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

UNIVERSITY OF CAMBRIDGE

Mike S. Wang // Summer, 2016

Faculty of Mathematics // Department of Chemistry

Random Matrix Theory meets Portfolio Theory?

Random Matrix Theory (RMT) studies the emergent behaviour in the *asymptotic limit* of many classes of matrices with *random variable* (r.v.) entries, when their dimensions tend to infinity. Introduced by Wishart in 1928, the theory gained popularity when Wigner applied it to nuclear physics.

Today, in modern portfolio theory, the covariance-correlation matrix is of fundamental importance to risk management and asset allocation; they belong to a large class of matrices in RMT called the *Wishart ensemble*.

In this study, our data set contains the standardised logarithmic returns for $P = 452$ stocks in the *Standard & Poor's* (S&P) *500* market over $T = 1258$ trading days.

A 'simple' spectrum prediction

If the standardised data matrix $X : T \times P$ has *independent identically distributed* (i.i.d.) r.v. entries with mean 0 and variance 1, then the *underlying* correlation matrix C would simply be the identity matrix I_P.

In such cases, a universal result below in RMT gives the limiting eigenvalue density function (e.d.f.) of the empirical correlation matrix $E \equiv T^{-1}X^TX$.

The *Marčenko-Pastur* (M-P) law

If $X : T \times P$ is a random matrix described as above, the limiting e.d.f. of matrix E is then given by

$$f(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{rx}$$

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

The M-P law provides a 'null hypothesis' distribution against which the observed e.d.f. of E can be compared. The eigenvalues not covered by it are signals that suggest the data are not truly random.

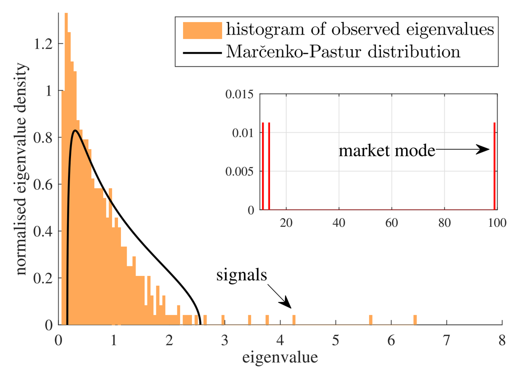

Fig. 1: Observed e.d.f. of E for the S&P 500 stock market. Large eigenvalues are shown in the inset.

This is unsurprising for stock market data, so we seek to improve the 'noise band' given by the M-P law.

Mode and clustering analyses

A feature of modes (eigenvalue-eigenvector pairs of the correlation matrix) is the localisation of their components, conveniently measured by the *inverse participation ratio* (IPR).

