18.325 Final Project : Noise Detection from Financial Correlation Matrices

Lee, Young Han

Massachusetts Institute of Technology,

Room 2-093,77 Massachusetts Avenue,

Cambridge, MA 02139

hahn@mit.edu

18th May, 2001

Abstract

The purpose of this paper is to distinguish true information from noise in financial correlation matrices and partially remove the noise. Thus the resulting *cleaned* matrices can be used in building an optimal portfolio with lower risk. Using several properties of random matrices and positive definite matrices such as 'spectral theorem' for symmetric matrices, we can identify the noise band of the eigenvalues of correlation matrix.

Key words and Phrases: Correlation matrices, positive definite, spectral theorem and Central Limit Theorem for eigenvalues

1 Introduction

Financial correlation matrices are widely used in financial sectors for building an optimal portfolio. However, it is indicated in the book 'Theory of Financial Risks' written by Dr. Jean-Philippe Bouchaud that you should be careful in constructing optimal portfolio, since it is likely contain much noise. According to the traditional view of Markowitz' portfolio theory, smaller eigenvalues correspond to smaller risk and thus more weight is given to the eigenvectors corresponding to them, but it turns out that they are mostly noise. In this paper, by mainly reproducing his results with recent data, some properties of financial correlation matrices are examined.

2 Construction of correlation matrix

2.1 Mathematical preliminaries

Correlation matrices have special properties. It is symmetric, all the diagonal elements are 1(correlation with itself) and other off-diagonal elements are between -1 and 1. Moreover, it is positive definite.

- **Definition 1** A real matrix $A \in R^{n \times n}$ is *symmetric* if it has the same entries below the diagonal as above: $a_{ij} = a_{ji}$ for all i, j, hence $A = A^T$. [2]
- **Definition 2** A square $n \times n$ real symmetric matrix A is said to be *positive definite* if $x^T A x > 0$ for all $x \in R^n$. [3]
- **Theorem 1: Spectral theorem**

  If the entries in an $n \times n$ matrix are real and the matrix is symmetric, then the eigenvalues are all real and the eigenvectors are orthogonal. Furthermore, we can construct an orthonormal basis consisting of eigenvectors $\{v_1, v_2, ...v_n\}$ (called 'spectral basis'), and the diagonal matrix with diagonals composed of eigenvalues. [3]
- **Theorem 2 : Positive eigenvalues**

  If a square matrix matrix A is symmetric and positive definite, all eigenvalues are positive.

  (pf)

  Let v be such an orthonormal eigenvector obtained by *spectral theorem.* Let λ be the corresponding eigenvalue. Then,

  $$v^T A v = v^T \lambda v = \lambda v^T v = \lambda > 0$$

2.2 Data for correlation matrices

Now, given stock market data, let's see how we construct a correlation matrix. First of all, 233 representative stocks from S&P500 are chosen and their prices between 1/2/1997 and 5/11/2001(1100 days)[1] were used for constructing the correlation matrix. [2]

Then the daily return is calculated, since what is important in construction of optimal portfolio is *the rate of return*, not the price itself.

[1] Provided by Sangil, Oum, Department of mathematics, Princeton university

[2] The reason that only 233 data were used is mainly because of the column size limit of 'Microsoft excel'(It can only express 256 columns). Actually, this is one limitation for the 'Microsoft excel' when used in financial purpose and I hear people in Wall Street complain about it.

Table 1: Real market data: Daily prices(in dollars) for 233 stocks in S&P500 for 1100 days

date	AA	AAPL	ABK	ABS	ABT	ABX	...
5/11/2001	40.27	22.85	53.86	31.86	49.26	17.18	...
5/10/2001	41.91	23	54.5	32.23	48.1	17.48	...
5/9/2001	41.38	23.98	54.24	32.3	48.36	17.42	...
5/8/2001	40.34	24.57	53.52	32.68	47.7	16.51	...
5/7/2001	41	24.96	53.72	33.4	47	16.71	...
5/4/2001	41.9	25.75	53.3406	33.36	46.59	16.27	...
5/3/2001	40.89	24.96	53.1209	33.15	45.81	16.28	...
5/2/2001	41.33	26.59	51.9726	32.65	45.37	16.95	...
5/1/2001	41.35	25.93	52.9811	33.18	45.8	16.96	...
4/30/2001	41.2504	25.49	53.73	33.4	46.38	16.44	...
4/27/2001	41.8382	26.2	54.1194	33.05	45.98	16.78	...
4/26/2001	41.9279	24.69	54.1693	32.15	44.86	16.37	...
.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	...
1/24/1997	15.8457	8.4375	21.2556	32.2944	25.3025	25.7899	...
1/23/1997	15.7288	8.625	21.3755	32.0645	25.1864	25.3101	...
1/22/1997	16.1673	8.5938	21.3755	32.4093	26.1149	25.4301	...
1/21/1997	16.5181	8.625	21.4554	32.9839	25.5926	26.2697	...
1/20/1997	16.9567	8.4688	21.3355	32.6392	25.6507	25.9099	...
1/17/1997	17.0736	8.375	21.3755	32.869	25.5926	26.3897	...
1/16/1997	16.752	8.375	21.6552	32.5242	25.7087	25.67	...
1/15/1997	16.6058	8.625	21.4953	33.0989	24.7221	24.9503	...
1/14/1997	16.6351	8.9375	21.5353	33.5586	24.7802	25.1902	...
1/13/1997	16.5474	9.0625	21.4554	33.3287	24.3159	26.0298	...
1/10/1997	16.6643	9.125	20.7762	33.5586	24.3786	25.9099	...
1/9/1997	16.2842	8.875	20.9759	33.4436	23.9164	26.0298	...
1/8/1997	15.758	8.8125	20.5764	33.3287	23.6853	25.3101	...
1/7/1997	15.7872	8.75	20.6563	33.6735	24.032	26.0298	...
1/6/1997	15.8749	8.9375	20.5764	33.5586	23.512	25.4301	...
1/3/1997	15.9042	10.875	20.3367	33.3287	23.8009	26.2697	...
1/2/1997	15.2318	10.5	20.5764	33.0989	23.281	26.7495	...

Table 2: Return of market data: Daily return for 233 stocks in S&P500 for 1100 days

