Eur. Phys. J. B **27**, 189–195 (2002)
DOI: 10.1140/epjb/e20020145

THE EUROPEAN
PHYSICAL JOURNAL B
EDP Sciences
© Società Italiana di Fisica
Springer-Verlag 2002

Eigen analysis of the stability and degree of information content in correlation matrices constructed from property time series data

W. Cook[a], C. Mounfield, P. Ormerod, and L. Smith
Volterra Consulting Ltd 5, The Old Power Station, 121 Mortlake High St, SW 14 8SN London, UK

Received 31 December 2001

Abstract. Property is an asset which forms part of the portfolios of many investors, particularly institutional ones, along with equities and bonds. Techniques from physics, particularly that of random matrix theory, have provided powerful insights into the behaviour of financial assets. A large database providing time series data for over 10,000 individual properties is available for the UK. Some of the data is available at an annual and some at a monthly frequency. However, even at the monthly frequency, only a relatively small number of observations is available, certainly in comparison with that available with financial assets. A key issue in translating methods of analysis in financial markets to property data is whether they are applicable given the small number of data points available. This paper addresses this issue. Using the tools of random matrix theory, we find that a great deal of information is contained within property data. The correlations between different types and geographical locations of property tend to have far more true information and be more stable over time than is the case with financial data, despite the large number of observations available with the latter.

PACS. 01.75.+m Science and society

1 Introduction

Property is an asset which forms part of the portfolios of many investors, particularly institutional ones, along with equities and bonds. Techniques from physics, particularly that of random matrix theory, have provided powerful insights into the behaviour of financial assets (for example, [1,8]).

A large database providing time series data for over 10,000 individual properties is available for the UK from IPD Ltd.. Property funds allow IPD to gather information on their individual properties, which IPD processes and turns into more aggregate indices of performance, over a range of property types and geographical locations. The aggregations are available at either annual or monthly frequency, depending upon the particular aggregation which is carried out.

However, even at the monthly frequency, only a relatively small number of observations is available, certainly in comparison with that available with financial assets. Monthly data is available from December 1986, and annual data from 1981.

A key issue in translating methods of analysis in financial markets to property data is whether they are applicable given the small number of data points available. This paper addresses this issue. This paper describes the application of techniques from Random Matrix Theory (RMT) in assessing the degree of information content (and its stability) contained within correlation matrices formed from time series of property price data. The analysis is undertaken for property data at both an aggregated regional level (using IPD data) and also at the individual property level using data from a major UK property fund.

2 The tools of random matrix theory

The techniques of Random Matrix Theory (RMT) have recently been applied to financial market data to analyse the true degree of information content contained within empirical correlation matrices formed from equity returns [1,8]. In addition to this the techniques have also been applied to macroeconomic data [9,10].

The essence of the RMT approach of assessing the degree to which an empirical correlation matrix is noise dominated lies in comparing the eigenspectra properties of the empirical matrix with the theoretical eigenspectra properties of a random matrix. Undertaking this comparison identifies those eigenstates of the empirical correlation matrix which contain genuine information content. These eigenstates are specific to the system under consideration and are indicative of the presence of collective modes of ‘motion’ of correlated groups of assets. The remaining eigenstates will be noise dominated and hence unstable over time. The stability of the information content of the eigenmodes (that is to say the stability of the

[a] e-mail: wcook@eigenrisk.com

correlations between the assets in the portfolio) can be assessed by analysing in more detail the precise structure of the information carrying eigenmodes (for example [6,7]).

To quantify the degree to which genuine information is contained within the correlation matrix, and the stability of that information, quantitative measures of the spectral properties of the correlation matrix are required.

In order to assess the degree to which an empirical correlation matrix is noise dominated one may compare the eigenspectra properties of the empirical matrix with the theoretical eigenspectra properties of a random matrix [11]. Undertaking this analysis will identify those eigenstates of the empirical matrix which contain genuine information content. The remaining eigenstates are understood to be noise dominated and hence potentially unstable over time. The eigenstates that contain genuine information content are specific to the system under consideration and are indicative of the presence of collective modes of motion.