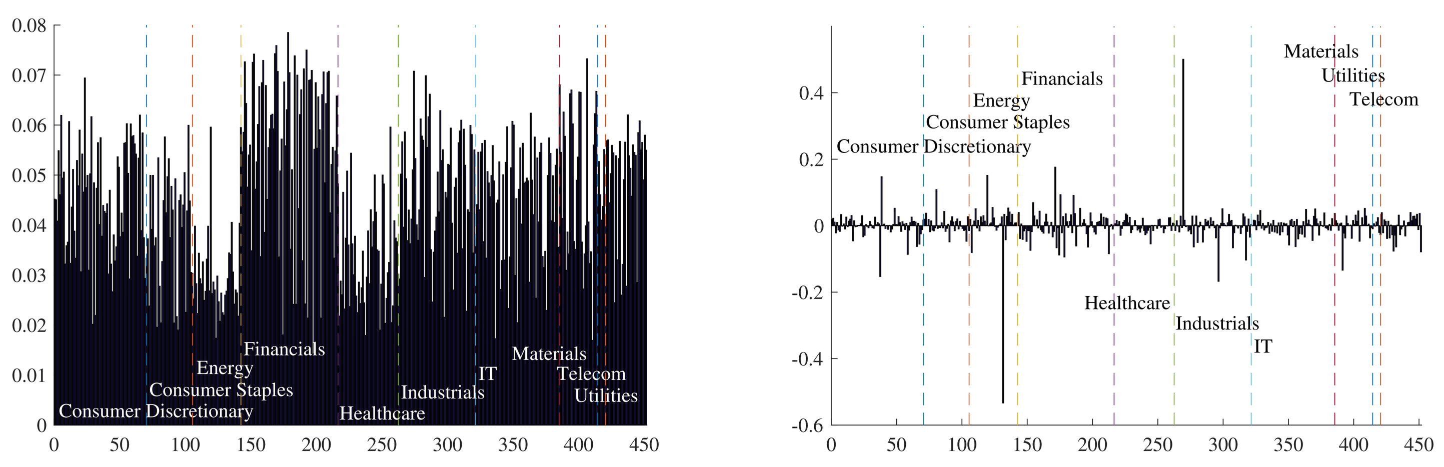

Fig. 2: The market mode (L) compoenents are uniform, as all stocks respond in a similar way to the overall market trend, whereas in the lowest mode (R) with high IPR, two stocks interact strongly.

Hierarchical clustering analysis can reveal the internal structure of the market. Here we have used the *average linkage method* with the *dissimilarity distance* measure.

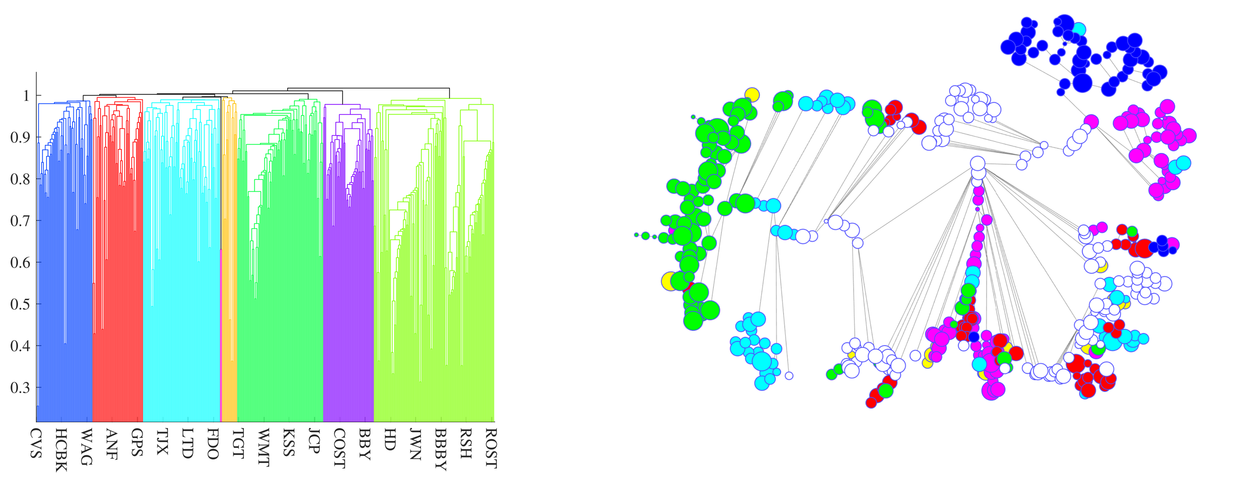

Fig. 3: The dendrogram generated by the average linkage clustering method and the minimum spanning tree (coloured by sectors) give complementary visualisations of the stock market internal structure.

A multi-layer correlation model & its predictions

Our analyses above help us construct a multi-layer model for the underlying correlation matrix C, whose off-diagonal entries are filled in by average interactions between stocks and (sub-)clusters.

Marčenko and Pastur have shown that for an underlying correlation matrix C with e.d.f. $\nu(\lambda)$, the *limiting* e.d.f. $f(\lambda)$ of the empirical correlation matrix satisfies the integral equation

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

where $G(z)$ is the *Stieltjes transform* of $f(\lambda)$, related by the transform pair

$$G(z) = \int_{-\infty}^{\infty} \mathrm{d}\lambda \frac{f(\lambda)}{\lambda - z}, \qquad f(\lambda) = \lim_{\epsilon\to 0} \operatorname{Im} G(\lambda + i\epsilon). \tag{2}$$

In our model, this results in a polynomial equation of degree equal to the number of distinct eigenvalues of C. Instead of solving it at high computational costs, simulating the empirical correlation matrix with random Gaussian data subject to C suffices.

Effect of layer division & an excellent match

Consider a two-sector toy model market whose correlation matrix is of the form

$$M = \begin{pmatrix} M_1 & B \\ B^T & M_2 \end{pmatrix}$$

where $M_{1,2} : m_{1,2} \times m_{1,2}$ have unit diagonals and constant off-diagonals $\alpha_{1,2}$, and $B : m_1 \times m_2$ have constant entries β.