date	AA	AAPL	ABK	ABS	
5/11/2001	-0.039131472	-0.006521739	-0.011743119	-0.011479988	...
5/10/2001	0.01280812	-0.040867389	0.00479351	-0.002167183	...
5/9/2001	0.025780863	-0.024013024	0.013452915	-0.011627907	...
5/8/2001	-0.016097561	-0.015625	-0.003723008	-0.021556886	...
5/7/2001	-0.021479714	-0.030679612	0.007112781	0.001199041	...
5/4/2001	0.024700416	0.031650641	0.004135849	0.006334842	...
5/3/2001	-0.01064602	-0.061301241	0.022094334	0.015313936	...
5/2/2001	-0.000483676	0.025453143	-0.01903509	-0.015973478	...
5/1/2001	0.002414522	0.017261671	-0.01393821	-0.006586826	...
4/30/2001	-0.014049362	-0.027099237	-0.007195202	0.010590015	...
4/27/2001	-0.002139387	0.061158364	-0.000921186	0.027993779	...
4/26/2001	0.039267593	-0.001213592	-0.019874176	0.033762058	...
4/25/2001	0.013263645	0.028714107	0.017276193	0.004846527	...
4/24/2001	-0.011379578	-0.009072165	-0.002383449	0.074652778	...
.	.	.	.	.	...
.	.	.	.	.	...
.	.	.	.	.	...
1/28/1997	0.020638404	0	0.003759009	0.010827807	...
1/27/1997	-0.016603874	-0.014814815	0	-0.014234666	...
1/24/1997	0.007432226	-0.02173913	-0.005609226	0.007169923	...
1/23/1997	-0.027122649	0.003630524	0	-0.010638922	...
1/22/1997	-0.021237309	-0.003617391	-0.003724004	-0.01742062	...
1/21/1997	-0.025865882	0.018444172	0.005619742	0.010560921	...
1/20/1997	-0.006846828	0.0112	-0.001871301	-0.00699139	...
1/17/1997	0.019197708	0	-0.012916066	0.010601337	...
1/16/1997	0.008804153	-0.028985507	0.007438835	-0.017363115	...
1/15/1997	-0.001761336	-0.034965035	-0.001857415	-0.013698426	...
1/14/1997	0.005299926	-0.013793103	0.003724004	0.006897959	...
1/13/1997	-0.007014996	-0.006849315	0.032691252	-0.006850703	...
1/10/1997	0.023341644	0.028169014	-0.00952045	0.003438625	...
1/9/1997	0.033392563	0.007092199	0.019415447	0.003447479	...
1/8/1997	-0.0018496	0.007142857	-0.003868069	-0.010239506	...
1/7/1997	-0.005524444	-0.020979021	0.003883089	0.003423862	...
1/6/1997	-0.001842281	-0.17816092	0.011786573	0.006897959	...
1/3/1997	0.044144487	0.035714286	-0.011649268	0.006942829	...

2.3 Formula for correlation matrices

Let the rate of daily return matrix from Table 2 be X. Then, correlation matrix C is given by

$$C = W^T W \tag{1}$$

Where W is the normalized rate of the return matrix such that

$$W_{ij} = \frac{x_{ij} - x_{.j}}{s_j} \tag{2}$$

where $x_{.j}$ and s_j indicates the mean and standard deviation of jth column of matrix X(jth stock), respectively.

Note that

$$\sum_{i=1}^{n} w_{ij} = 0, \quad \sum_{i=1}^{n} {w_{ij}}^2 = 1$$

- **Theorem 3**

  Correlation matrix C is positive definite.

  pf)

$$x^T C x = x^T W^T W x = (Wx)^T (Wx) = \| Wx \|^2 \geq 0$$

Thus, all the eigenvalues of correlation matrix would be positive and the corresponding eigenvectors will be orthonormal.

In fact, *Matlab* provides a built-in function named *corrcoef* so that you can easily construct the correlation matrix.

2.4 Central Limit Theorem for eigenvalues of random matrices

There is a **Central Limit Theorem** concerning the distribution of eigenvalues of a symmetric matrix of a sufficiently large size. In the limit of very large matrices, the distribution of its eigenvalues has universal properties, which are to a large extent independent of the distribution of the elements of the matrix. [1]

Especially, when the concerned matrix is a financial correlation matrix such that $C = W^T W$, CLT can be stated in the following way.

- **Theorem 4 : Central Limit Theorem for eigenvalues**

  Let the correlation matrix $C = W^T W$ be such that W is a rectangular matrix of size $M \times N$, so that C is $M \times M$. M will be the number of assets, and N, the observations(days). Let $\rho(\lambda)$ be the *density of eigenvalues* of C.

  Then, in the limit $N, M \longrightarrow \infty$, with a fixed ratio $Q = N/M \geq 1$, one has:

$$\rho(\lambda_c) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda_c)(\lambda_c - \lambda_{min})}}{\lambda_c} \tag{3}$$

  where

$$\lambda_{min}^{max} = \sigma^2 (1 + 1/Q \pm 2\sqrt{1/Q}) \tag{4}$$

  with $\lambda \in [\lambda_{min}, \lambda_{max}]$ and where σ^2 is the variance of elements of W(=1 in our case).

Although this results are only valid in the limit $N \longrightarrow \infty$, we will see that the random matrices with the same size of the stock market data will fit the formula well.

3 Numerical results

3.1 Eigenvalues of correlation matrices

Figure 1 indicates the distribution of eigenvalues for the correlation matrix constructed from 233 stocks in S&P500 for 1099 days(from 1/2/1997 to 5/11/2001). Notice that most of them are between 0 and 2. There are several data which are a lot larger than others. The largest one is 42.3759. This indicates the movement of market itself.

Now, consider the case of a pure random matrix. What will happen if matrix X is not rates of return of stock prices, but just a pure random matrix(say, Gaussian or uniform distribution)? Will the distribution be just the distribution predicted in CLT theorem in section 2.4? Will the matrix size be big enough to satisfy the CLT?

To answer the question, consider any random matrix of the same size with the correlation matrix for stock data(generated from matrix of the size 1099×233). I generated two kinds of random matrices, whose elements are normally distributed and uniformly distributed, respectively. [3] The result is amazing.

Figure 2 shows 4 kinds of generated eigenvalue distribution. The dashed(-.-.-) line is the theoretical result from equation 3. The circled(o) line is the eigenvalue distribution of correlation matrix, whose underlying matrix X is Gaussian distribution. The star(*) indicates that of uniform distribution. Notice that both of the results from random matrices

[3] I wanted to try with other distributions such as Levy distribution, but Matlab only provides *rand* and *randn* only.

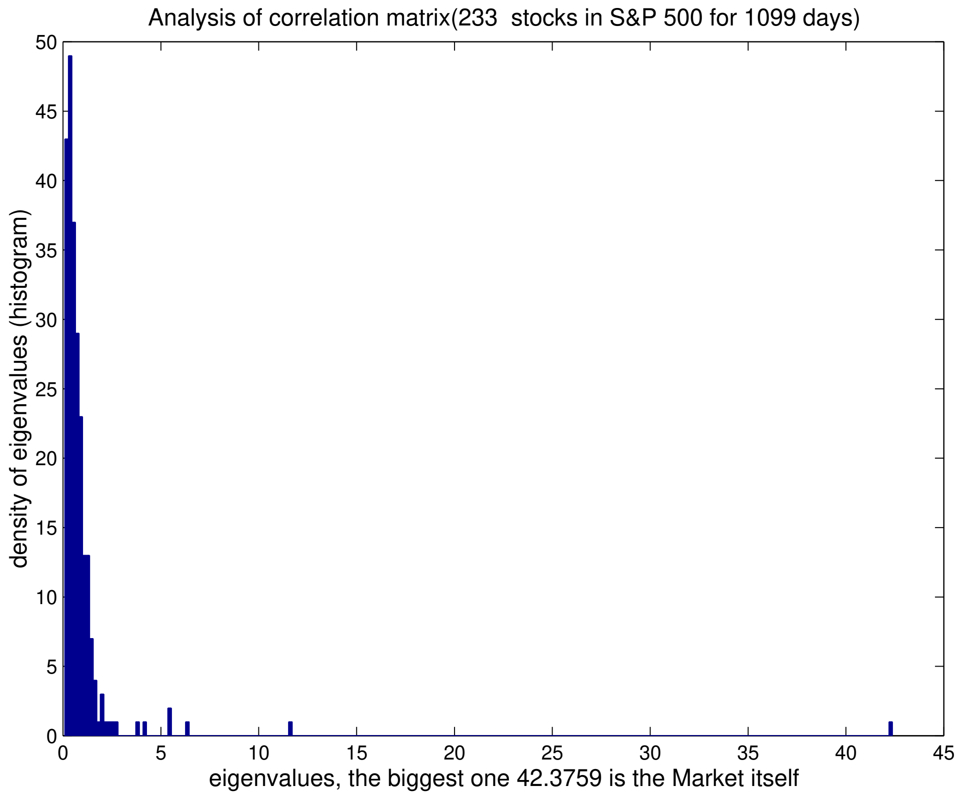

Figure 1: Histogram of eigenvalues: Density of eigenvalues for correlation matrix. The biggest eigenvalue is about 42, which is more than 20 times bigger than other eigenvalues.