Consider a matrix $\underline{\underline{M}}$ of T observations of the prices of N assets (at a frequency of *e.g.* monthly observations). In the context of property data the N assets could correspond to regional IPD indices or to individual properties. If the inter-period logarithmic returns are defined as

$$M_i(t) = \ln P_i(t) - \ln P_i(t-1)$$

then the correlation matrix measuring the correlations between the N assets is given by

$$\underline{\underline{C}} = \frac{1}{T-1}\underline{\underline{M}}\underline{\underline{M}}^T \cdot$$

If the T observations are i.i.d random variables then in the limit $N \to \infty$ and $T \to \infty$ the density of eigenvalues, λ , of the random correlation matrix $\underline{\underline{C}}$ is given by

$$\rho_C(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max}-\lambda)(\lambda-\lambda_{min})}}{\lambda}$$

for $\lambda \in [\lambda_{min}, \lambda_{max}]$ where $Q = T/N \geq 1\cdot$

The upper and lower bounds on the theoretical eigenvalue distribution are given by,

$$\lambda_{max} = \sigma^2\left(1+\frac{1}{\sqrt{Q}}\right)^2$$

$$\lambda_{min} = \sigma^2\left(1-\frac{1}{\sqrt{Q}}\right)^2$$

(σ^2 is the variance of the elements of $\underline{\underline{M}}$, usually rescaled to unity).

This range of eigenvalues corresponds to a random, noisy subspace band where the postulates of RMT hold. That is to say, the eigenvectors corresponding to eigenvalues within $\lambda_{min} < \lambda < \lambda_{max}$ contain no genuine information and the components of the associated eigenvectors are indistinguishable from random noise.

The eigenvalue distribution of the empirical correlation matrices can be compared to this 'null-hypothesis' distribution and thus, in theory, if the distribution of eigenvalues of an empirically formed matrix differs from the above distribution, then that matrix will not have completely random elements. In other words, there will be structure present in the correlation matrix. Each isolated eigenstate outside of the RMT bounds represents a correlated group of assets whose size and participants are obtained from the eigenvalue and eigenvector respectively.

When the dimensions of the random matrix under consideration are finite (but still 'large') this has the effect of broadening the spectral distribution. However in these instances Monte-Carlo simulation can generate what the broadened eigenvalue distribution is expected to be.

To analyse the structure of the eigenvectors of the empirical correlation matrix the inverse Participation Ratio (IPR) may be calculated. The IPR is commonly utilised in localisation theory to quantify the contribution of the different components of an eigenvector to the magnitude of that eigenvector (thus determining if an eigenstate is localised or extended) [6].

Component i of an eigenvector ν_i^α corresponds to the contribution of time series i to that eigenvector. That is to say, in this context, it corresponds to the contribution of asset i to eigenvector α. In order to quantify this we define the IPR for eigenvector α to be

$$I^\alpha = \sum_{i=1}^{N}(\nu_i^\alpha)^4 \cdot$$

Hence an eigenvector with identical components $\nu_i^\alpha = 1/\sqrt{N}$ will have $I^\alpha = 1/N$ and an eigenvector with one non-zero component will have $I^\alpha = 1$. Therefore the inverse participation ratio is the reciprocal of the number of eigenvector components significantly different from zero (*i.e.* the number of assets contributing to that eigenvector).