By studing its characteristic polynomial with linear algebra identities such as the *Sherman-Morrison* formula, we find the eigenvalues $1 - \alpha_{1,2}$ of respective multiplicities $m_{1,2} - 1$ are always present whatever β is, but the other two eigenvalues $\lambda_{1,2}$ from the roots of $p(\lambda) = 0$ are perturbed.

This means for our more complex correlation model with fundamental structures just like this, layer division results in perturbations that push the eigenvalues $\lambda_{1,2}$ apart, giving a very large and a very small eigenvalue.

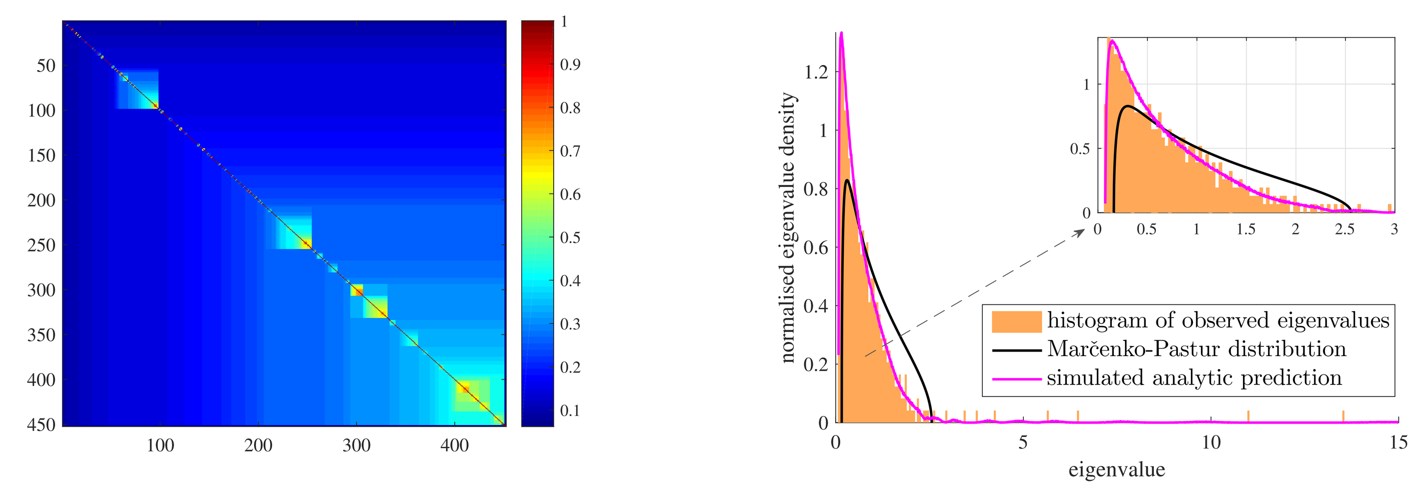

Fig. 4: Left: a 148-layer model *with the market mode* for the underlying correlation matrix of the S&P 500 market analysed. Right: new predicted empirical spectral distribution of the 148-layer model.

Our stock market empirical correlation matrix has many near-zero and a number of very large eigenvalues, so by increasing the number of layers in our model, we obtain an analytic prediction of the limiting e.d.f. of E that is an excellent match with the observations (see top right figure).

Summary & further developments

Through our analyses, we have constructed a multi-layer structured correlation model of the S&P 500 market that match the observations very well. Further developments could include:

- edge asymptotics. Our correlation matrix is finite-dimensional, but the analytic prediction is derived in the asymptotic limit. Its resulting artefacts can be studied with the *Tracy-Widom law*.
- time evolution. The underlying data have been, probably incorrectly, assumed to have a stationary distribution in time.
- …

The author is a Part III Mathematical Tripos undergraduate. This project is part of the Summer Undergraduate Research Opportunities Programme (SUROP) and is generously supported by the Bridgwater Scheme.

Many thanks goes to the project supervisor, Dr Lucy Colwell, and her PhD student Chongli Qin at the Molecular Informatics Centre, Department of Chemistry, for their kind support in this study.

The MATLAB Live Scripts are published at people.ds.cam.ac.uk/sw664/SUROP Project 2016/SUROP.html.

LaTeX TikZposter