Econophysics - Term project
Time-lagged partial correlations of financial time series with high dimensional conditions

Sebastian Gemsheim

April 30, 2015

Contents

1 Introduction

In every field of science associated with the observation of systems, regardless if they are on the scale of the universe or on quantum mechanical scales, one strives to understand the interplay between units or parts of this system and the underlying rules for these interactions. These investigations will then lead to a discrete set of observations collected on a finite time scale, i.e. discrete data points. In order to analyze the data obtained from the experiment one has to use sophisticated methods to understand the relationships inside the investigated system. Later, they can then be tested against mathematical models or can yield to the knowledge about relevant parameters of a system.
One speaks about a time series if one observable was measured with successive measurements at different times. The so-called "time series analysis" is a very valuable and often used tool in almost every field of sicence. This includes fields like neuroscience, geophysics, atmospheric physics, economics, network theory, signal processing, astronomy, etc.

In the case of this term paper the experiment will be the New York Stock Exchange and the units of the system are different stocks. The data is obtained on a time scale of 3 years and we wish to understand the relations between different stocks and the interplay of economic sectors and subsectors. The concept of correlation is used to identify significant relations between different stocks. Once it is clear which units of the system interact with each other it hopefully enables us to understand the nature of these relations.

A lot of work has been done already for the observed correlations of two observables and their associated time series, but only recently the influence of third observables on such correlations has been studied. Therefore it is of great interest to understand this topic furter, not only for economics. The main idea of this paper is to use all available data of all units of a system and include them into the determination of the correlation. If one pictures such relations as a correlation based network where all units are somehow connected (or sets of disconnected networks), it becomes clear that only the consideration of all observables at once can lead to an understanding of the single links. This in turn can allow us to build models and try to predict the future outcome, once we know the 'pure' relations.

The paper is organized as follows: In Section 2 I introduced the data under scrutiny. Sections 3 and 4 will give brief definitions of the concepts of correlation and previous work in this area. Afterwards, I compare the time-lagged correlation matrix with the time-lagged partial correlation matrix for synchronous times and for different lags in Section 5 and 6. Section 7 will be about the temporal behaviour of correlations and the eigenvalue spectra of such matrices is investigated in Section 8. Finally I conclude in Section 9 and give an outlook to future work.

2 Data set under investigation

The dataset in use contains the price return values of the 100 largest capitalized stocks in New York Stock Exchange (NSYE) for the period of 2001 - 2003. It contains 748 trading days with data points in 5 minutes intervals, where 78 data points account for one single trading day with 390 minutes trading time. Therefore, the total length of one return time series is 58344 entries. The price return time series is defined as the difference between two adjacent price logarithms in time.

$$\tilde{x}_i(t) = \ln(P_i(t) - \ln(P_i(t-1))$$

Without loss of generality, I rescaled every return time-series to mean zero and unit variance.

$$x_i(t) = \frac{\tilde{x}_i(t) - \mu_{\tilde{x},i}}{\sigma_{\tilde{x},i}}$$

Here, $\mu_{\tilde{x},i}$ is the mean value and $\sigma_{\tilde{x},i}$ the standard deviation (or square root of the variance) of the time series for the ith stock. In general, after the analysis one could transform it back to the original values. Later this is useful to compare the resulting correlations with findings of Random Matrix Theory (RMT). In Section 3 the mean value and the variance are explained in more detail. The data is organized in the data matrix X with dimension $N \times T$.

The distribution of normalized returns for all 100 stocks is shown in a semilog plot in Figure 1 and has a tent shape. A closer look at the distribution around zero shows some anomaly which I account to round-off errors and finite precision in the price values. This behaviour is presented in Figure 2, but not further investigated.

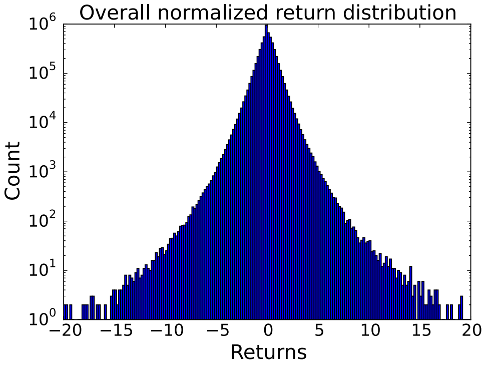

Figure 1: Distribution of price returns for all 100 stocks in semilog histogram.

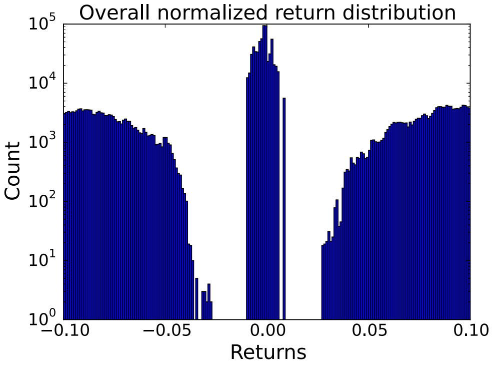

Figure 2: Anomaly in distribution of price returns for all 100 stocks.

3 Correlation and previous work

The investigation of covariances obtained from the time series under scrutiny has a long history. In general it shows how two quantities are related and how they change mutually, i.e. if one quanity increases who much does the other change. For example, if one increases and the other decreases the covariance will be negative. If the two quantities have no causal relation then the covariance and also the correlation is zero. The reverse is not necessarily true and one has to be aware of this. The covariance accounts only for linear dependencies and that has to be kept in mind.

It is defined as

$$\mathrm{Cov}(x,y) = \sigma(x,y) = \mathbb{E}\left[(x - \mathbb{E}(x))\,(y - \mathbb{E}(y))\right],$$

or the covariance matrix

$$\mathrm{Cov}(X,Y) = \Sigma_{XY} = \mathbb{E}\left[(X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y))^T\right] = \Sigma_{YX}^T,$$

for vector quantities $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. The expectation value is denoted by $\mathbb{E}(x)$ and can be estimated with the average value of a finite time series with

$$\mathbb{E}(x) = \frac{1}{T} \sum_{t=1}^{T} x(t).$$

If the time series is long enough this value will be sufficiently close to the real expectation value. The quantity $\text{Cov}(x, x) \equiv \sigma^2(x)$ is refered to as "variance" and $\sigma(x)$ is the associated "standard deviation".

The covariance quantifies the amount of mutual change whereas the correlation gives a qualitative measure of how strong two quantities are related. The correlation is defined as a normalized version of the covariance

$$\text{corr}(x, y) = \rho(x, y) = \frac{\sigma(x, y)}{\sqrt{\sigma(x, x)\sigma(y, y)}} \in [-1, 1]$$

and its values range from -1 to $+1$. For our normalized returns the covariance and the correlation have the same numerical value. The covariance estimate, and for normalized data also the correlation estimate, is obtained via

$$\Sigma(X, X) = \frac{1}{T} \sum_{t=1}^{T} X(t) X^T(t) = \frac{1}{T} X X^T$$

To great extent the correlations of synchronous time series have been studied in different field of researches. Especially the eigenvalue decomposition of the correlation matrix for multiple interacting quantities, e.g. fincancial stocks , has been proven useful. The eigenmodes can be related to the identification of clusters and subcluster in different sectors of economy.

Lately in finance [1, 2, 3] and other sciences [4, 5, 6], also the partial correlation has been taken into account since it is a tool to measure the underlying interactions in a correlation based network. The mathematical framework for it is explained in Section 4. Even fewer has be done for time-lagged correlations where one looks at $\rho\left(x(t), y(t+\tau)\right)$ for some lag τ. In the case of time-lagged correlation matrices for multiple quantities it is not symmetric as in the case for synchronous correlation matrices. This introduces another difficulty since many results of RMT cannot be applied and, additionally, the corresponding eigenvalues and eigenvectors will be complex. I will address this problem more in Section 8. Recent publications are [7, 8, 9].