For those eigenvectors that deviate from the theoretically predicted bounds of RMT it is important to quantify the degree of stability of the information content of the eigenmode (*i.e.* the stability of the correlations between the assets). This is necessary since spurious correlations may be introduced by a particular choice of data to calculate the correlation matrix from. We may assess this stability by calculating the scalar product of eigenvectors in non-overlapping analysis periods (for an application to macro-economic data of this concept see [10]). That is for two analysis periods T_A and T_B we form the overlap matrix

$$\underline{\underline{O}}(T_A, T_B) = \begin{pmatrix} \underset{\rightarrow}{\nu}^N(T_A)\cdot\underset{\rightarrow}{\nu}^N(T_B) & \dots & \underset{\rightarrow}{\nu}^N(T_A)\cdot\underset{\rightarrow}{\nu}^1(T_B) \\ \vdots & \ddots & \vdots \\ \underset{\rightarrow}{\nu}^1(T_A)\cdot\underset{\rightarrow}{\nu}^N(T_B) & \dots & \underset{\rightarrow}{\nu}^1(T_A)\cdot\underset{\rightarrow}{\nu}^1(T_B) \end{pmatrix}.$$

Hence if the eigenvector structure remains perfectly stable in time (*i.e.* the correlations between the assets contributing to that eigenvector remain stable from period to period) then each element of the overlap matrix would be equal to $O_{ij}(T_A, T_B) = \delta_{ij}$. No inter-period stability would imply that $O_{ij}(T_A, T_B) = 0$.

The overlap matrix (the matrix of dot products of each eigenvector with every other eigenvector in 2 non-overlapping periods) is therefore a means to quantify the degree of temporal stability in the correlations between the assets.

3 Application to property data

3.1 IPD monthly regional data

The analysis is undertaken using the IPD monthly regional data. This data extends from Dec. 1986 (at a benchmark index level of 100) until Feb. 2001. There are thus 171 observations. These observations are made for the three property types (office, retail and industrial) and are aggregated at a regional level. 35 different asset types were used for the analysis (*e.g.* South East retails). The data measures the total return on the assets in a particular region. The logarithmic returns are calculated (reducing the number of observations to 170) and the spectral properties of the empirical correlation matrix calculated.

In order to analyse this dataset it is first segregated into a number of non-overlapping analysis periods. Given that we have 170 observations we may segregate the data into two non-overlapping periods of 85 observations each (corresponding to the periods Jan. 1987–Jan. 1994 and Feb. 1994–Feb. 2001) or into 3 non-overlapping periods of 56 observations each (corresponding to the periods Mar. 1987–Oct. 1991, Nov. 1991–Jun. 1996, Jul. 1996–Feb. 2001).

For the 2 non-overlapping periods case we form the correlation matrix from the relevant observations of the price changes of the 35 regions and calculate the spectral properties of the correlation matrix. In this case the eigenvalue statistics are:

	Period 1	Period 2
Theoretical minimum eigenvalue	0.123	0.123
Theoretical maximum eigenvalue	2.72	2.72
Number of eigenvalues below theoretical minimum	12	12
Number of eigenvalues above theoretical maximum	2	1
Number of eigenvalues in the noise band	21	22

The corresponding eigenvalue statistics for the 3 non-overlapping periods case are,

	Period 1	Period 2	Period 3
Theoretical minimum eigenvalue	0.041	0.041	0.041
Theoretical maximum eigenvalue	3.23	3.23	3.23
Number of eigenvalues below theoretical minimum	5	10	7
Number of eigenvalues above theoretical maximum	2	1	1
Number of eigenvalues in the noise band	28	24	27

In both cases we observe that there are a significant number of eigenvalues which lie outside of the noise sub-space band. These eigenvalues demonstrate that there is present within the correlation matrix genuine information to be extracted.

For the 2 analysis period case we observe that the largest eigenvalue is approximately 22, implying that 62% of the total information within the correlation matrix is contained within the market eigenmode. The corresponding IPR for this eigenvalues eigenvector is 0.029, which is very close to the value 1/35 which we would expect for an eigenmode where all of the assets are contributing equally (confirming that this eigenmode corresponds to the correlated movements of the market as a whole). Similar results are observed for the 3 non-overlapping case.

In order to verify that the RMT is detecting genuine information we may repeat the analysis using the same dataset which has been shuffled at random. That is each of the 35 time series of 170 observations is shuffled at random 10,000 times. This has the effect of preserving the properties of the distribution of returns for each time series (for example the mean and variance) but if there is any temporal correlation in the asset price dynamics this information will be completely removed from the time series. In addition to this any cross-correlations between the equal-time asset price dynamics will be removed.

