Extracting information from noisy time series data

Paul Ormerod (Pormerod@volterra.co.uk)

Volterra Consulting Ltd

Sheen Elms
135c Sheen Lane
London SW14 8AE

December 2004

Abstract

A question which is prior to any analysis of co-movements in economic time series is the extent to which the series contain genuine information rather than noise.

In this paper, I describe a methodology for distinguishing between signal and noise, and illustrate its application both with respect to the degree of convergence in European business cycles (the co-movements between real GDP growth in the various countries) and to understanding the poor practical record of macro-economic forecasts (the co-movements between variables within economies).

By way of illustration, the degree of synchronisation of the business cycles may be quantified by calculation of the correlation matrix of the matrix of observations formed from the time series of GDP growth for each economy.

However due to the finite size of the number of variables (which corresponds to the number of economies) and the number of observations (which is the number of observations of GDP) then a reliable determination of the correlation matrix may prove to be problematic. The structure of the correlation matrix may be dominated by noise rather than by true information.

In order to assess the degree to which an empirical correlation matrix is noise dominated we can compare the eigenspectra properties of the empirical matrix with the theoretical eigenspectra properties of a random matrix. Undertaking this analysis will identify those eigenstates of the empirical matrix who contain genuine information content. The remaining eigenstates will be noise dominated and hence unstable over time.

1. Introduction

The distribution of the eigenvalues of *any* random matrix has been obtained analytically (Mehta, 1991). In particular, the theoretical maximum and minimum values can be calculated. We compare the eigenvalues of the correlation matrix of the data series in which we are interested with the theoretical maximum and minimum values of those of a random matrix of similar dimension.

Empirical correlation matrices will in general contain both noise and true information. If such a matrix is dominated by noise, its eigenspectra will resemble closely those of a purely random matrix, and we will not be able to draw any meaningful inferences from such a matrix, regardless of the technique which we use.

We illustrate this with two empirical applications. First, the degree of convergence of the business cycles of the main EU economies. Secondly, we combine this approach with the method of delays from the dynamic time-series literature (for example, Mullin 1993) and examine the extent to which consistently successful macro-economic forecasts can be made in principle.

2. Random matrix theory

Quarterly data exists for most of the EU economies over the past twenty years or so for the level of real output in the economy (GDP). We can therefore calculate annual growth rates quarter-by-quarter. The correlations between these growth rates for the various economies will inform us about the extent to which their business cycles are in synchronisation.

In other words, the degree of synchronisation of the business cycles may be quantified by calculation of the correlation matrix of the matrix of observations formed from the time series of GDP growth for each economy.

If $\underline{\underline{M}}$ is an N x T rectangular matrix (T observations of the GDP growth of the N economies) and $\underline{\underline{M}}^T$ is its transpose, the correlation matrix $\underline{\underline{C}}$ as defined below is an N x N square matrix

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

However due to the finite size of N (which corresponds to the number of economies) and T (which is the number of observations of GDP) then a reliable determination of the correlation matrix may prove to be problematic. The structure of the correlation matrix may be dominated by noise rather than by true information.

In order to assess the degree to which an empirical correlation matrix is noise dominated we can compare the eigenspectra properties of the empirical matrix with the theoretical eigenspectra properties of a random matrix. Undertaking this analysis will identify those eigenstates of the empirical matrix who contain genuine information content. The remaining eigenstates will be noise dominated and hence unstable over time. This technique has recently been applied by many researchers to financial market data (for example, Mantegna et al 1999, Laloux et al 1999, Plerou et al 1999, Gopikrishnan et al 2000, Plerou 2000, Bouchaud et al 2000, Drozdz et al 2001).

For a scaled random matrix **X** of dimension N x T, (i.e where all the elements of the matrix are drawn at random and then the matrix is scaled so that each column has mean zero and variance one), then the distribution of the eigenvalues of the correlation matrix of **X** is known in the limit T, N $\rightarrow \infty$ with Q = T/N $\geq$ 1 fixed (Sengupta et al 1999). The density of the eigenvalues of the correlation matrix, λ, is given by:

$$\rho(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_{\max}-\lambda)(\lambda-\lambda_{\min})}}{\lambda} \qquad \text{for } \lambda \in [\lambda_{\min}, \lambda_{\max}] \qquad (1)$$

and zero otherwise, where $\lambda_{\max} = \sigma^2 (1 + 1 / \sqrt{Q})^2$ and $\lambda_{\min} = \sigma^2 (1 - 1 / \sqrt{Q})^2$ (in this case $\sigma^2 = 1$ by construction).