4 Partial Correlation

The partial correlation is the correlation between the residuals of two time series $X = \{x_1, x_2\}$ after the linear regression to a set of conditional quantities/time series $Y = \{y_1, \ldots, y_p\}$. It represents the mutual dependence of two quantities on each other without the influence of possible other correlated quantities. This means that the variables x_1 and x_2, respectively, will be projected on the linear space of Y and the difference between this "conditional mean" and the original time series x_1 and x_2, respectively, gives the residual [10]. In the case of only one conditional quantity $Y = \{y\}$ the conditional mean writes

$$\hat{x}(y) = \mathbb{E}(x) + \frac{\sigma(x, y)}{\sigma(y, y)} \left(y - \mathbb{E}(y)\right)$$

$$= \sigma(x,y) \cdot y \qquad \text{for unit variance and zero mean,}$$

and is exactly a linear regression. The partial variance between the two residuals is then given by

$$\sigma(x_1, x_2|y) = \sigma_{12|y} = \mathrm{Cov}\left(x_1 - \hat{x}_1(y), x_2 - \hat{x}_2(y)\right).$$

For a more general case with an arbitrary finite number of y_i the conditial mean is written as (zero mean)

$$\hat{X}(Y) = \Sigma_{XY}\Sigma_{YY}^{-1}Y$$

We can then write down the full formula for the partial covariance matrix $\Sigma_{XX|Y}$ for X given Y and make use of the bilinearity of the covariance and the symmetry property of $\Sigma_{YY} = \Sigma_{YY}^T$, and therefore also its inverse $\Sigma_{YY}^{-1} = \left[\Sigma_{YY}^{-1}\right]^T$.

$$\begin{aligned}
\Sigma_{XX|Y} &= \mathrm{Cov}\left(X - \Sigma_{XY}\Sigma_{YY}^{-1}Y, X - \Sigma_{XY}\Sigma_{YY}^{-1}Y\right) \\
&= \underbrace{\mathrm{Cov}\left(X, X\right)}_{=\Sigma_{XX}} + \mathrm{Cov}\left(\Sigma_{XY}\Sigma_{YY}^{-1}Y, \Sigma_{XY}\Sigma_{YY}^{-1}Y\right) - \mathrm{Cov}\left(X, \Sigma_{XY}\Sigma_{XY}^{-1}Y\right) \\
&\quad - \mathrm{Cov}\left(\Sigma_{XY}\Sigma_{YY}^{-1}Y, X\right) \\
&= \Sigma_{XX} + \mathbb{E}\Big(\Sigma_{XY}\Sigma_{YY}^{-1}Y\underbrace{\left(\Sigma_{XY}\Sigma_{YY}^{-1}Y\right)^T}_{=Y^T\Sigma_{YY}^{-1}\Sigma_{YX}}\Big) - \mathbb{E}\left(\Sigma_{XY}\Sigma_{YY}^{-1}YX^T\right)^T - \mathbb{E}\left(\Sigma_{XY}\Sigma_{YY}^{-1}YX^T\right) \\
&= \Sigma_{XX} + \Sigma_{XY}\Sigma_{YY}^{-1}\mathbb{E}\left(YY^T\right)\Sigma_{YY}^{-1}\Sigma_{YX} - \left(\Sigma_{XY}\Sigma_{YY}^{-1}\mathbb{E}\left(YX^T\right)\right)^T - \Sigma_{XY}\Sigma_{YY}^{-1}\mathbb{E}\left(YX^T\right) \\
&= \Sigma_{XX} + \Sigma_{XY}\underbrace{\Sigma_{YY}^{-1}\Sigma_{YY}}_{=\mathbb{1}}\Sigma_{YY}^{-1}\Sigma_{YX} - \left(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right)^T - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}
\end{aligned}$$

With $\Sigma_{XY} = \Sigma_{YX}^T$ the last line reduces to

$$\Sigma_{XX|Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = \begin{pmatrix} \sigma_{11|Y} & \sigma_{12|Y} \\ \sigma_{21|Y} & \sigma_{22|Y} \end{pmatrix}.$$

The result holds also for the case of non-vanishing mean values and variances unequal one for X, Y [10]. Another way of looking at it is as a regression model like

$$x_i(t) = \alpha_i + \beta_i Y(t) + \epsilon_i(t) \qquad \text{with} \quad \alpha \in \mathbb{R} \quad \beta, Y(T) \in \mathbb{R}^p$$

and the coviarance between the residuals

$$\mathrm{Cov}(x_i, x_j|Y) = \mathbb{E}\left(\epsilon_i(t), \epsilon_j(t)\right).$$

The coefficients can be identified as $\alpha_i = \mathbb{E}(x_i)$ and $\beta_i = \Sigma_{XY}\Sigma_{YY}^{-1}$. For example, whenever somebody applies autoregressive models or removes the market mode his intention is to idenitify the underlying raw relationships between single participants on the stock market. The partial correlation can then be calculated from

$$\rho_{12|Y} = \frac{\sigma_{12|Y}}{\sqrt{\sigma_{11|Y}\sigma_{22|Y}}}$$

Again, a partial correlation of zero does not necessarily imply causal independence [10].

5 Synchronous correlation matrix

One of the main ideas of my work was to include all available data into the calculation of the partial correlation matrices, i.e. for every pair i, j of stocks the condition is a $N-2$ dimensional vector for synchronous correlation matrices. For time-lagged correlation matrices with lag τ the dimension will be $\tau N - 2$. For large lags this will be computationally expensive, but since we have the fortune to live in a time of cheap computational power the calculation time for small and medium lags ($\tau = 1, \ldots, 15$) is quite short.

In order to evaluate this idea I compare it to the factor model Capital Asset Pricing Model (CAPM).

5.1 CAPM and market mode removal

The CAPM is a linear regression of the returns of one stock onto the so-called market mode x_m.

$$x_i(t) = \alpha_i + \beta_i x_m(t) + \epsilon_i(t)$$

The market mode results from a coordinate transformation into a space where the data is "orthogonalized" and has no correlations. This is obtain throught the diagonalization of Σ_{XX}, or r_{XX} respectively. Here X denotes the whole data set. The covariance matrix can then be written as $\Sigma = U \Lambda U^T$ where Λ is a diagonal matrix with the eigenvalues of Σ in the diagonal. The columns of U are the corresponding unit eigenvectors. The dataset can then be transformed via $\tilde{X} = XU$. The idea is that the eigenvalue spectrum typically has one very large separated eigenvalue and will be the main mode of the market. Assuming that the eigenvalues are ordered like $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ the market mode is calculated as [7]

$$x_m = \sum_{j=1}^{N} u_{1j} x_j$$

It was shown that the entries of eigenvector u_{1j} are almost equal and, hence, show market behaviour. After the linear regression onto the market mode only the residuals are considered and investigated.

I think that this can only approximate the raw partial correlations between two stocks without the influence of third stocks. If the number of considered time series is low (e.g. in other fields outside of finance) than the estimate will get worse since the market mode depends also on the stock x_i itself ($x_m \sim x_i$).

The residual time series are then arranged in the data matrix X_{res}. An example of how a time series looks against the market mode is illustrated in Figure 3

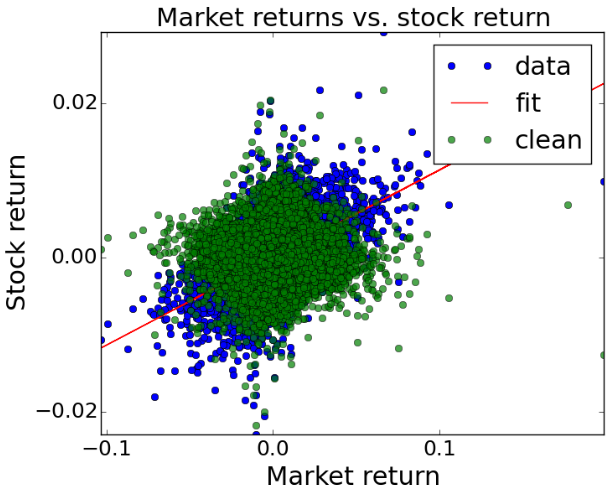

Figure 3: Scatter plot of Market mode against one stock time series (blue) and its associated linear regression. The green dots show the scatter plot for the residuals of this time series after the market mode removal.