Undertaking this analysis for the 2 and 3 non-overlapping periods yields the results that in every analysis period all 35 eigenvalues of the empirical correlation matrix lie within the noise sub-space band. This demonstrates very clearly that there are non-trivial temporal correlations present within the original data set.

To assess how stable these correlations remain over time we may calculate the overlap matrix between the eigenvectors in each analysis period. For the 2 non-overlapping periods the diagonal elements of the overlap

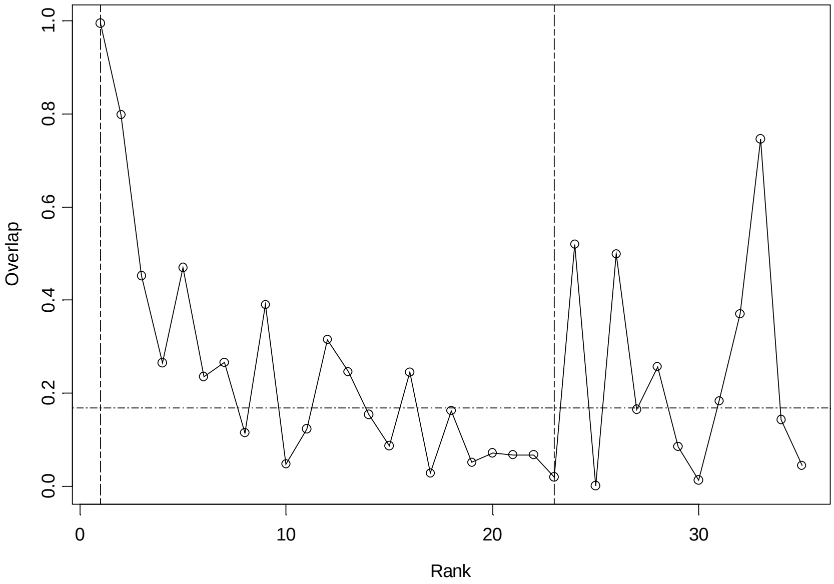

Fig. 1. The overlap between eigenvectors for 2 non-overlapping analysis periods.

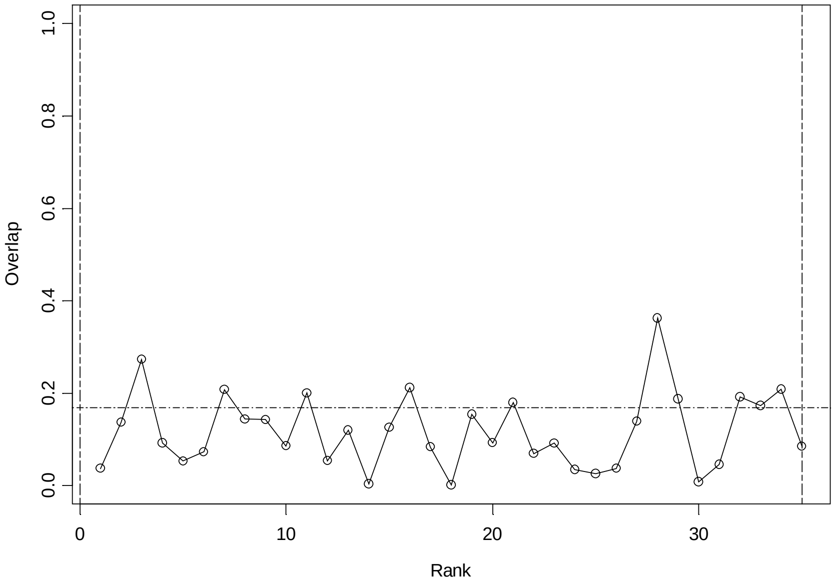

Fig. 2. Overlap matrix between 2 non-overlapping periods (randomly shuffled data).

matrix are plotted against the rank of the eigenvalue in Figure 1. The rank corresponds to the eigenvalue *e.g.* rank 1 in these examples corresponds to the largest eigenvalue. Rank 35 corresponds to the smallest eigenvalue. The horizontal dotted line corresponds to the 'noise' level. This is defined as $1/\sqrt{N}$ where N is the number of variables (distinct time series) used in the analysis, which in this case is equal to 35. Any measurement lying below this line is indistinguishable from noise.