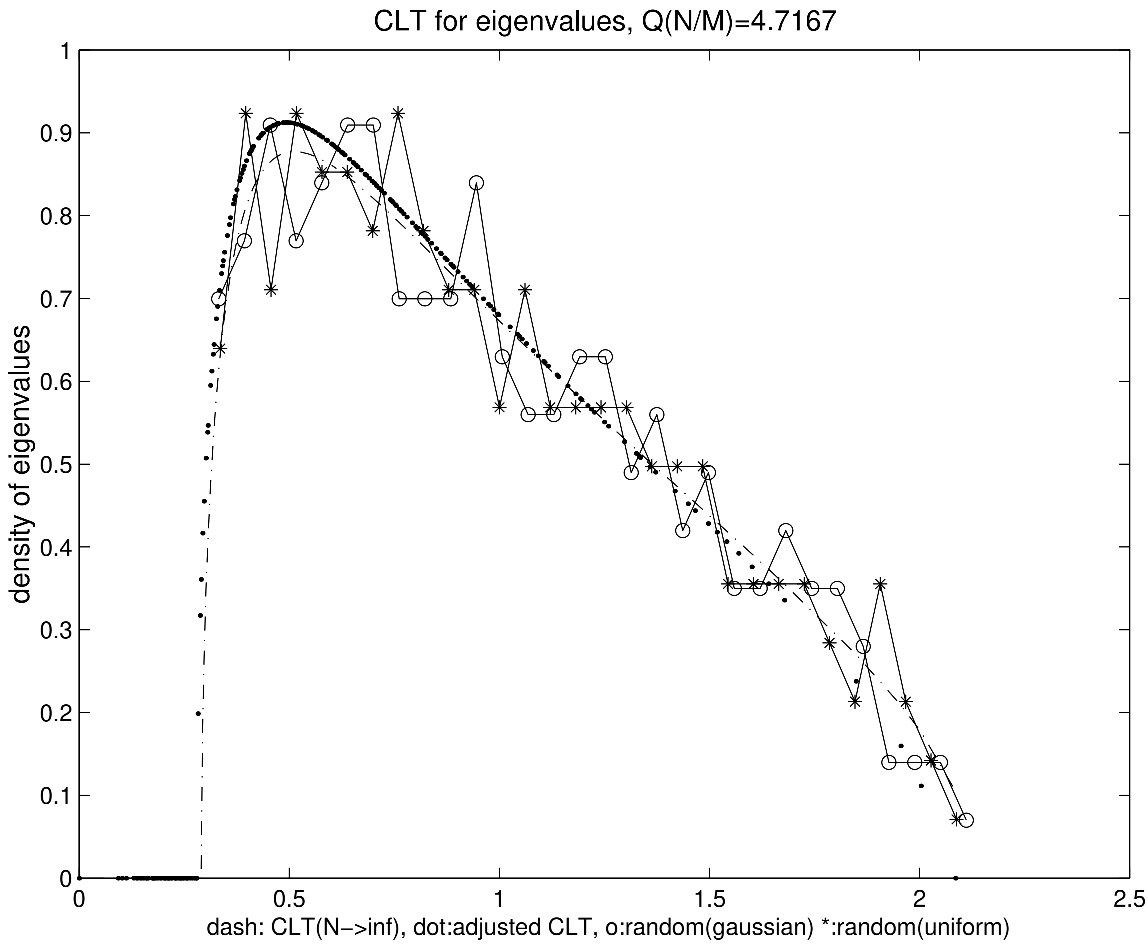

Figure 2: CLT for random matrices: The dashed(-.-.-) line is the theoretical result from equation 3. The circled(o) line is the eigenvalue distribution of correlation matrix, whose underlying matrix X is Gaussian distribution. The star(*) indicates that of uniform distribution. The dots express the adjusted CLT

overlaps well with the theoretical CLT. Thus, we can check three results here. First, CLT works well with size of 1099×233 random matrices. Second, it works well regardless of the way underlying random elements are generated. Note that even if we generated an arbitrary random matrix of size 1099×233, we still need to *normalize* this and construct the symmetric correlation matrices with diagonal 1. Lastly, it is known that *adjusted CLT* is more suitable than pure CLT for stock correlation matrix. Thus, we consider the effect of deviated large eigenvalues and adjust them. That is, instead of using $\sigma^2 = 1$, we adjust

$$\sigma^2_{adjust} = 1 - \frac{\sigma^2_{max}}{\sum_{i=1}^{M} \sigma_i^2 - \sigma^2_{max}} = 1 - \frac{\sigma^2_{max}}{M - \sigma^2_{max}} = 1 - \frac{\lambda_{market}}{M - \lambda_{market}} \tag{5}$$

where $\sigma^2_{max} = \lambda_{market} = 42.3759$ indicates the variance caused by the largest eigenvalue from others. Note that the eigenvalues represent the (unscaled) variance and the summation of all the eigenvalues are just $M = 233$.[4]

The dotted(...) graph indicates the distribution of adjusted CLT.

In the next picture, we will compare the adjusted distribution with eigenvalues of correlation matrix of real data.

The conjecture is that most of the small eigenvalues are just noise. To see this, we compare them with the result of CLT given in section 2.4. Figure 3 shows the distribution of lower part(except the deviations from large eigenvalues) of the correlation matrix and theoretical prediction of CLT. Under the assumption that financial correlation matrix is purely random, the graph should coincide with that of CLT. We can see that they fits approximately, not exactly. [5]

Note that 12 eigenvalues are bigger than theoretical maximum($\lambda_{max} = 2.0507$) from adjusted CLT. We will refer them as *information part* from now on. The ratio of real largest eigenvalue(market eigenvalue) versus λ_{max} is $19.8677 \simeq 20$.

The most unsure part is those below theoretical minimum($\lambda_{min} = 0.2799$). 45 eigenvalues are below this value in my case, which would not occur in original paper.

Should I treat this part as just noise? Or Should I think it contains information, because they are also deviated from CLT region? At this point, I will just regard them as being in the noise band. (In case of pure random matrices, the eigenvalues are all above the theoretical minimum. See Figure 2)

3.2 Eigenvectors of correlation matrices

Now, we turn to the eigenvector analysis.

In case of correlation matrices(C) based on random matrices(X), the components of an eigenvector itself is normally distributed. It does not make any difference whether the random matrix is generated from Guassian or uniform distribution.

[4]The mathematical proof for why the summation of all the eigenvalues of correlation marices are the same with th matrix size will follow in the revised paper.

[5]Actually, this is one thing that I got a little different result from Dr. Bouchaud's [1], which shows that the two graphs almost overlap each other.