5.2 Comparison to removed market mode time-series

In order to compare the performance of my method I calculated the following synchronous correlation matrices:

- No removal: ρ
- Since we have a finite data set there will be noise terms. I scrambled the data matrix X with random permutations and, therefore, destroyed all the correlations between the different stocks. It is denoted by ρ_{scr}
- Market mode removed: ρ_{res}
- Partial correlation with condition on all $N-2=98$ other stocks: ρ_p

First I check if the random permutations of the data destroyed the correlations. This is the case as can be seen in Figure 4 and the distribution function of the off-diagonal elements is Gaussian. This matrix is also called Wishart ensemble and its associated Wishart distribution. For correlation matrices, and also time-lagged correlation matrices, of a finite data set with T observations for gaussian variables the upper noise limit for uncorrelated data is [11] $(N=100, T=58344)$

$$\rho_{\max} \sim \sqrt{2\ln\left(N^2\right)/T} = 0.01777.$$

As stated before for Figure 1, the distribution function has rather a tent shape than a gaussian distribution. But as a rough limit approximation this limit will be sufficient.

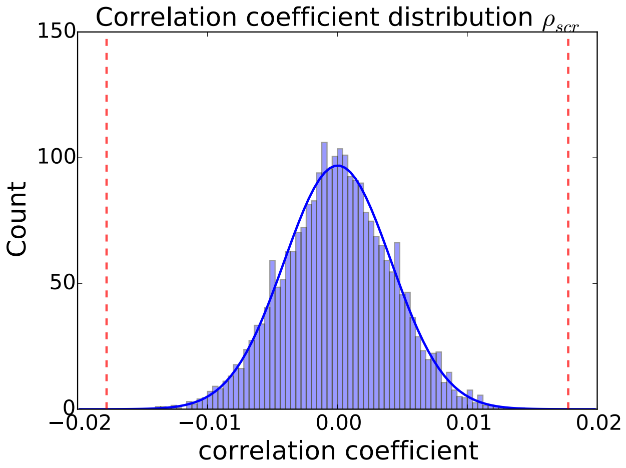

Figure 4: Distribution of correlation coefficients for the scrambled data. The fit shows that the distribution is gaussian. The red dashed lines indicate the upper noise limit and it can be seen that the distribution lies within these limits.

All correlation matrices are shown in Figure 5. The distinct dark blue line (almost no correlation) in the untouched data indicates a stock almost decoupled from the rest of the stocks. This stock belongs to the Newmont Mining Corporation which is one of the largest producers of gold. Therefore, it will be closely related with the gold price which is known to be very stable. Hence, it is not very correlated with the rest of the stocks. In comparison with the removed market mode data it can be seen that most of the cross-correlations in the untouched data are due to the market mode. But there is also a noticable difference between ρ_{res} and ρ_p. The distributions of the correlation coefficients is presented in Figure 6

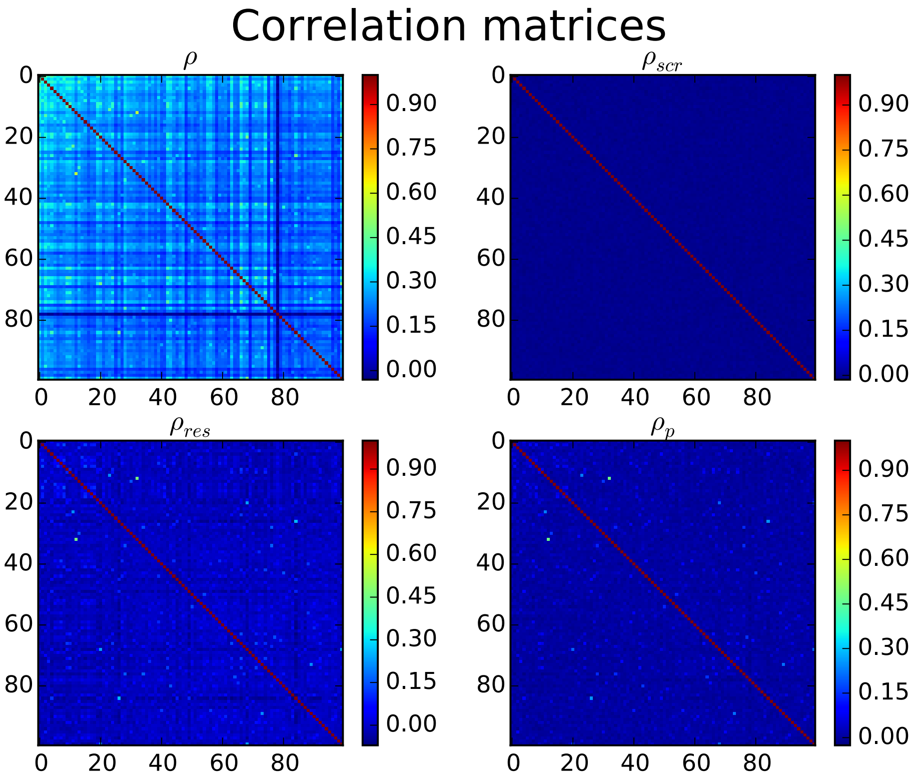

Figure 5: Synchronous correlation for untouched data, scrambled data, market mode removed data and the partial correlation. There is only very little negative correlation in all unscrambled correlation matrices.

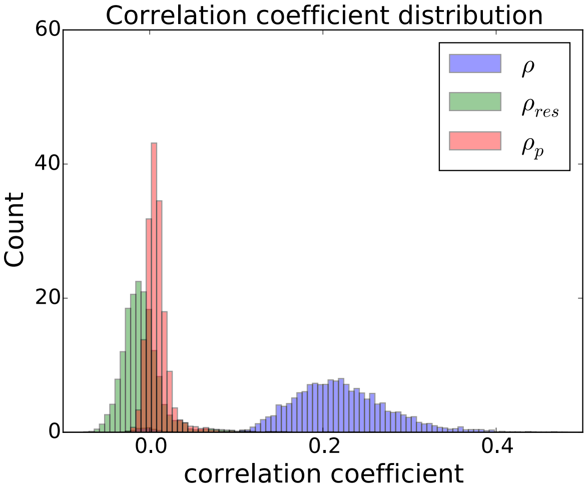

Figure 6: Distribution of synchronous correlation coefficients without autocorrelations. All three distributions show fat tail behaviour. There is a distinct difference between ρ_{res} and ρ_p.

In the next step I compare the real-valued eigenvalue spectrum of ρ_{res} and ρ_p in order to see difference. This is of particular interest since the eigenvalue outside of the noise spectrum are used to identify clusters and subclusters in the correlation based networks of stocks. The eigenvalue spectrum of a real symmetric random gaussian matrix was pointed out by Wigner [12]. If the matrix has the size $N \times N$ the spectrum is given by

$$f(x) = \frac{\sqrt{4N\sigma^2 - x^2}}{2\pi N\sigma^2}$$

In order to test the data the eigenvalue spectrum of ρ_{scr} is compared to Wigner's semicircle law in Figure 7. The variance σ^2 of the gaussian distribution is empirically obtained from the fit in Figure 4. Notice that the diagonal elements of the correlation matrix were set to zero, or in other words, I substracted the identity matrix $\mathbb{1}_N$. This accounts simply to a shift of the spectrum of one. This can be seen easily from the characteristic equation:

$$\det((\rho - \mathbb{1}) - \lambda\mathbb{1}) = \det(\rho - (1+\lambda)\mathbb{1}) = \det(\rho - \tilde{\lambda}\mathbb{1}) = 0 \quad \Rightarrow \quad \lambda = \tilde{\lambda} - 1$$

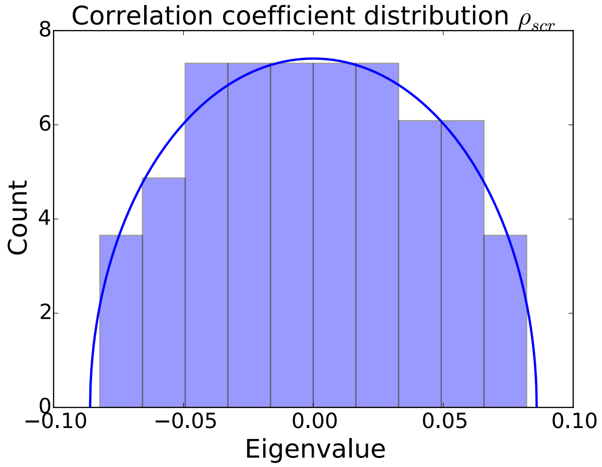

Figure 7: Eigenvalue distribution of the correlation matrix for scrambled data. The blue solid line describes Wigner's semicircle law for the variance fitted from the correlation coefficient distribution fit. They are in good agreement.

