic equations, we see that now $\lambda'_1 + \lambda'_2 = \lambda_1 + \lambda_2 = 2 + (m-2)\alpha$.

What this means is that in increasing the layer number by perturbing $\beta = \alpha$ in M to $\beta \ll 1$ in M' (as observed in the computed model of C), we have decreased the product of the roots $\lambda_1\lambda_2$ to $\lambda'_1\lambda'_2$ while their sum must be kept the same. Intuitively and conclusively, this tells us more layers in our model result in greater abundance of very large eigenvalues like λ'_1 and positive eigenvalues λ'_2 really close to zero.

Indeed, in Figure 6 where we compare the predicted empirical spectral density functions of a 10-layer correlation model and a 148-layer correlation model, we see that the latter is a closer match for the small positive eigenvalues. In fact, the latter is an excellent match with the observed eigenvalues, the best we have achieved!



(a) 10 layers (b) 148 layers

Figure 6: Comparison of the predicted empirical spectral density functions of a 10-layer correlation model and a 148-layer one with the market mode.

9 Alternative Model with Prior Sectoring

We saw in Section 6 that the new classification of sectors was not robust, whereas in Section 4 we saw at the market mode level the distinction between pre-assigned sectors was already obvious. We wonder if this prior information on sectoring could be incorporate in our model. Here we propose the following construction based on the observations.

We will assume the pre-assigned sectors have zero (little) correlations, whereas inside each sector

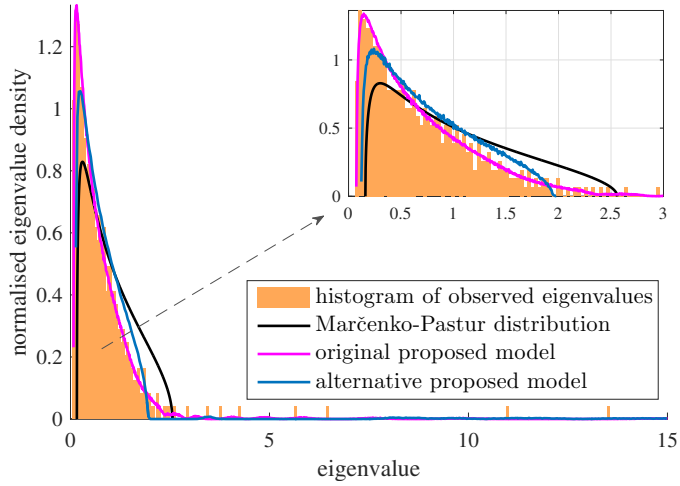


Figure 7: Predicted empirical spectral density functions of the alternative model with prior sectoring, superposed with the M-P law and the previous analytic prediction.

the stocks have equal mutual correlations. The underlying correlation matrix then takes the block-diagonal form

$$\begin{pmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_s \end{pmatrix}$$

where $M_i \equiv M_{m_i}$ as defined in equation (10), and $s = 10$ is the number of pre-assigned sectors (as in Figure 2). The data are processed without the removal of the market mode.

We repeat the procedure as in Section 7.2 to plot the predicted empirical spectrum of the correlation matrix in Figure 7.

We see that this model, even without detailed layer division, is a reasonable fit to the observed eigenvalue density, but it resembles in shape more of the M-P distribution than the observed distribution, or the previous 148-layer model.

10 Summary and Further Developments

Through mode and clustering analyses, we have been able to construct a multi-layer structured correlation model for the S&P 500 stock market. By analysing the dependence of the spectrum of our model on the layer depth, we have shown analytically that increasing the number of layers improves the match between (simulated) prediction of the empirical spectral density function with observed eigenvalues of the empirical correlation matrix.

This results in a reliable estimation for the underlying correlation structure of the market analysed, which may then have a positive impact on investment portfolios.

However, our study of the stock market correlations can be further developed by considering:

1) edge asymptotics;

Our correlation matrix is finite-dimensional, which means eigenvalues are expected to leak out of the edges of the predicted distribution of the empirical correlation matrix spectrum, which is derived

in the asymptotic limit. This leakage effect could be studied using the *Tracy-Widom law*.

2) time evolution;

The underlying stock market data have been assumed to have a stationary distribution in time, and this is unlikely a good assumption. We need to build the time parameter/variable t into our model.

3) fine-tuning.

We have performed the hierarchical clustering analysis with average linkage to build a binary tree structure. The correlation model can be improved with tailored clustering method suited for the stock market data.

Acknowledgements

Many thanks goes to my project supervisor, Dr Lucy Colwell, and her PhD student Chongli Qin at the Department of Chemistry, University of Cambridge, whose guidance and help have been crucial to this research project. I am also grateful for the generous support by the Bridgwater Scheme.

References

[1] S. Pafka, M. Potters, and I. Kondor. Exponential Weighting and Random-Matrix-Theory-Based Filtering of Financial Covariance Matrices for Portfolio Optimization. *arXiv:cond-mat/0402573*, February 2004.

[2] S. Wang. SUROP Project. http://sw664.user.srcf.net/SUROP%20Project%202016/SUROP.html, 2016. [On-line, updated and accessed 25/09/16].

[3] N. C. Snaith, P. J. Forrester, and J. J. M. Verbaarschot. Developments in Random Matrix Theory. *arXiv:cond-mat/0303207*, March 2003.

[4] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 1967.