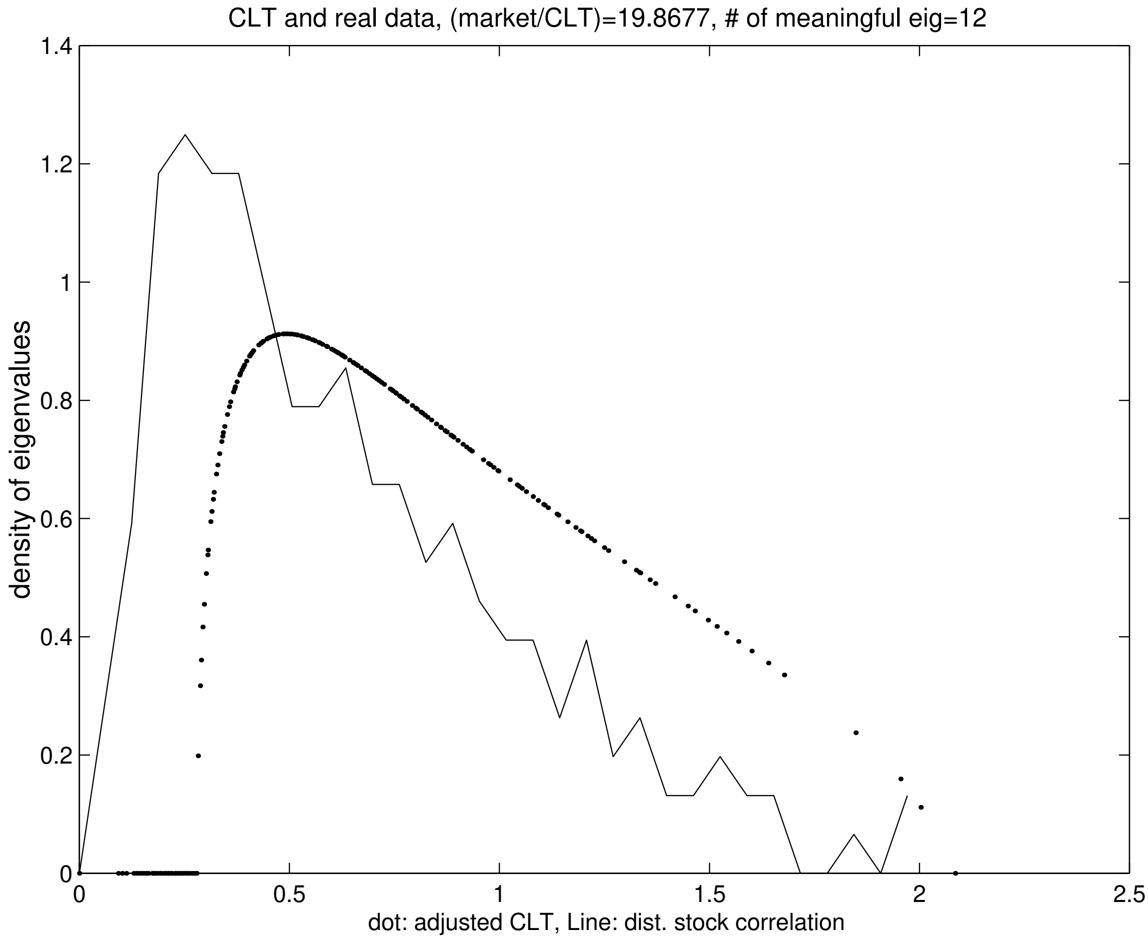

Figure 3: CLT and original eigenvalues: Distribution of small eigenvalues for correlation matrix is compared with theoretical prediction. My graph is somewhat different with theoretical CLT.

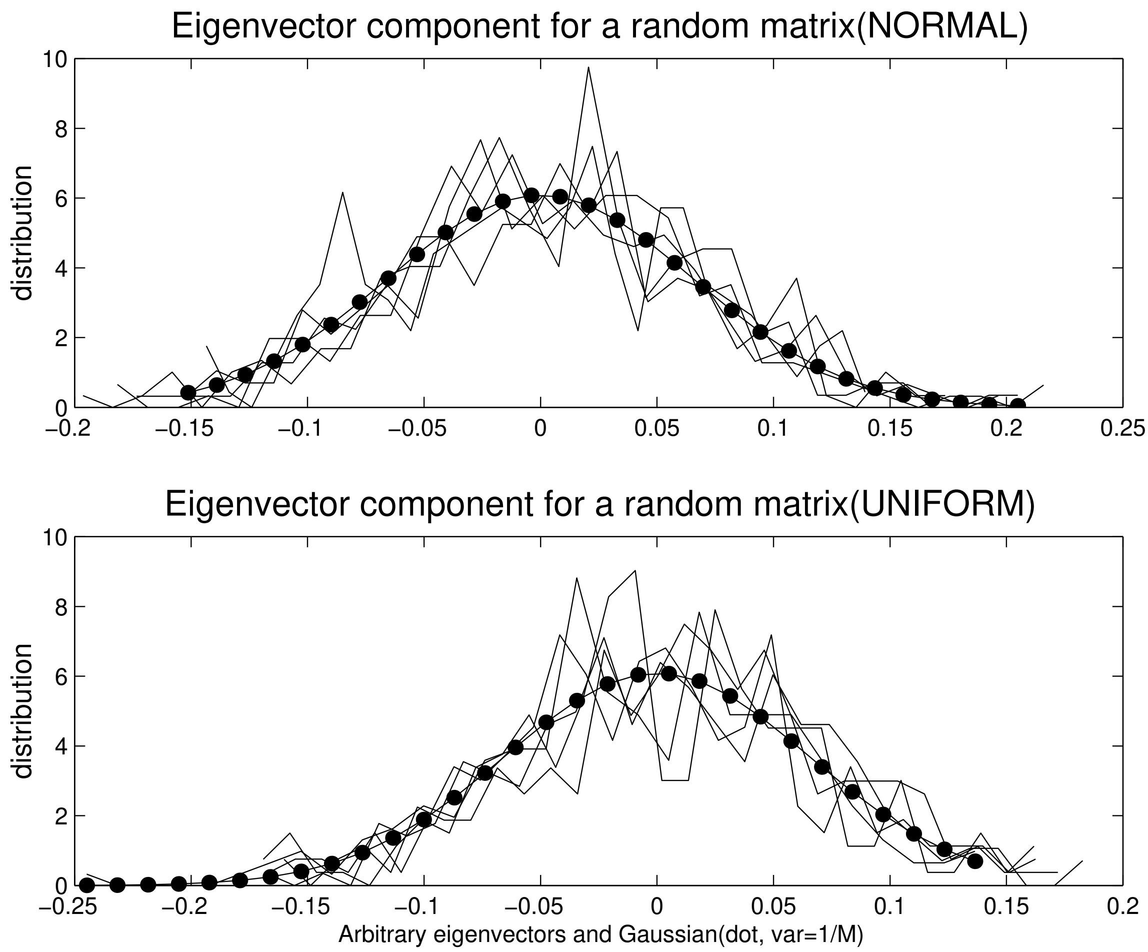

Figure 4: Distribution of components of 5 eigenvectors randomly selected for correlation matrices with underlying random matrices.

Figure 4 guarantees that the components of each eigenvectors are also normally distributed with variance $1/M$. The reason for scaling factor $1/M$ is so simple. Each element of the component just follows standardized normal distribution. When we consider the density of them, we accumulate them and divide by M, just like the case of *mean* in order statistics. This scaling holds for other coming distributions also. Thus, when we compare the distribution of components of eigenvectors with Gaussian distribution, all of them would have the scale of variance $1/M$.

However, if we consider the case of correlation matrix from stock market data, the result is different. We observe that the eigenvectors corresponding to a few of the largest eigenvalues have different distribution.

Upper part of Figure 5 shows the deviation of the components of *market eigenvector*. They form a peak within a short interval (0.04, 0.07), assuring the prediction that the market eigenvector has approximately equal component value around $1/M{=}1/233 = 0.0043$.