The two eigenvalue spectrums are shown in Figure 8. Both distributions show them same overall structure of large positive eigenvalues. But there are still differences, for example the value of the largest eigenvalue differs from each other. This will result and a different correlation based network and, maybe, also in different sector clustering. I will address this issue in the summary.

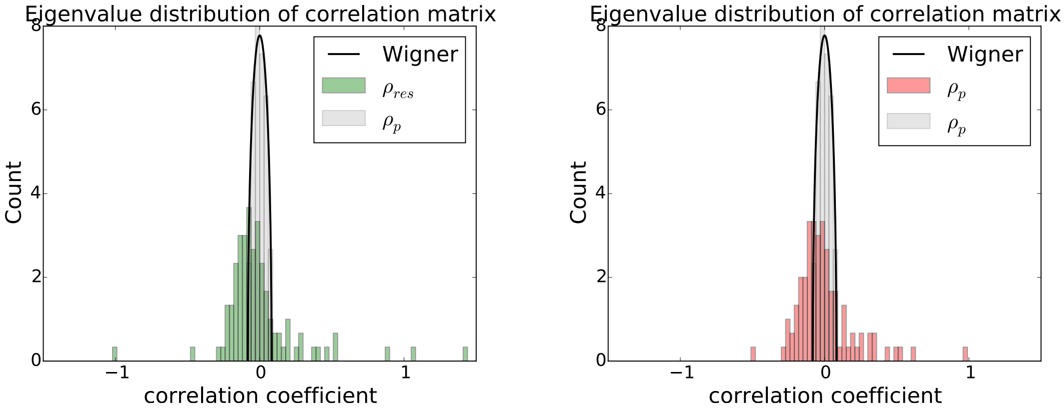

Figure 8: Eigenvalue spectrum of ρ_{res} and ρ_p. Wigner's semicircle law and the eigenvalue distribution for ρ_{scr} is also included to idenitify the eigenvalues outside the noise spectrum.

6 Time-lagged correlation matrix

The next step in my analysis is the investigation of time-lagged partial correlations (LPCs). This includes again a comparison with for the market mode removed data. First, I show the general structure of such asymmetric matrices for a lag of unit time and, secondly, the evolution of the matrix structure. The time-lagged correlation matrix for the market removed data set is obtained with

$$C^{\tau}_{res} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} X_{res}(t) X^{T}_{res}(t+\tau)$$

For the time-lagged partial correlation C^{τ}_{p} I use the same formula as for the synchronous partial correlation, but include every lagged time series (up to lag τ) in the conditions as well.

$$\begin{aligned} X &= \{x_i(t), x_j(t+\tau)\} \\ Y &= \Big\{ x_1(t), \dots, x_{i-1}(t), x_{i+1}(t), \dots, x_N(t), \dots, x_1(t+(\tau-k)), \dots, x_N(t+(\tau-k)), \dots, \\ &\quad x_1(t+\tau), \dots, x_{j-1}(t+\tau), x_{j+1}(t+\tau), \dots, x_N(t+\tau) \Big\} \end{aligned}$$

The matrix C^{τ}_{p} can then be calculated from the partial covariance matrix as explained in Section 4. Again, Y has a large dimension: $(\tau N - 2) \times (T - \tau)$.

6.1 Cause and reaction

The investigate the structure of LPCs for a lag of one unit time I define two quantities to measure how much one stock can influence other stocks after one unit time and how much it gets influenced from other stocks, or how much it reacts to other stocks.

$$\begin{aligned} \text{cause}(x_i) &= \sum_{n \neq i} C^{1}_{in} \\ \text{reaction}(x_i) &= \sum_{n \neq i} C^{1}_{ni} \end{aligned}$$

The autocorrelation entries are excluded because they give the same contribution to both measures and only the cross-correlation are of interest.
The three time-lagged correlation matrices C^1, C^1_{res} and C^1_p including cause and reaction are shown in Figures 9, 10 and 11.

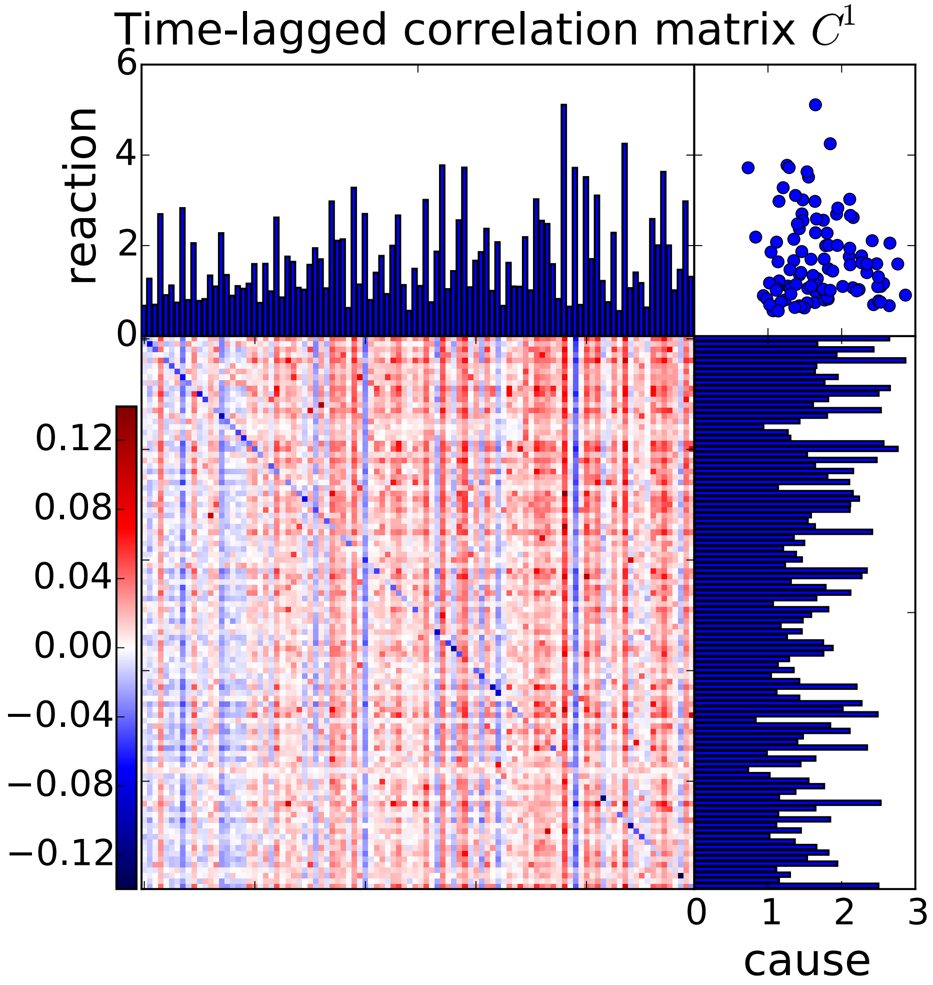

Figure 9: Time-lagged correlation matrix for the untouched data including cause and reaction. In the lower left plot blue pixels indicate negative correlation and red pixels represents positive correlation. A very distinct feature of the plot is the pattern of vertical lines which indicate that every stocks gets influenced almost equally by every other stock after one time lag. This accounts to the market mode. The upper right subplot shows the distribution of reaction versus cause. It can be seen that the distribution is stretched out due to the vertical pattern. Also the autocorrelations in the diagonal entries can be distinguished from the rest and show mostly negative correlations.