The eigenvalue distribution of the correlation matrices of matrices of actual data can be compared to this 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 random elements. In other words, there will be structure present in the correlation matrix.

To analyse the structure of eigenvectors lying outside of the noisy sub-space band 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) (Plerou et al 1999).

Component i of an eigenvector v_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 economy i to eigenvector α. In order to quantify this we define the IPR for eigenvector α to be

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

Hence an eigenvector with identical components $v_i^\alpha = 1/\sqrt{N}$ 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 economies contributing to that eigenvector).

3. The data and the results: European business cycles

Quarterly levels of real GDP over the period 1977Q1 - 2000Q3 are analysed from the OECD database for the largest EU economies, France, Germany, Italy, Spain and the UK. The first three plus the Benelux countries are widely regarded as forming the EU 'core', being the founder members of the (then) European Economic Community. Quarterly data is available for the Netherlands but not for Belgium, and we include the former in the 'core' group[1].

We analyse the correlation matrix of real GDP growth rates for the following groups of countries:

- EU 'core' i.e. France, Germany, Italy and the Netherlands

- EU core plus Spain
- EU core plus the UK

As a comparator, we also analyse the EU core plus a random data series generated from a large number of random shuffles of the data for Germany. This sets the benchmark of what we would expect to see if an economy were added to the EU core data set whose short-term growth rates over the business cycle are by construction not correlated with those of the core members.

In terms of the EU core, there is a large amount of genuine correlation between the growth rates of the economies over the business cycle. Further, there is a substantial degree of stability of these correlations over the 1978-2000 period.

The theoretical range of the eigenvalues for a random matrix of the relevant order is between 0.62 and 1.46. The eigenvalues of the empirical correlation matrix of annual growth rates over the 1978Q1 - 2000Q3 period are 2.68, 0.69, 0.39 and 0.24, indicating the presence of a large amount of true information in the correlation matrix.

In terms of those eigenvalues which lie outside the noisy sub-space band the most important from a macroeconomic perspective is the largest eigenvalue. The application of these techniques to equities traded in financial markets have demonstrated that this eigenmode corresponds to the 'market' eigenmode (e.g. Gopikrishnan et al, 2000). In this context the largest eigenvalue will inform us as to the degree to which the movements of the EU economies are correlated.

The contribution which each of the core economies makes to eigenvector 1 can be seen from calculating the IPR. The components are in fact (0.49, 0.55, 0.51, 0.44), which gives a calculated value of the IPR of 0.256, indicating that all four economies are contributing approximately equally to this eigenvector.

[1] The Luxemburg economy is trivially small

The fact that the individual elements are almost identical in size also shows that this vector corresponds to a collective motion of all of the GDP growth time series. It is therefore a measure of the degree to which the growth of different countries in the EU core is correlated.

The trace of the correlation matrix is conserved, and is equal to the number of independent variables for which time series are analysed. That is, for the core EU correlation matrix the trace is equal to 4 (since there are 4 time series). The closer the 'market' eigenmode (i.e. eigenmode 1) is to this value the more information is contained within this mode i.e. the more correlated the movements of GDP. The market eigenmode corresponds to the largest eigenvalue. The degree of information contained within this eigenmode, expressed as a percentage, is therefore $100\lambda_{max}/ N$.

To follow the evolution of the degree of business cycle convergence over time we may analyse how this quantity evolves temporally. The analysis is undertaken with a fixed window of data. Within this window the spectral properties of the correlation matrix formed from this data set are calculated. In particular the maximum eigenvalue is calculated. This window is then advanced by one period and the maximum eigenvalue noted for each period.

The choice of an appropriate window to span the periodicity of what constitutes the business cycle is not completely straightforward. Business activity is influenced by a very large number of events, and these events may be very diverse in character and scope. Individual cycles therefore vary both in terms of amplitude and period. We initially carried out results for a window of 10 years, although the results for a window of 8 years are virtually identical, and it is these which we present here. The results are in fact robust to the choice of window, until a window as short as 5 years is chosen, when greater instability begins to be introduced, due to measurement noise induced by the reduced number of observations.