However, as we move on eigenvectors corresponding to smaller eigenvalues, the distribution becomes Gaussian, indicating that they are just noise.(Lower part of the picture)

Figure 6 shows the actual components of the first 4 largest eigenvalues. Note that eigenvector components of 4th largetst eigenvalue is already almost a gaussian distribution.

Isn't this amazing that the distribution of eigenvector components are Gaussian in case of no information?

3.3 Duration of information

In previous sections, we observed eigenvalues and eigenvectors and figured out that the large eigenvalues are correspondent to information part and others to noise. Will the information be time dependent?

To check this, we divide the time interval into two sub periods. ($1099/2 \simeq 549$ days) Then the eigenvalues and eigenvectors for each period are calculated. If we sort the eigenvectors as the way of descending magnitude of corresponding eigenvalues, we can check the relationship them by dot product(inner product). If the eigenvectors remain consistent, the dot product should be significant. If they are relatively independent, the dot product will be near 0.

Figure 7 shows the result. For the information part(approximately first 10 eigenvectors), the values of inner product are of significant size, implying certain duration of information. Especially the eigenvectors corresponding to the market have inner product 0.96196, showing that they are almost identical. (Market remains the same.)

However, for the noise part, corresponding eigenvectors are uncorrelated to each other so that the inner products fall into the noise level($1/\sqrt{M} = \pm 0.0655$). However, this would not guarantee that the information will be consistent in the future, because it is suggested that there appears to be a genuine time dependence in the structure of the meaningful correlations. [1]

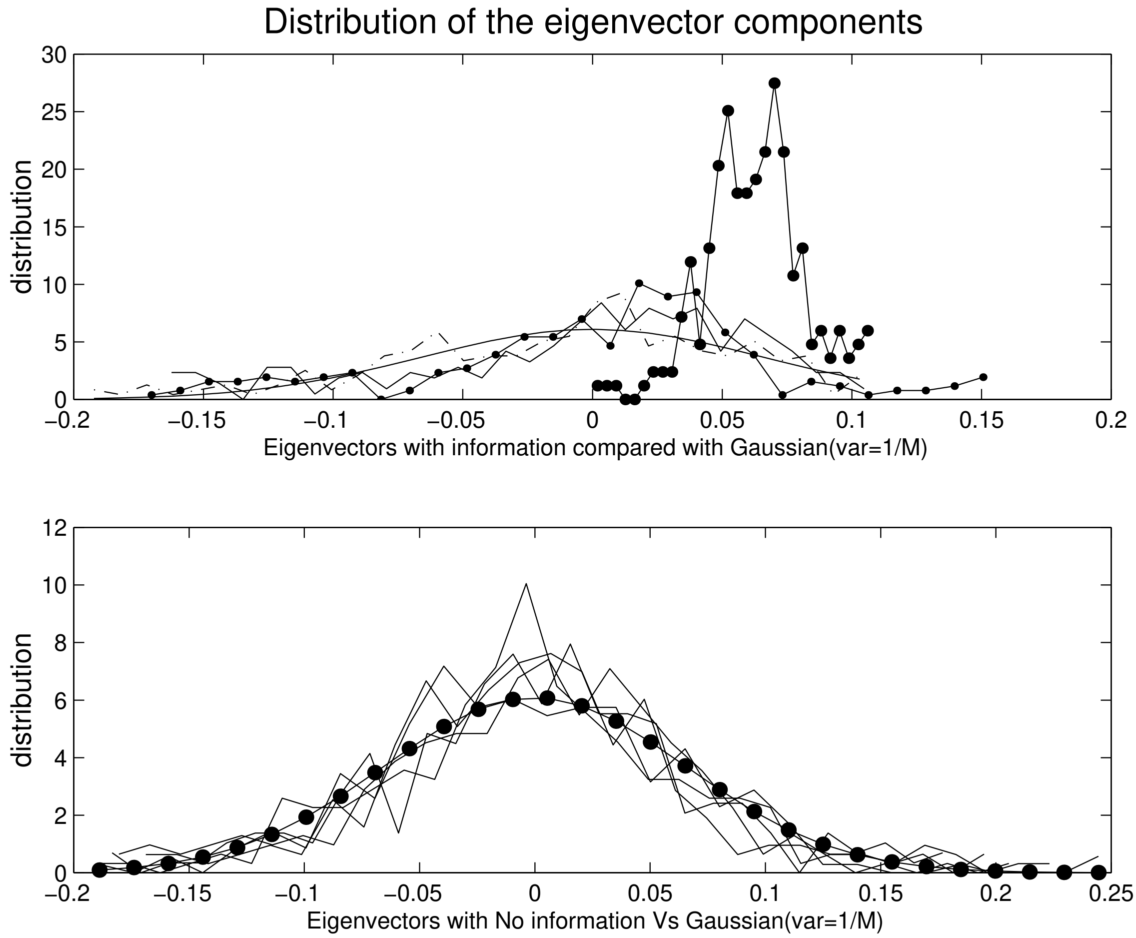

Figure 5: Distribution of components of eigenvectors correlation matrices from stock market data: Upper part indicates that 1st, 2nd and 3rd eigenvector(according to the magnitude of the eigenvalue) are not Gaussian distribution. They certainly contains information! Meanwhile, lower part shows that eigenvectors in the middle have just Gaussian distribution for their components