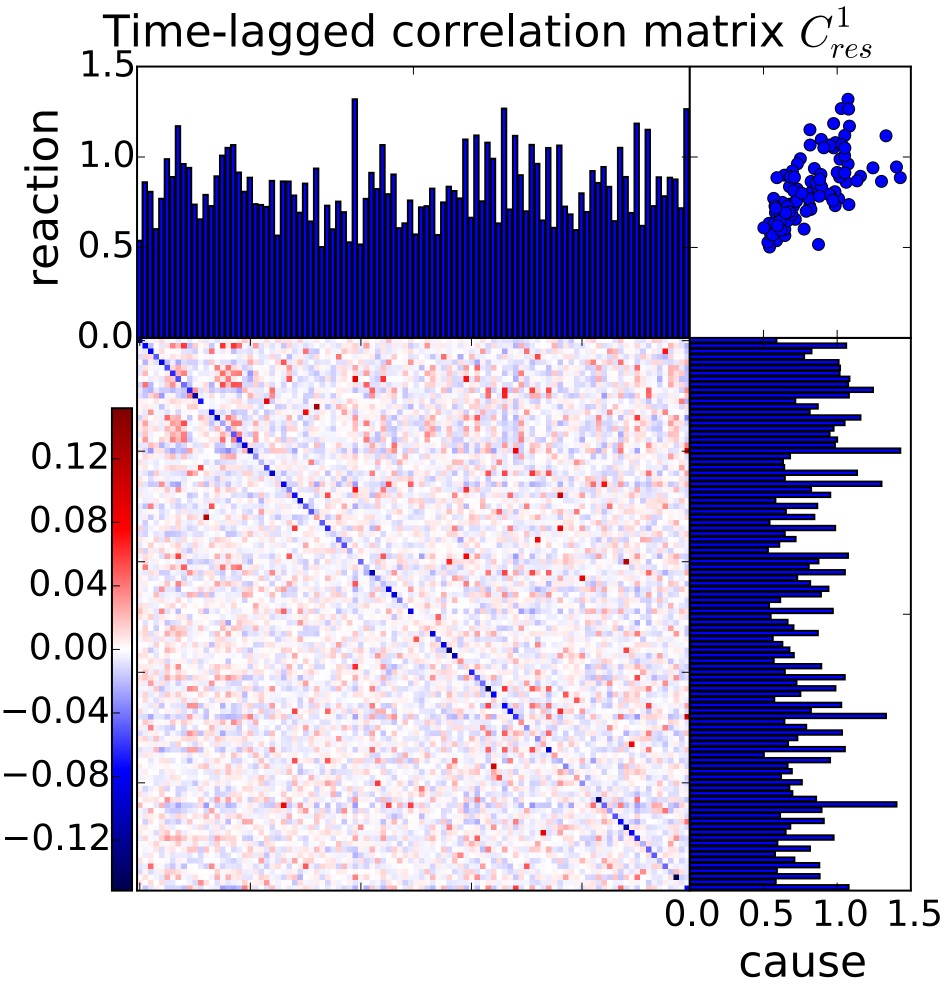

Figure 10: Time-lagged correlation matrix for the market mode removed data including cause and reaction. The pattern of vertical lines related to the market mode is not existing anymore. The autocorrelations still exhibit negative behaviour. Interestingly the distribution of cause and reaction shows a linear behaviour and seems to be correlated. Thus, stocks with a strong future influence on other stocks also tend to be more influenced by stocks in the past.

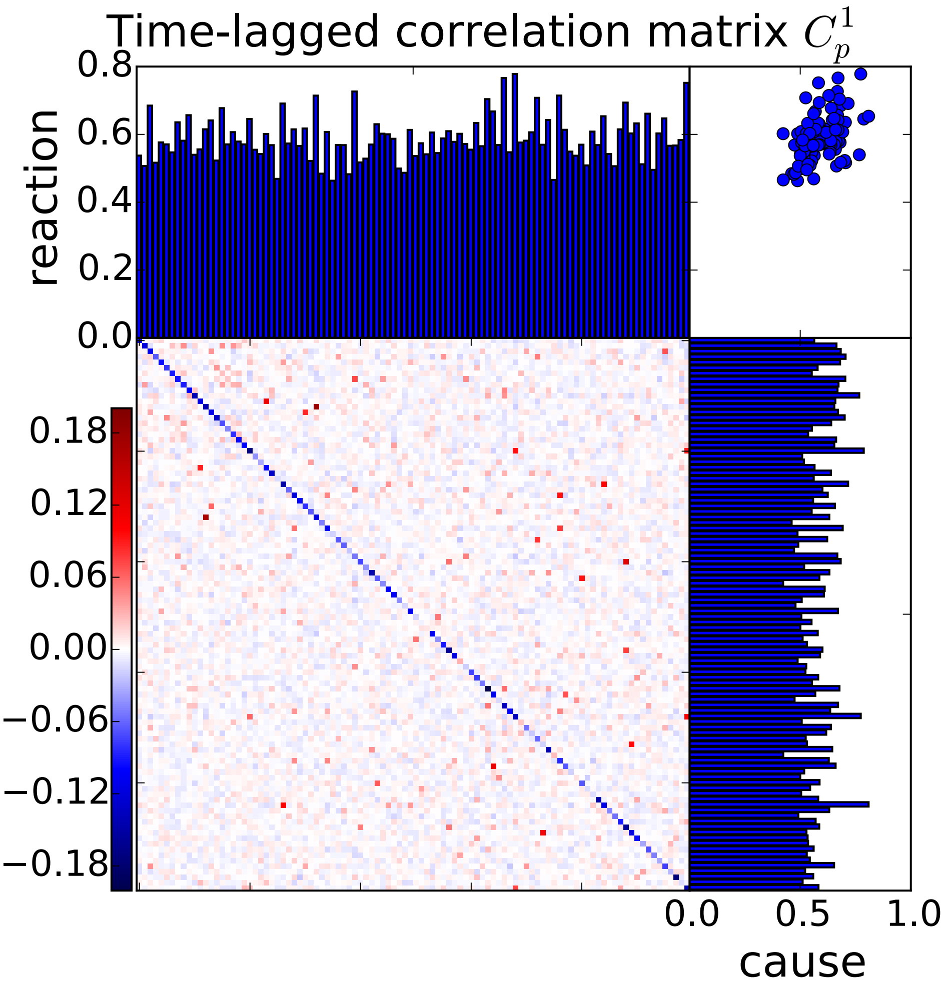

Figure 11: Time-lagged partial correlation matrix including cause and reaction. In comparison with C^1_{res} the lower left plot of C^1_p shows more distinct single pixels of strong correlation. A comparison of the scale indicates that the numerical values for the market removed data are lower. Therefore, the influence of other stocks seems to dampen the raw correlation between two stocks. As before, almost all of the autocorrelations are in a negative range and also for the LPCs cause and reaction seems to be positive correlated.

6.2 Lag evolution

Just a remark on the time scale of computation: With the high-level scripting language PYTHON and an ordinary computer the calculation of the time-lagged partial correlation matrix up to lag 10 is on a minute time-scale. For the range up to lag 30 it operates on an hour time scale. I believe the use of a programming language like C++ and parallel computing would speed up the calculation significantly.

In the following plots I compare the time-lagged correlation matrices of different lags for the untouched, scrambled and market mode removed data with the time-lagged partial correlation (Figures 12, 13, 14 and 15).

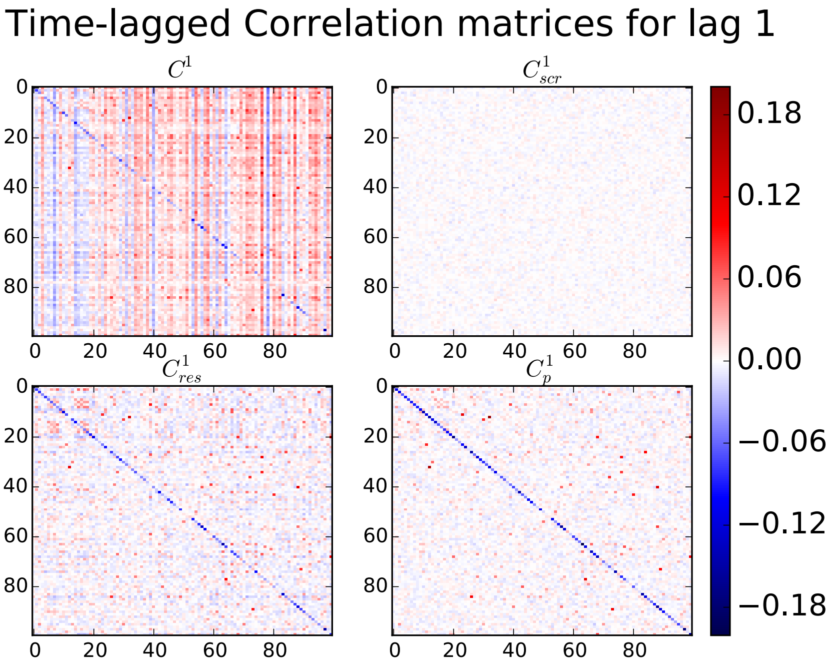

Figure 12: Time-lagged correlation for C^1, C^1_{res}, C^1_{scr} and C^1_p for lag 1.