The two dotted vertical lines correspond to the boundaries of the noisy sub-space band. That is to say the eigenvalues below the lower vertical line correspond to those eigenvalues which lie above the theoretical maximum for a random matrix and the eigenvalues above the upper vertical line are those eigenvalues which lie below the theoretical minimum for a random matrix. Between these lines it is to be expected that the overlaps between the eigenvectors in the analysis periods will be subject to stochastic fluctuations.

The noise band in this case is fairly narrow. Eigenvectors 1 and 2 have overlaps which are very large. We would of course expect eigenvector 1 to have a significant degree of stability since this corresponds to the 'market' eigenmode.

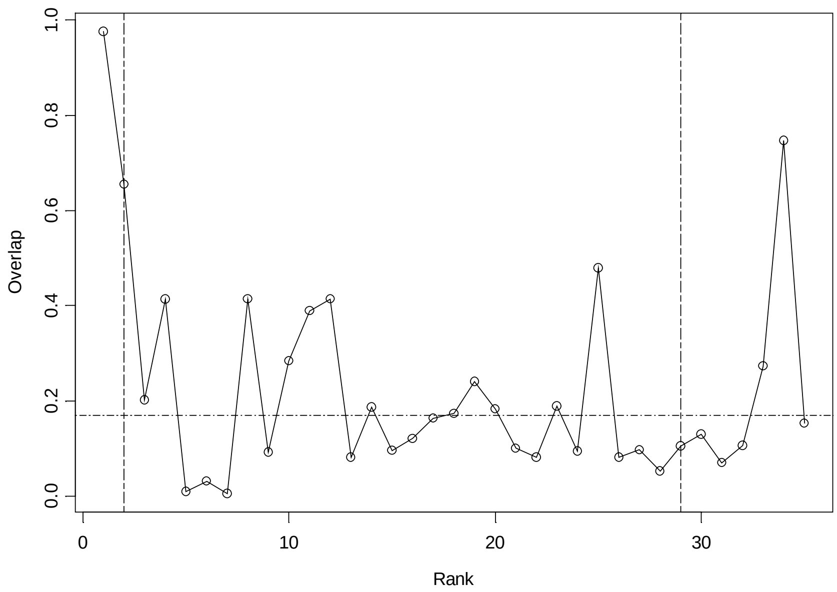

Fig. 3. The overlap between eigenvectors for 3 non-overlapping analysis periods (periods 1 and 2).

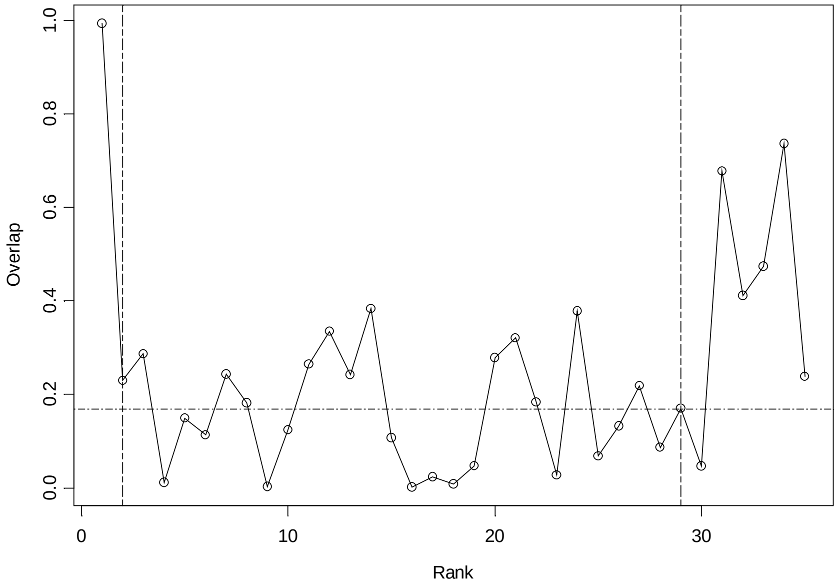

Fig. 4. The overlap between eigenvectors for 3 non-overlapping analysis periods (periods 2 and 3).