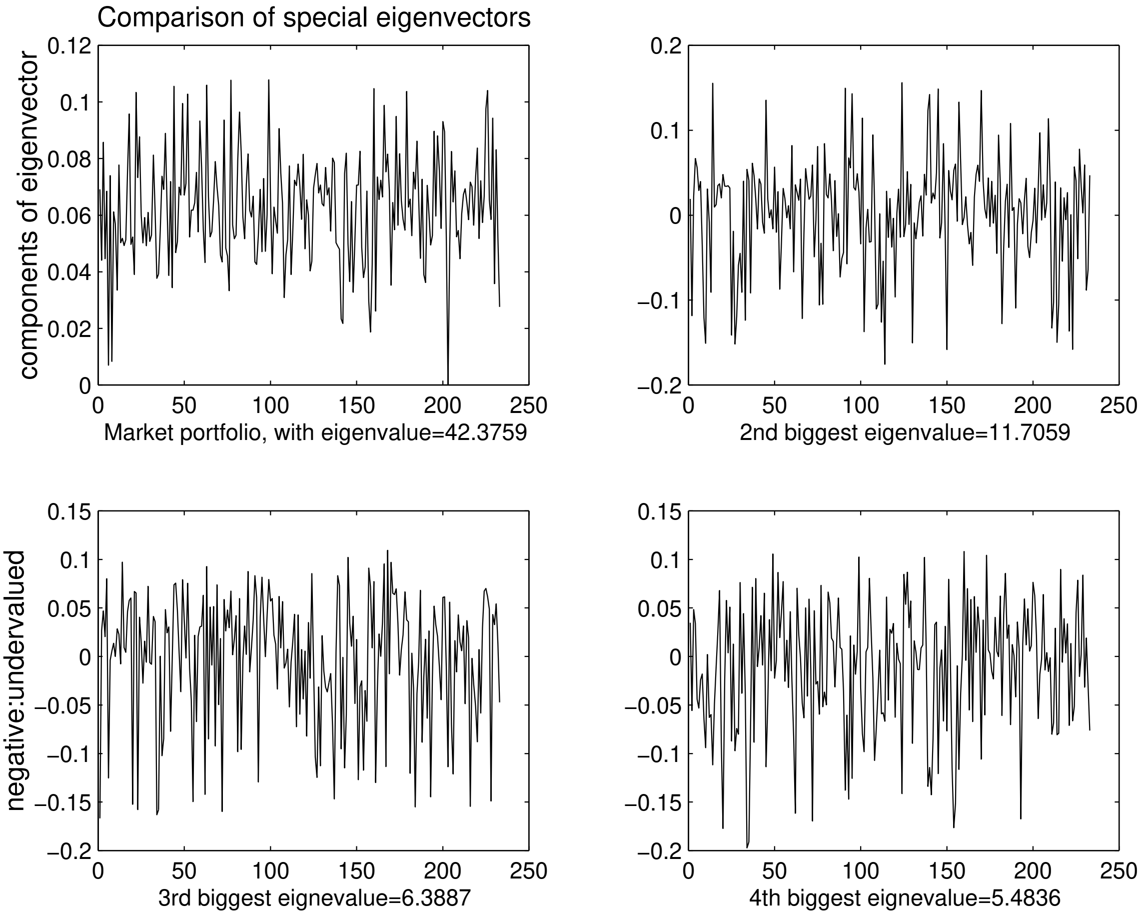

Figure 6: The actual components of the first 4 largest(42.3759, 11.7059, 6.3887, 5.4836) eigenvalues: The eigenvector components diffuses(?) into Gaussian distribution

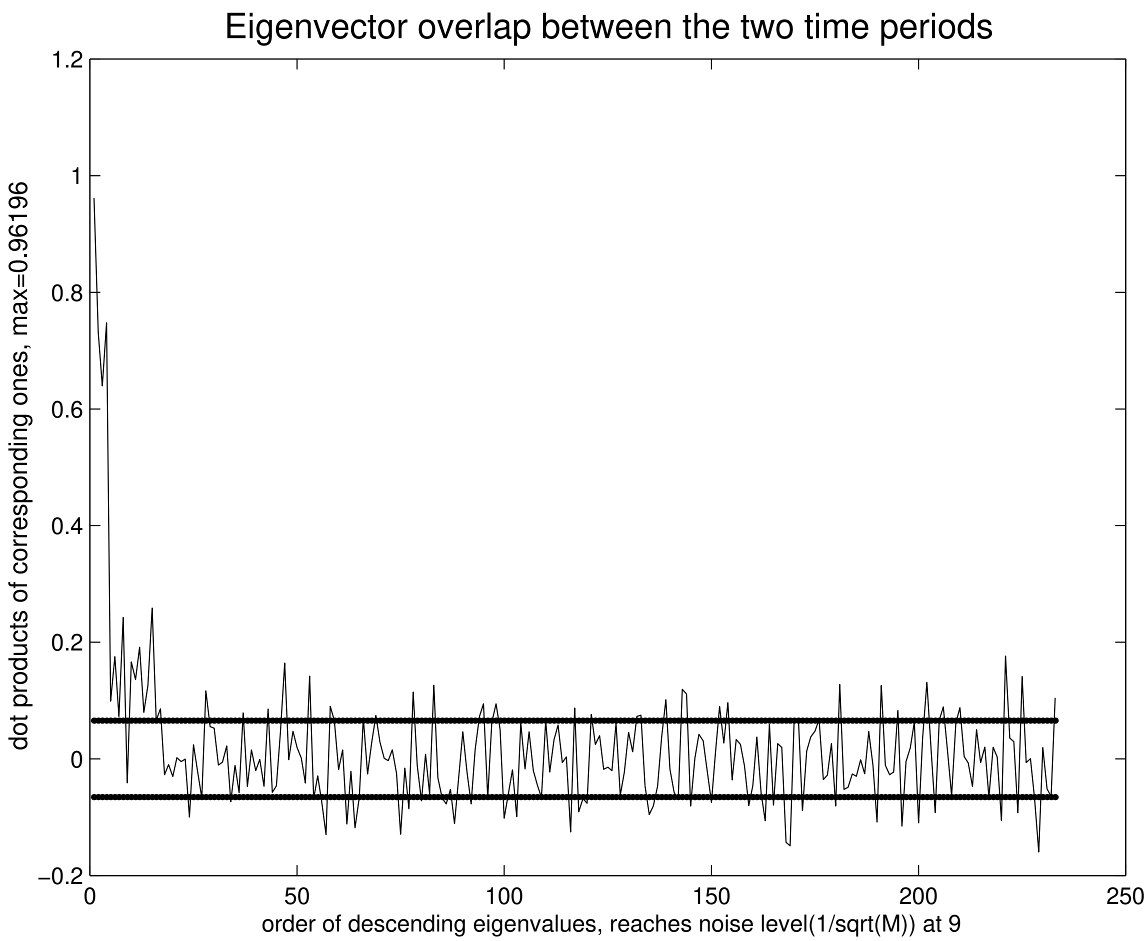

Figure 7: Eigenvector overlap between two time period: The first one(market) has 0.96196, indicating persistent eigenvector. But at 9th eigenvector, we reach noise level.

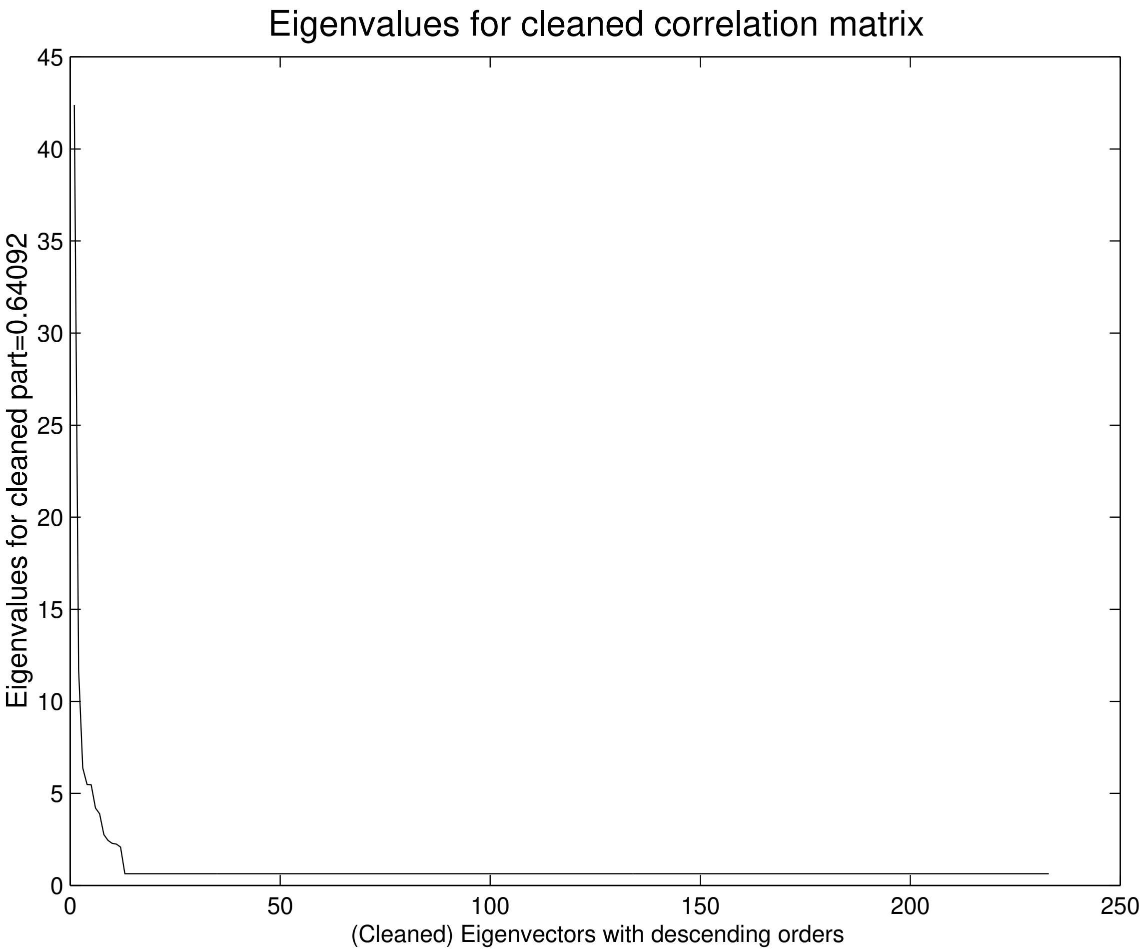

Figure 8: Flattening of eigenvalues for noisy part: The 12 largest values are untouched, but other parts are replaced by their mean