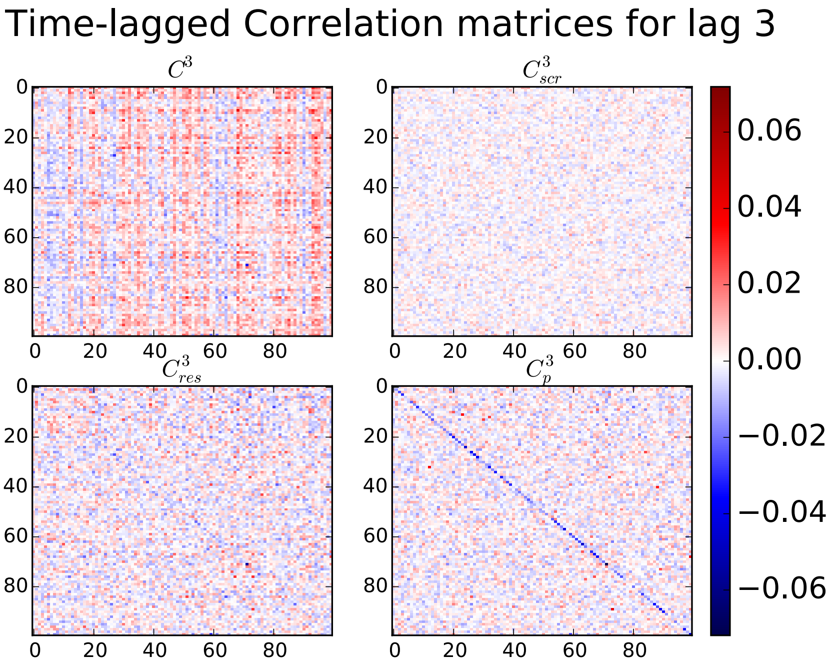

Figure 13: Time-lagged correlation for C^3, C^3_{res}, C^3_{scr} and C^3_p for lag 3. The time-lagged autocorrelations of C^3_{res} cannot be distinguished by eye anymore, whereas this is the case for C^3_P.

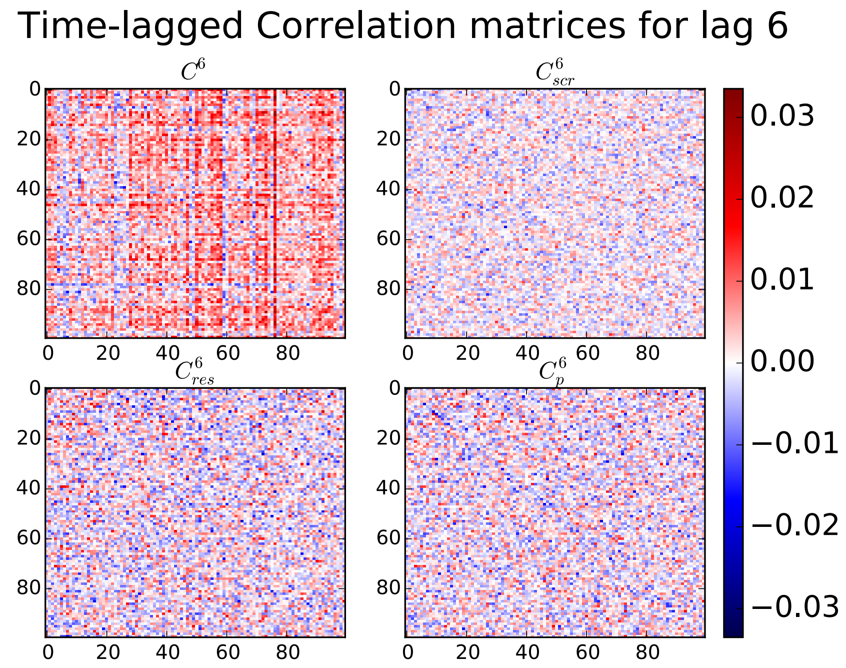

Figure 14: Time-lagged correlation for C^6, C^6_{res}, C^6_{scr} and C^6_p for lag 6. The time-lagged partial autocorrelations cannot be distinguished by eye anymore from the surrounding partial cross-correlations.

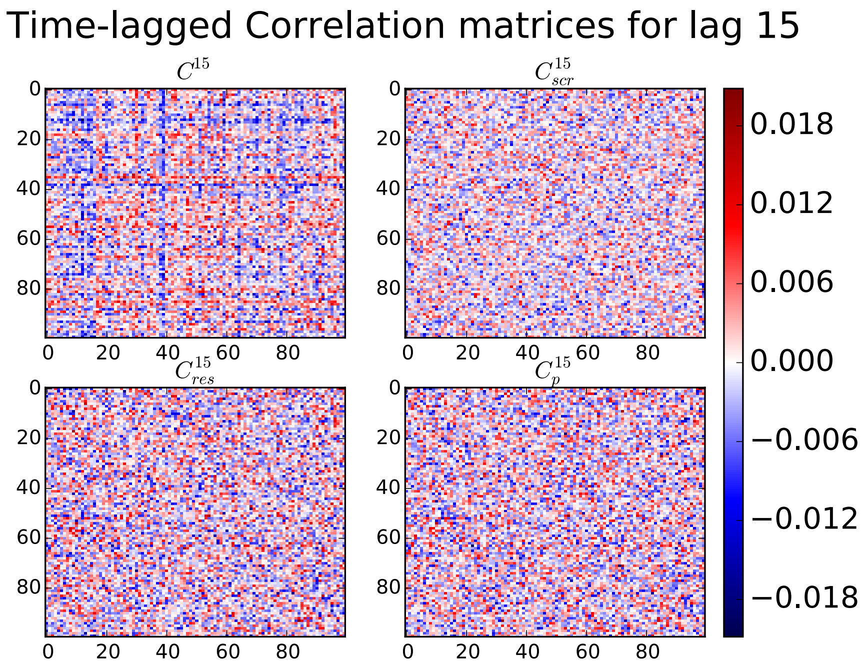

Figure 15: Time-lagged correlation for C^{15}, C^{15}_{res}, C^{15}_{scr} and C^{15}_p for lag 15. There is no distinct structure anymore for C^{15}_{res} and C^{15}_p.

Especially the case of $\tau = 3$ gives evidence that the raw correlations between two stocks are suppressed by the influence of many other influential stocks. For the removed market mode data the autocorrelations shift towards the noise spectrum, whereas the partial autocorrelation are still outside the noise spectrum.

7 Time evolution of autocorrelations and strong cross-correlations

In this section I want to have a closer look at the time evolution of single strong time-lagged cross-correlations (LPCCs) and of the partial autocorrelations. As pointed out in the section above the time scale of the decay time will be different for the time-lagged correlations of the market mode removed data and the LPCs because of the damping influence, due to correlations with all other stocks.

7.1 Partial autocorrelations

The autocorrelations of the market mode removed data and the partial autocorrelations is shown in Figures 16 and 17. As mentioned above for C^{τ}_{res} almost all autocorrelations evolved into the noise region at $\tau = 3$. Whereas for the partial autocorrelations of C^{τ}_{p} the same behaviour happens at $\tau = 6$. This shows again how the raw correlations to other stocks dampen. The noise limit given in [11] gives as very good approximate limit.