The results for the core EU economies are set out in Figure 1. Each window contains 32 quarterly observations, and so we have 60 windows in total. The period 1978Q1 - 1985Q4 corresponds to the first data point in Figure 2, 1978Q2 - 1986Q1 to the second, and so on through to 1992Q4 - 2000Q3.

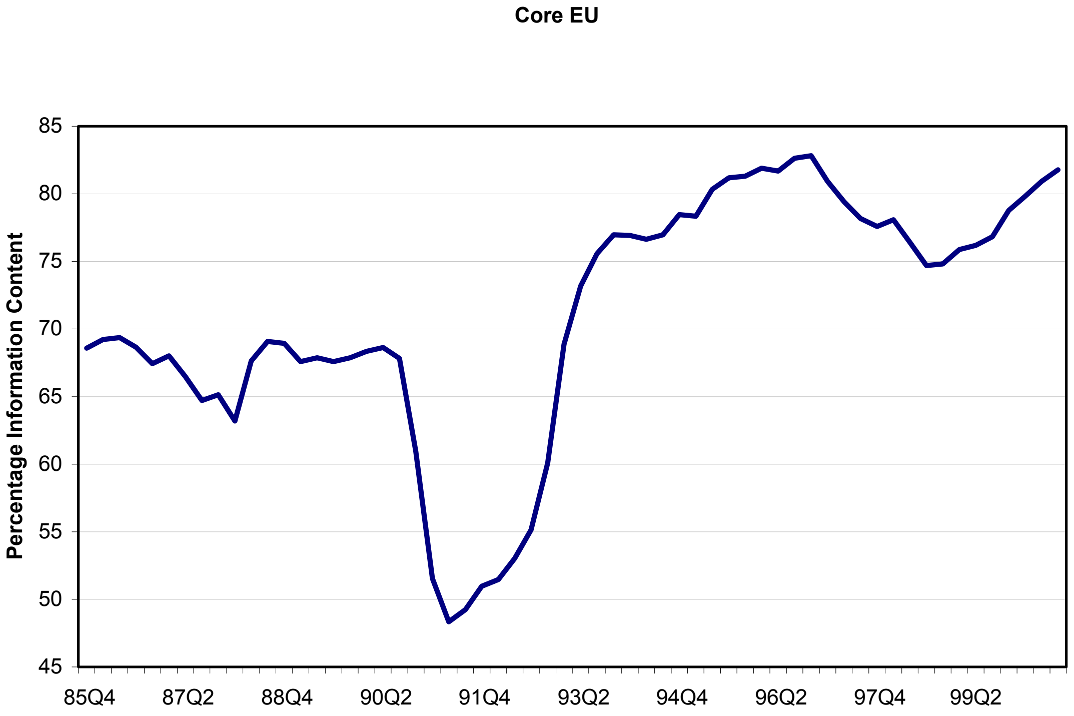


Figure 1

The temporal evolution of the degree of information content in the maximum eigenvalue of the empirical correlation matrix formed from the time series of quarterly GDP growth for the core EU economies of France, Germany, Italy and the Netherlands.

Even in the early part of the period, the 'market' eigenvalue took up some 70 per cent of the total of the eigenvalues, indicating a strong degree of convergence of the business cycles of the EU core economies. There was a temporary reduction of convergence around the time of German re-unification in the early 1990s, but the economies rapidly re-converged and the principal eigenvalue now accounts for some 80 per cent of the total information content within the correlation matrix, indicating a movement towards even greater convergence of the business cycles of the EU core economies over time.

As a benchmark for comparison, Figure 2 illustrates the effect of adding a purely random series to the core EU data set. This consists of the German quarterly growth rate data shuffled 100,000 times (thereby destroying any temporal correlations in the data).