3.4 Noise removal from correlation matrix

Now that we know what is information and what is noise, how should we *clean*(or remove the noise from) the correlation matrix? Dr. Bouchaud suggests to flatten the noise part by replacing it into identity matrix, still keeping the trace. Maintaining the same trace is important, since the trace or the sum of all eigenvalues are always equal to $M = 233$, which are the number of stocks. Thus, we replace the noise part to the mean.

Figure 8 shows this simple operation.

Then we construct *cleaned* correlation matrix by multiplying the original eigenvector matrices.

$$C_{cl} = VD_{cl}V^T \tag{6}$$

where V is the matrix having eigenvectors in their column such that $C_{cl} \times V = V \times D$

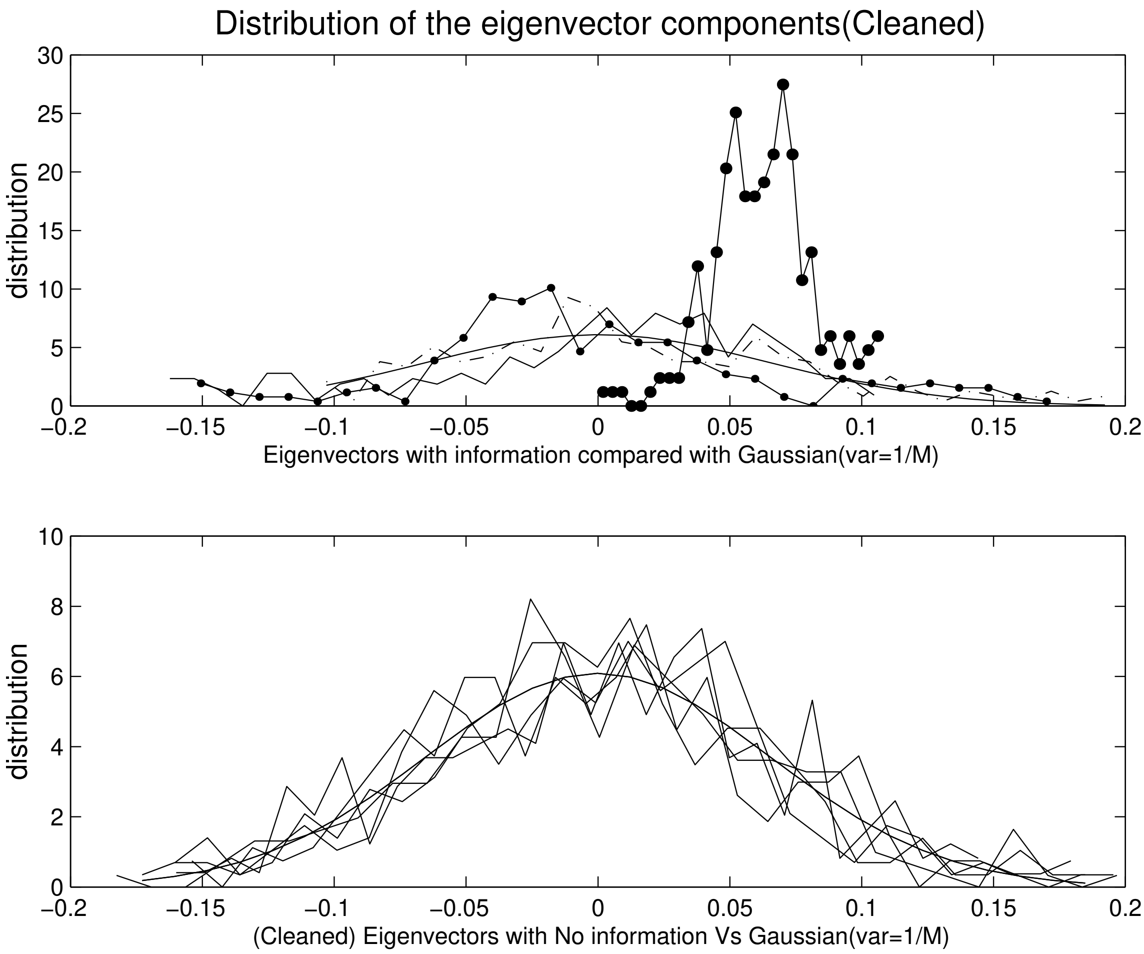

Figure 9: Distribution of components of eigenvectors correlation matrices from stock market data : almost the same as Figure 5