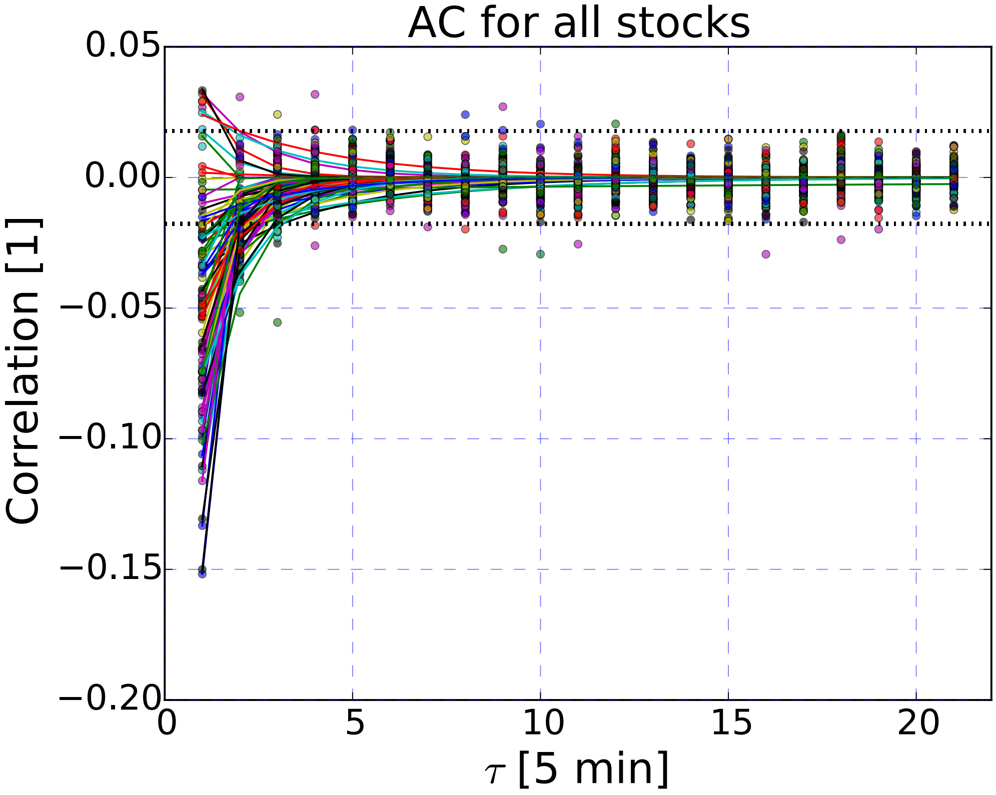

Figure 16: Temporal evolution of the autocorrelations of C^{τ}_{res} and their exponentials fits. The dashed black lines correspond to the noise limit ρ_{max}.

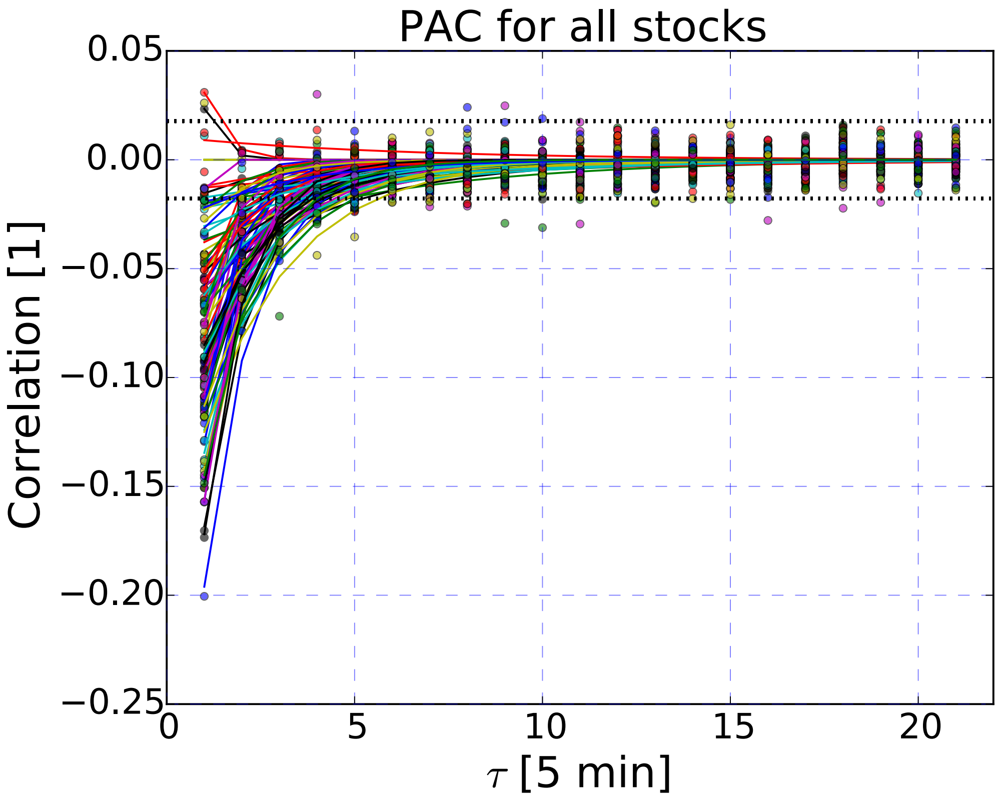

Figure 17: Temporal evolution of the autocorrelations of C_p^τ and their exponentials fits. The dashed black lines correspond to the noise limit ρ_{max}.

Interestingly, the are a few autocorrelations outside the noise region for $\tau = 8, 9, 10, 11, 16, 18, 19$ for both, the autocorrelations of C_{res}^τ and the partial autocorrelations. This could indicate significant autocorrelations not described by a simple exponential decay.

The distributions of the decay time fit parameter are presented in Figure 18. For the autocorrelations of C_{res}^τ the peak of the distribution is below $\tau = 1$ and peaks at approximately $\tau = 0.5 = 2.5\,\text{min}$. The peak for the partial autocorrelation is apprixmately center around $\tau = 1.4 = 7\,\text{min}$.

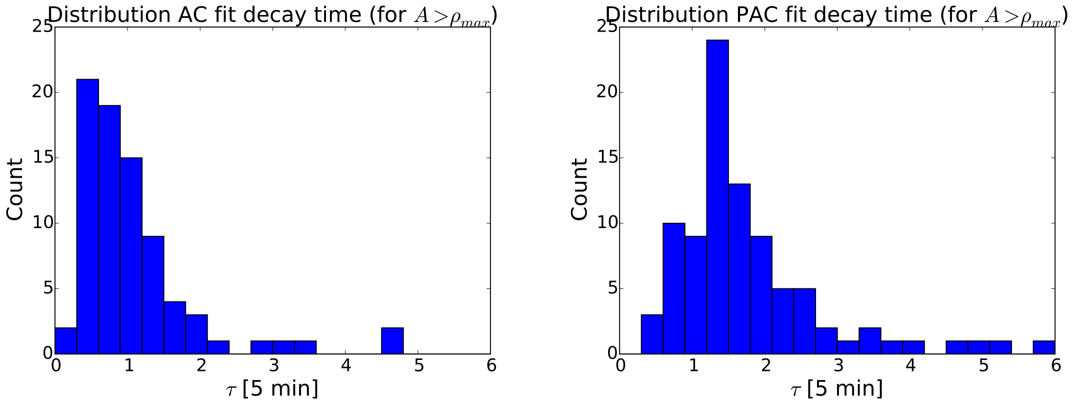

Figure 18: The plot on the left shows the distribution of the decay time constant of the exponential fits to the temporal evolution of the autocorrelations of C^{τ}_{res}. Only time constants are considered for which the associated amplitude fit parameter A lies outside the noise spectrum. The plot on the right side corresponds to the partial autocorrelations.

7.2 Partial cross-correlations

The time-lagged cross-correlations can extend the analysis of cluster identification for synchronous correlations. It gives insight how stocks influence each other in the future. I expect stocks from the same sector will have strong time-lagged correlations to each other and appear as symmetric elements in the time-lagged correlation matrix ($c^{\tau}_{ij} \approx c^{\tau}_{ji}$). Asymmetric entries will represent a directed future influence of stock i to stock j without getting influenced by stock j. Here only the time-lagged partial correlations are considered in order to test their typical time scale. I applied a threshold of $\pm 0.05 \approx 3\rho_{max}$ to the off-diagonal entries of C^1_p to filter the strongest time-lagged partial cross-correlations. Figure 19 shows the resulting matrix after the filtering. Most entries in the lower left triangle below the diagonal have a symmetric partner in the upper right triangle. Also entries in the upper right triangle without a symmetric partner can be seen.