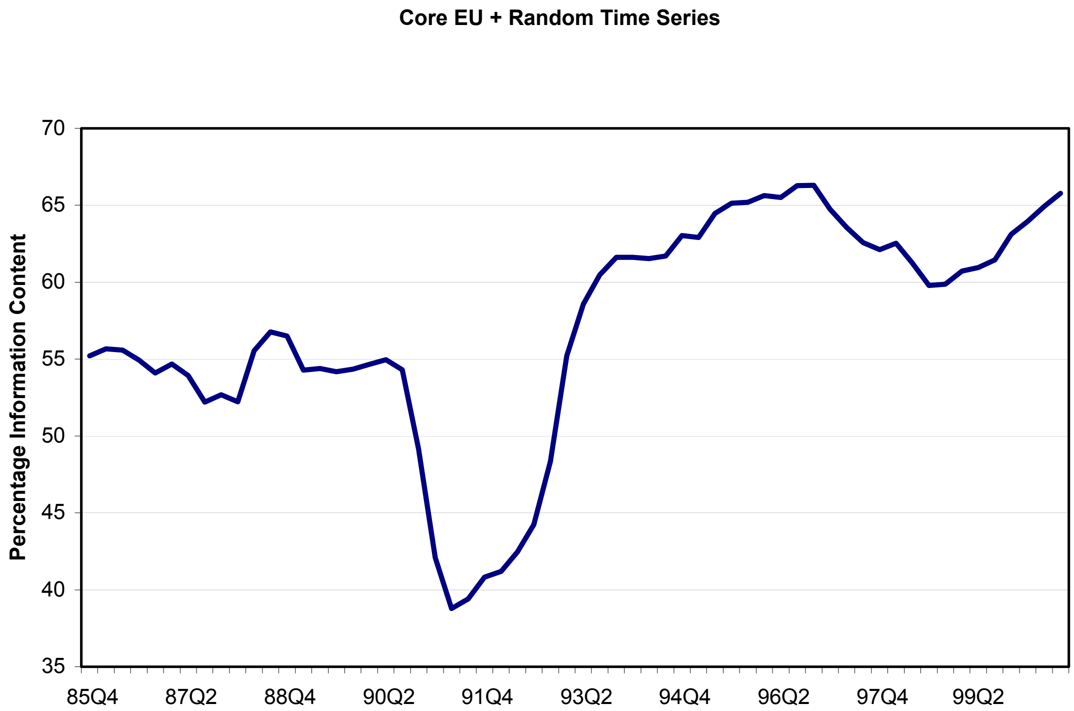


Figure 2

The temporal evolution of the degree of information content in the maximum eigenvalue of the empirical correlation matrix formed from the time series of quarterly GDP growth for the core EU economies of France, Germany, Italy and the Netherlands plus a time series formed from 100,000 random shuffles of the German GDP time series

.

The pattern of movement is almost identical to that of Figure 1. The important point to note from this is the scale over which the contribution of the maximum eigenvalue moves. There are now 5 series in the data set rather than 4, so the sum of the eigenvalues is now 5. Essentially, the data plotted in Figure 3 is the data in Figure 2 multiplied by 4/5.

In other words, Figure 2 represents what we would observe if an economy whose business cycle were completely uncorrelated to that of the EU core were added to the data set.

We now move on to examine the case of Spain. After many years isolated under dictatorship, the Spanish authorities have attached great importance to modernising their economy and society in a European context. Policy has been strongly supportive of European integration. The extent to which business cycle convergence has been achieved with the EU core is plotted in Figure 3.