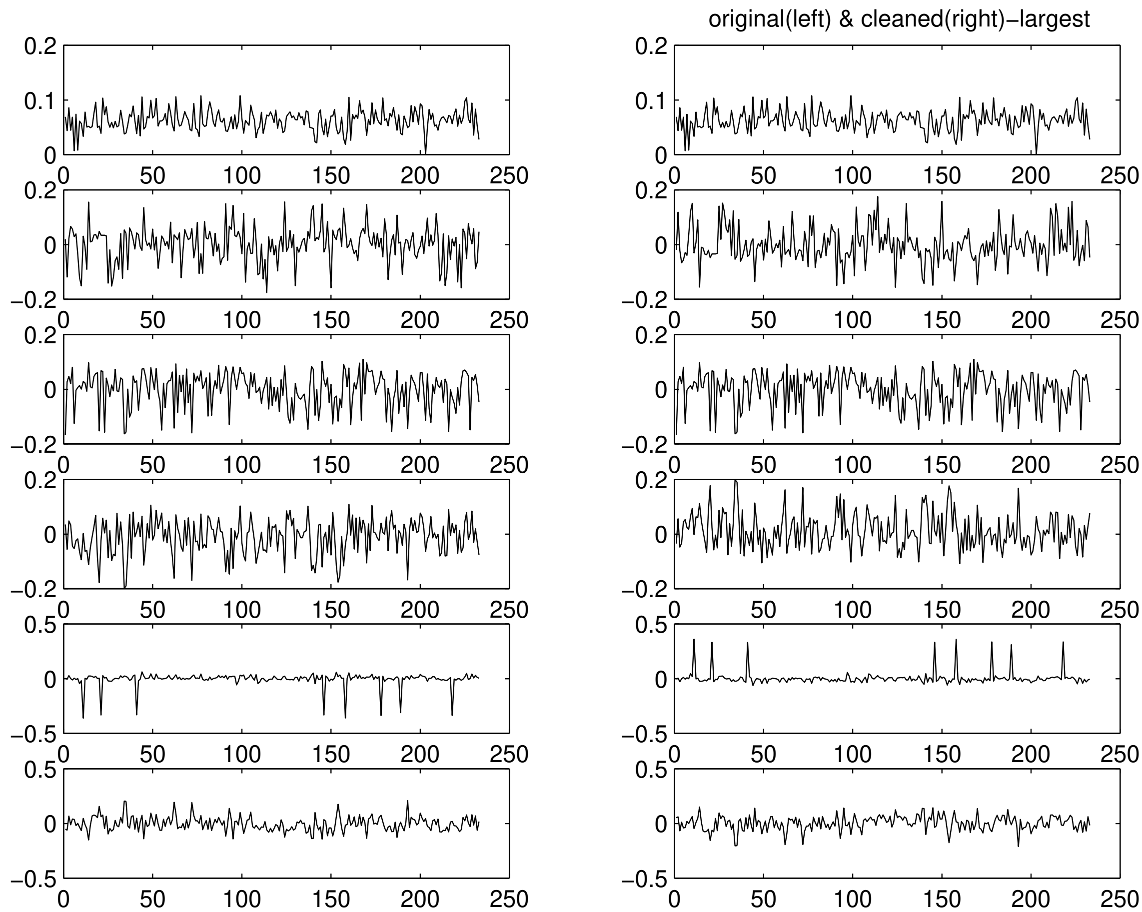

Figure 10: Comparison of the 6 largest eigenvector components between original and cleaned one: They are the same up to the sign

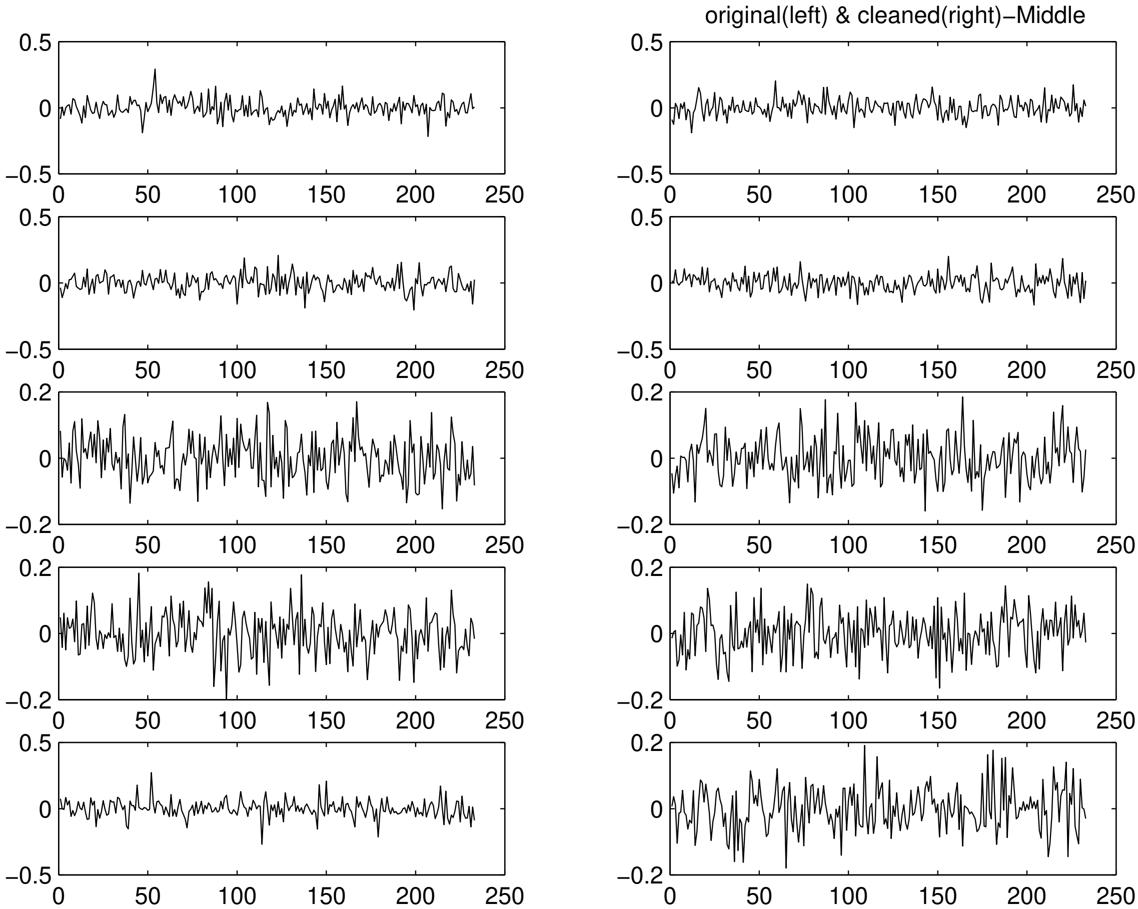

Figure 11: Comparison of the eigenvector components lying in the middle between original and cleaned one: The remain similar, but changed(sample collected by 40th, 80th, 120th, 160th and 200th, descending magnitude of eigenvalues)

Note that the eigenvectors corresponding to the information part(intact part) remains the same. (Figure 9, Figure 10)

However, the eigenvectors corresponding to the noise part are not the same anymore. See Figure 11. [6]

3.5 Construction of optimal portfolio

Now we cleaned the correlation matrix. How do we construct the optimal portfolio from this? This is another problem(named *quadratic programming*) and already widely known and practiced in industry. We do not pursue more in this paper. However, the idea is as follows:

Given expected return, we minimize the risk(Variance). Or given risk(variance) tolerance, you maximize the return. [7]

An email from Dr. Bouchaud The optimal composition is a *SUPERPOSITION* of different eigenvectors, not of one them in particular (in the general case). [8]

4 Conclusion

By reproducing Dr. Bouchaud's original paper with recent data(Jan 1997 to May 2001), we recapitulated some properties of random matrices and correlation matrices, approaching to their eigenvalues and eigenvectors. Matrix obtained from the stock data is different from just random matrices. However, many part of its eigenvalues and eigenvectors are noisy. Large eigenvalues apart from the region for CLT contains information and it is durable although it is still time dependent. Typical number of significant eigenvalues about 10. Market eigenvalue is about 20 times bigger than other eigenvalues in noise band. Typically, eigenvalues corresponds to the variance caused by the eigenvector in the correlation matrix.

CLT for eigenvalue distribution of random correlation matrices fits well for large N, M.

Market eigenvector has approximately equally weighted eigenvector components of magnitude $1/M$, where M is the number of stocks. Eigenvectors in noise band are composed of elements from Gaussian distribution. In fact, regardless whatever random matrices you generate(either Gaussian or Uniform distribution), the eigenvector components form Gaussian distribution.

By taking average of eigenvalues in noise band, we can partially remove the noise and construct a *cleaned* correlation matrix. It has the same eigenvectors for *information part*, but has different(however, still similar) eigenvalues in the *noise band.*

[6] If they remain the same, there is no reason for cleaning, since we will have the same eigenvectors anyway.

[7] What I witnessed during my internship in a fixed income company really showed the trade-off between *risk* and *return*. The way they try to get a little bit higher return than the benchmark was, after all, enduring a little bigger risk under controllable limit.

[8] Sat, 12 May 2001

Acknowledgement It is a pleasure to thank Dr. Jean-Philippe Bouchaud for discussion and help.

References

[1] Jean-Philippe Bouchaud(2000), *Theory of Financial Risks*, 1st ed. Cambridge University press

[2] LLoyd N. Trefethen, David Bau, III(1997) *Numerical linear Algebra*, 1st ed. Siam

[3] Serge Lang(1987), *Linear Algebra*, 3rd ed. Springer-Verlag