We may compare this with the overlap matrix calculated for the randomly shuffled dataset. This is shown in Figure 2.

It is clear that there is a significant difference between the two cases. In particular for the shuffled data there is no structure in the correlations apparent whatsoever (indeed most of the overlaps lie below the noise threshold as we would expect). For the original data however it is apparent that the high and low-lying eigenmodes have a structure (*i.e.* the overlaps between these eigenvectors are significantly different from pure noise)

Similar results are obtained for the case of 3 non-overlapping periods. Shown in Figures 3 and 4 are the overlaps between periods 1 and 2 and 2 and 3 respectively.

As with the 2 non-overlapping periods case it is apparent that there is significant structure in the information carrying eigenmodes. Repeating this analysis with the randomly shuffled data yields the expected result that there is no consistent structure in the overlaps between eigenmodes.

3.2 IPD annual regional data

We may repeat the previous analysis using the annual IPD regional data. In this case there are now only 20 yearly observations of total return on properties in the 35 regions.

Because of the small number of observations we will only consider segregating the data into 2 non-overlapping

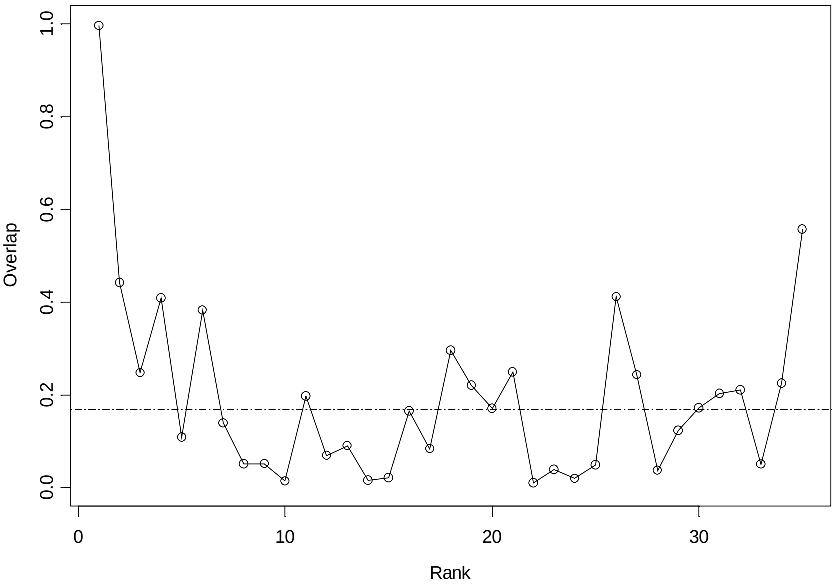

Fig. 5. The overlap between eigenvectors for 2 non-overlapping analysis periods (annual regional data).