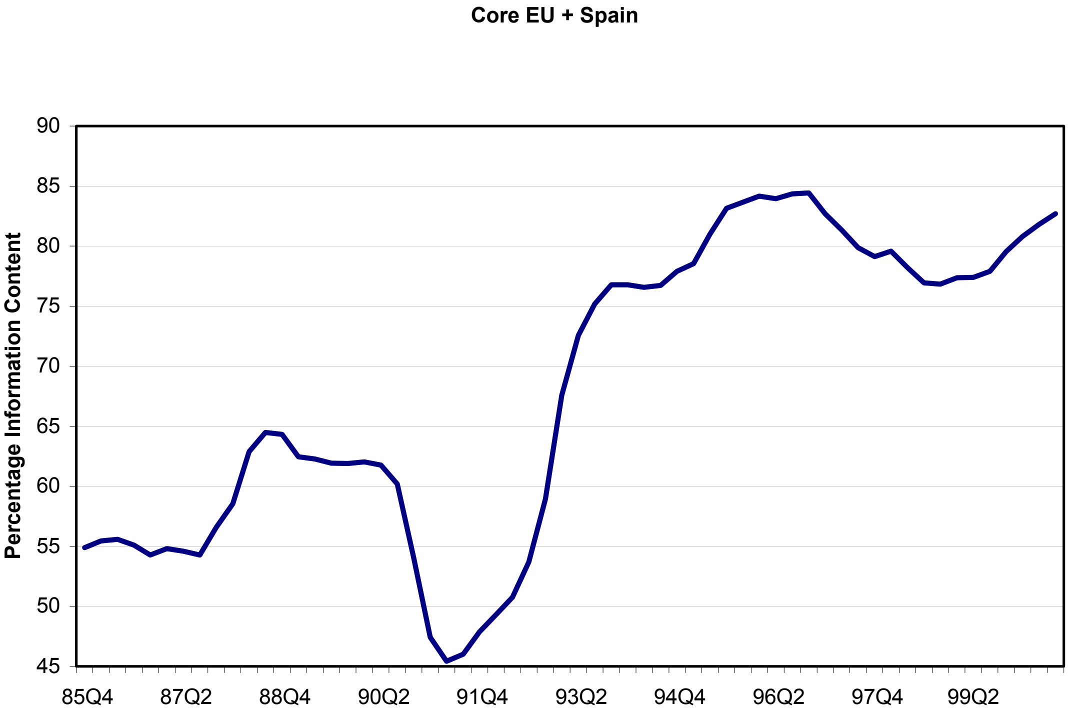


Figure 3

The temporal evolution of the degree of information content in the maximum eigenvalue of the empirical correlation matrix formed from the time series of quarterly GDP growth for the core EU economies of France, Germany, Italy and the Netherlands plus the time series of GDP growth for the Spanish economy.

Qualitatively, the pattern over time is similar to that of Figure 1, reflecting, for example, the temporary impact of German re-unification. But there is a very clear upward trend in these results. In the early parts of the window, the value of $100\lambda_{max}/N$ is around 55, very similar to that of the core EU plus a random data series. However, by the end this has risen to around 80, very similar to that of the core EU alone.

In other words, this suggests strong evidence to support the view that the Spanish economy has become closely converged with the core EU economies in terms of its movements over the business cycle.

In contrast, Figure 4 shows the results for the core EU plus the UK.

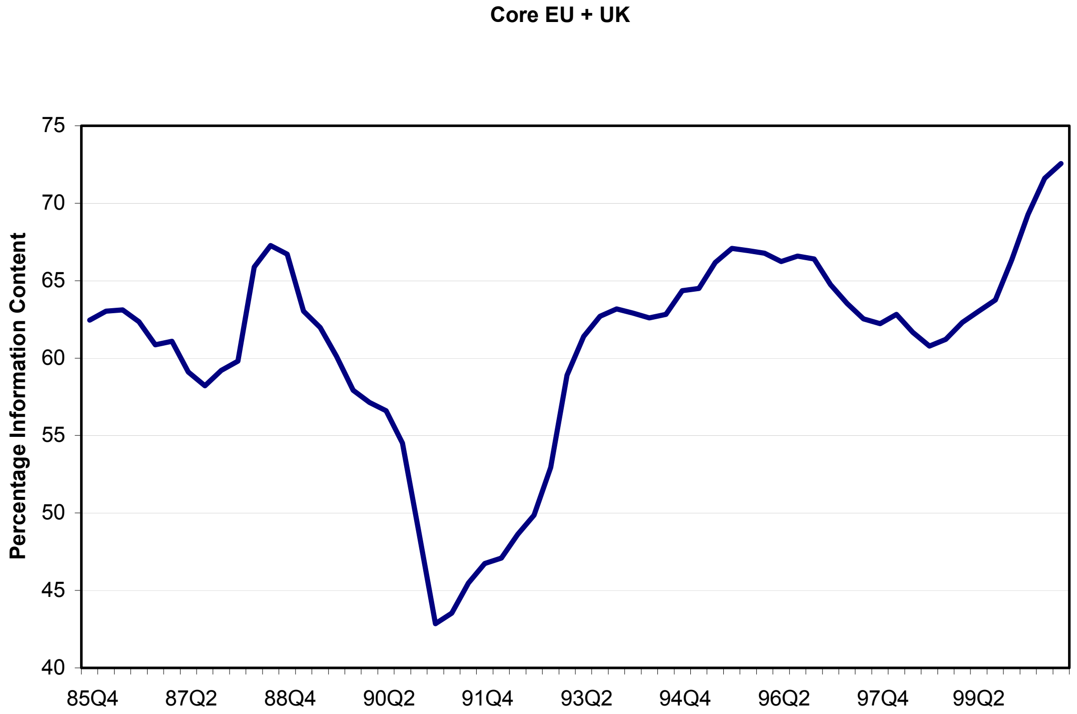


Figure 4

The temporal evolution of the degree of information content in the maximum eigenvalue of the empirical correlation matrix formed from the time series of quarterly GDP growth for the core EU economies of France, Germany, Italy and the Netherlands plus the time series of GDP growth for the UK economy.

In the early part of the analysis period, the value of $100\lambda_{max}/N$ for the EU core plus UK is around 65, less than for the core EU itself, but distinctly higher than for either the core EU plus Spain or the core EU plus a random series. However, subsequently the value shows no clear trend, unlike the case when Spain is added. At the very end of the analysis period, there is a rise to just over 70, but this remains well below the value for the EU core and the EU core and Spain, and indeed may simply be a temporary fluctuation around an average value of some 65.

Table 1 summarises these findings

	Average Information Content in Market Eigenmode in First 20 Periods	**Average Information Content in Market Eigenmode in Last 20 Periods**
Core EU	68%	79%
Core EU + Random	55%	63%
Core EU + Spain	59%	81%
Core EU + UK	61%	65%

Table 1

4. The data and the results: macroeconomic time series

A standard method in the time series analysis of dynamic systems is to form a delay matrix from the original series (e.g. Mullin 1993). Let **x**(t) be a T x 1 vector of observations of the rate of growth of GDP at time t, where t runs from 1 to T. We form a delay matrix, **Z**, such that the first column of **Z** is **x**(t), the second **x**(t-1), and so on through to **x**(t-m) in the (m+1)th column.

By suitable choice of m, the delay matrix can span what is usually thought of as the time period of the business cycle in economics, in other words the period over which any regularity of behaviour of the growth of GDP might be postulated to exist. Individual cycles vary both in terms of amplitude and period.

Carefully constructed annual data on GDP per capita in 17 advanced capitalist economies from 1870 to 1994 is provided in Maddison (1995)[2]. These include the US, the UK, Germany, France, Italy and Japan. For each of these economies, after calculating the annual percentage rate of growth and taking lags up to 12 years, we obtain a delay matrix of dimension 112 x 13.

For a random matrix of this dimension, (1) indicates that the eigenvalues of its correlation matrix should fall in the range 0.435 to 1.797 (where the variables in the matrix are independent identically-distributed normal random variables scaled to have constant unit volatility). However, (1) only holds in the limit, and so we examined the possible existence of small-sample bias. Computing the eigenvalues of the correlation matrix of 5,000 such random matrices did in fact suggest a slight bias, with the range falling between 0.329 and 2.005. The summary statistics for the 65,000 eigenvalues computed are as follows:

min 0.329; 1st quartile 0.727; mean 1; 3rd quartile 1.247; max 2.005

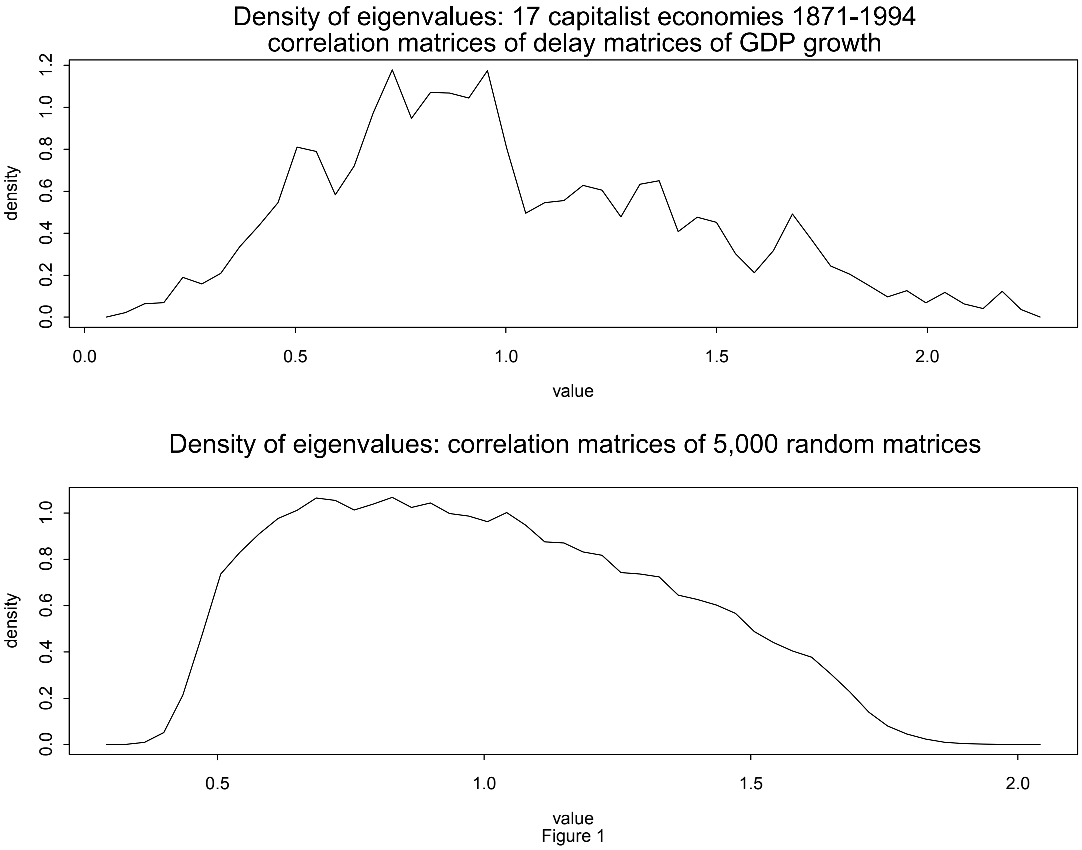
[2] For Japan, estimates are available from 1885, and for Switzerland from 1900

The eigenvalues of the correlation matrix of the delay matrix formed from the annual rate of growth of GDP can be compared with these theoretical and empirical ranges. For no less than 11 out of the 17 countries, the eigenvalues all lie within the theoretical limits given by (1). France and Denmark each have one eigenvalue which lies within the theoretical maximum value in the limit and the empirical maximum (1.803 and 1.828 respectively), and Canada (1.960 and 1.882) and New Zealand (1.955 and 1.887) each have two.

Only in the case of the US and the UK do we find eigenvalues outside the empirically calculated range of those for a random matrix. Even here, they are not large, being 2.174 and 2.051 for the US and 2.181 and 2.055 for the UK. A potential economic explanation of this finding is that over the 1871-1994 period, only the UK and, latterly, the US have been the dominant world economic power, and therefore able to operate, albeit to a very limited extent, as an isolated system.

As a comparator, we took a time series where it is known that accurate genuinely ex ante predictions can be obtained, namely the mean annual sunspot data. This is shown, for example, with a low-dimensional non-linear autoregressive model (e.g. Tong , 1990). We took data for the same length as that available for annual GDP growth, and calculated a delay matrix again using 12 lags. The resulting largest eigenvalues were 5.391 and 4.126, far larger than the empirical maximum. The third largest eigenvalue was also greater than the latter value, at 2.213. The smallest eigenvalues were far below the minimum, being close to zero. In other words, the technique shows that the correlations between lagged values of the sunspot series do contain genuine information. This is consistent with the fact that the series can be predicted with systematic accuracy over time.

Figure 6 plots the density of the eigenvalues calculated for all 17 economies, and the density of the eigenvalues of the correlation matrices of 5,000 random matrices discussed above. The latter looks much smoother when plotted, for there are 65,000 such eigenvalues compared to only 221 for the economic data.


Density of eigenvalues: 17 capitalist economies 1871-1994
correlation matrices of delay matrices of GDP growth

Density of eigenvalues: correlation matrices of 5,000 random matrices

Figure 6

On a formal Kolmogorov-Smirnov goodness-of-fit test, the hypothesis that the cumulative density function of the economic data eigenvalues is not significantly different from that of the cumulative density function of the random data eigenvalues is rejected at a p value of 0.073. In other words, at the conventional level of $p = 0.05$, the null hypothesis is not rejected. However, the lack of rejection is not dramatic, and careful inspection of the economic data eigenvalues reveals that almost all countries have at least one eigenvalue close to the theoretical maximum for a random matrix - as the small peak at around 1.7 in Figure 6 shows. In other words, the results suggest a small amount of genuine information in the movements of GDP.