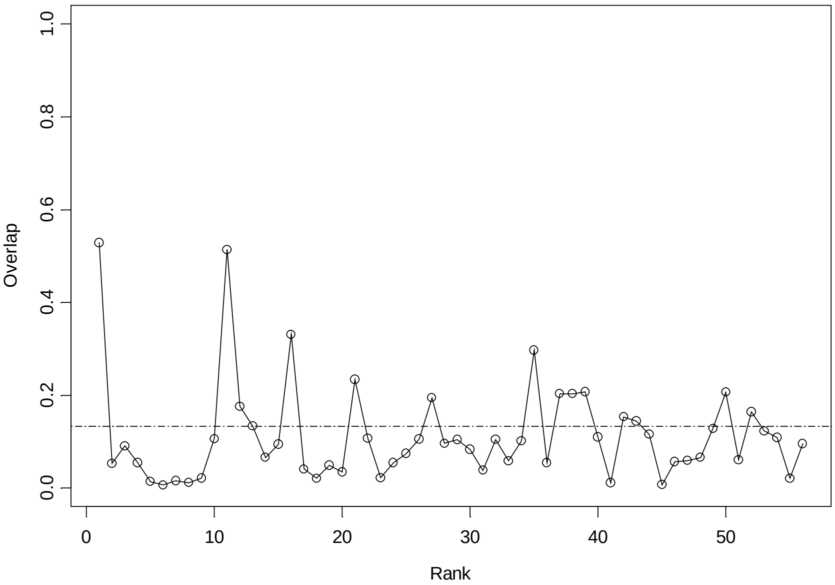

Fig. 6. he overlap between eigenvectors for 2 non-overlapping analysis periods (annual individual property data).

periods (period 1 corresponding to returns in the period 1981–1990 and period 2 corresponding to 1991–2000). Shown in Figure 5 is the overlaps between the eigenvectors in the two periods as a function of the eigenvalue rank.

The graph is qualitatively the same as for the monthly data. That is the largest eigenvectors remain highly stable – the macro level, the mid – range eigenvectors are noisy and the low lying eigenvectors display a degree of stability that is significantly different from noise.

4 Application to individual property data

We may of course undertake the same analysis for the regional data at the individual property level. This particular dataset corresponds to 56 individual properties taken from a major property fund[1] for which there exist a time series of annual observations of total return for the period 1982–2000.

This dataset is segmented into two non-overlapping periods and the overlap between the eigenvectors in the two periods plotted in Figure 6.

Comparing this figure with that for the annual regional data it is apparent that there is less stability in the correlations between the price movements at an individual property level than there is for the regional data. However the largest eigenmodes do display a degree of stability that is

[1] A condition of using this data is that the fund must remain anonymous.

significantly different from the noise threshold suggesting that there is indeed genuine information contained within the eigenmodes of the correlation matrix.

5 Conclusions

Despite the small number of observations available on the returns on property, a great deal of information is contained within the data. The correlations between different types and geographical locations of property tend to have far more true information and be more stable over time than is the case with financial data, despite the large number of observations available with the latter.

An important reason for this is that a single factor exercises a powerful influence over the property market, namely the state of the business cycle. Properties of all types and in all locations are influenced by this. Of course, many other factors influence the property market, but thebusiness cycle effect is strong. Almost all property returns tend to rise sharply in a boom, and fall in a recession. This gives considerable structure and stability to the correlations within property portfolios.

References

1. J.-P. Bouchaud, M. Potters, *Theory of Financial Risks – From Statistical Physics to Risk Management* (Cambridge University Press, 2000).
2. S. Drozdz, J. Kwapien, F. Grummer, F. Ruf, J. Speth, Physica A **299**, 144 (2001).
3. P. Gopikrishnan, B. Rosenow, V. Plerou, H.E. Stanley, *Identifying Business Sectors from Stock Price Fluctuations*, `cond-mat/0011145`, (2000).
4. L. Laloux, P. Cizeau, J.-P Bouchaud, M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).
5. R.N. Mantegna, H.E. Stanley, *An Introduction to Econophysics* (Cambridge University Press, 2000).
6. V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).
7. V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Physica A **287**, 374 (2000).
8. B. Rosenow, V. Plerou, P. Goprikrishnan, H.E. Stanley, *Portfolio Optimization and the Random Magnet Problem*, paper presented at the conference in honour of Gene Stanley, Sicily, December 2001 (2001).
9. P. Ormerod, C. Mounfield, Physica A **280**, 497 (2000).
10. P. Ormerod, C. Mounfield, Physica A **307**, 494 (2002).
11. M. Mehta, *Random Matrices* (Academic Press, 1991).