Calculation of the inverse participation ratio of the eigenvectors also reveals a certain amount of information in the data. (see Ormerod and Mounfield , 2000)

We also considered the quarterly rate of growth of GDP in the US and the UK. Data on the economy as a whole is not in general available at a higher sampling frequency (for example, monthly). Quarterly data is available only for the post-war period, in the case of the US from 1947Q1 through 1999Q3, and for the UK from 1955Q1 through 1999Q1.

The theoretical range of eigenvalues for the correlation matrix of the US data delay matrix - again using lags of up to 12 years - is from 0.209 to 2.380. A calculation using 5,000 random matrices of the same dimension produced a range of 0.154 to 2.541. The actual eigenvalues calculated from US data lie somewhat more decisively outside this range than those calculated from the longer run of annual data. The largest two are 3.32 and 3.16, and the third largest is 2.451. This suggests a certain degree of information in the data correlations, and hence in principle a certain degree of predictability. The Federal Reserve do appear to have had some success in actual prediction during the 1990s.

For the UK, the results with quarterly data suggest less information content in the post-war period than is the case with the longer run of annual data. The range of eigenvalues from random matrices of the same dimension as the UK delay matrix is from 0.118 to 2.551. The largest eigenvalue calculated from the UK data is, by coincidence, also 2.551. These findings for the US and UK in the post-war period are consistent with the hypothesis formulated above to account for the longer run results, namely the ability of the dominant world economic power to operate as a partially isolated system. In the post-war period, this has not been true of the UK, but certainly is the case for the US.

The implication here is that the poor macro-economic forecasting record is due to inherent properties of the data, and cannot be overcome regardless of the economic theory or statistical technique which is applied to the problem.

5 Conclusion

We illustrate the application of random matrix theory to investigate the extent to which economic data being analysed contains true information or is composed of noise. Such investigation is logically prior to any statistical modelling of the data.

We show that the co-movements over time between the growth rates of the EU economies do contain a large amount of genuine information. In contrast, the co-movement over time between lags of such macro-economic data is strongly dominated by noise. This suggests both that the macro-forecasting record of such series is poor for reasons inherent to the data. Further, econometric models of such series will, when confronted with genuine out-of-sample data, break down much more frequently than conventional statistical theory suggests.

References

J.-P. Bouchaud and M. Potters, *Theory of Financial Risks – From Statistical Physics to Risk Management*, Cambridge University Press (2000)

S. Drozdz, J. Kwapien, F. Grummer, F. Ruf, J. Speth, *Quantifying the Dynamics of Financial Correlations*, cond-mat/0102402 (2001)

P. Gopikrishnan, B. Rosenow, V. Plerou, and H.E. Stanley *Identifying Business Sectors from Stock Price Fluctuations*, cond-mat/0011145 (2000)

L. Laloux, P. Cizeau, J.-P Bouchaud and M. Potters *Noise Dressing of Financial Correlation Matrices*, Phys Rev Lett **83**, 1467 (1999)

R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics*, Cambridge University Press (2000)

M. Mehta, *Random Matrices*, Academic Press (1991)

P. Ormerod and C. Mounfield, *Random Matrix Theory and the Failure of Macroeconomic Forecasts*, Physica A **280**, 497 (2000)

P Ormerod and C Mounfield, *The Convergence of European Business Cycles 1978-2000*, Physica A **000**, 000 (2002)

V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral and H.E. Stanley *Universal and Non-universal Properties of Cross-correlations in Financial Time Series*, Phys Rev Lett **83**, 1471 (1999)

V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral and H.E. Stanley *A Random Matrix Theory Approach to Financial Cross-Correlations*, Physica A **287**, 374 (2000)

A.M Sengupta and P. P. Mitra, Phys Rev E **60** 3389 (1999)

A. Maddison , *Monitoring the World Economy 1820-1992*, OECD, Paris (1995)

H.Tong, *Non-linear Time Series: A Dynamic Systems Approach*, Oxford Statistical Science Series, Clarendon Press, Oxford (1990)