Similarly to the analysis above I study the evolution of the correlations over the lag time. The result is depicted in Figure 20. The temporal dependence is almost the same as for the partial autocorrelations and the decay time distribution peak has the same position at approximately $\tau = 1.4 = 7\,\text{min}$. Therefore, all correlation in C^{τ}_p decay on the same timescale.

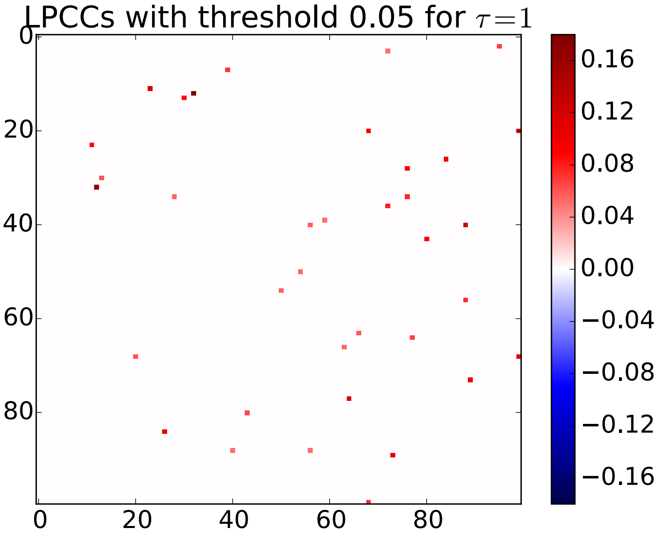

Figure 19: Time-lagged partial correlation matrix without diagonal elements and a filter threshold of 0.05.

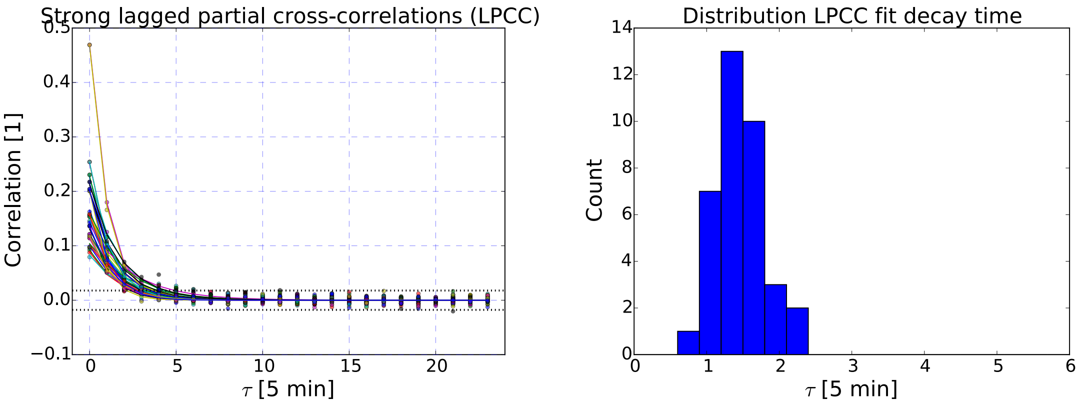

Figure 20: Temporal evolution of the strong time-lagged partial cross-correlations of C_p^τ and their exponential fits in the left plot. The dashed black lines correspond to the noise limit ρ_{max}. The right plot shows the distribution of the fit parameter for the decay time constant of the exponential fit function.

In Figure 21 I present LPCCs which exhibit peaks at lags different than zero or one. They weren't investigated further but in order to measure its significance a more thorough analysis of the extension of the noise region is needed.

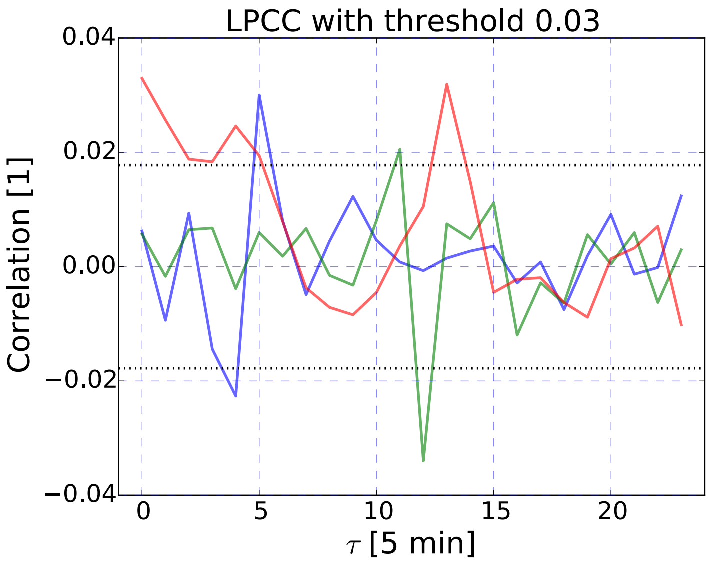

Figure 21: Time-lagged partial cross-correlations with correlation peaks above 0.03 for lags greater than one. The dashed black lines correspond to the noise limit ρ_{max}.

8 Eigenvalue distribution of time-lagged partial correlation matrices

In [7] the eigenvalue decomposition analysis was extended to time-lagged correlation matrices. Following this approach the eigenvalue spectrum is complex, because in general the time-lagged correlation matrices will be asymmetric. As in the case for synchronous correlation matrices one wants to distinguish the relevant eigenvalues from the noise spectrum.

8.1 Random Matrix theory for asymmetric real random matrices

For asymmetric random gaussian matrices the eigenvalue density in the complex plane for $N \to \infty$ has the form of a circle [13]

$$f(\omega = x + iy) = \begin{cases} \frac{1}{\pi a^2} & x^2 + y^2 \leq a^2 \\ 0 & \text{otherwise} \end{cases}$$

A proper scaling with $a = \sqrt{N}\sigma$ gives an estimate for the support of the noise spectrum for eigenvalues of C^{τ}_{scr}. To verify the goodness of this estimation I generated asymmetric random matrices (dimension $N \times N$) with uncorrelated gaussian entries and compared the eigenvalue spectrum to the density given above. As shown in Figure 22 the agreement is good.

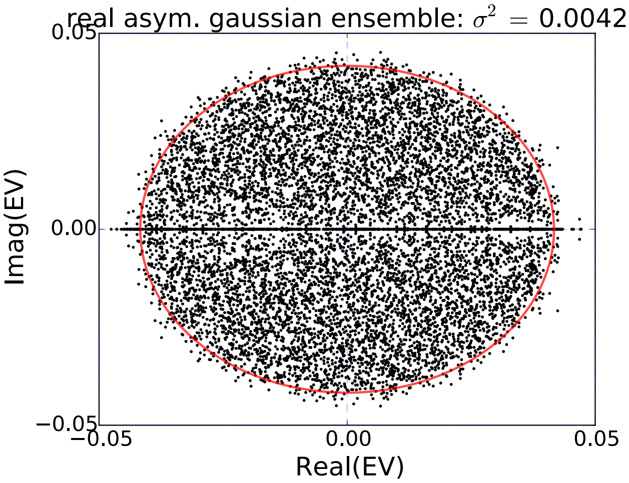

Figure 22: Eigenvalue distribution of 100 asymmetric random gaussian matrices with dimension $N \times N$. The red solid lines shows the theoretical support for $N \to \infty$. The distinct gap of imaginary parts in the vicinity of the real axis and the higher density for real eigenvalues is due to higher level repulsion near the real axis as stated in [13]

A more thorough analysis is given in [7], where the authors extended the results of [14] for symmetric cases to asymmetric real random matrices.

8.2 Eigenvalue distribution comparison

The eigenvalue spectra for different lags($\tau = 1, 3, 14$) are shown in Figures 23, 24 and 25. For the non-scrambled time-lagged correlation matrices the theoretical noise support is shifted by the average value of the diagonal elements. As shown in Section 5.2 the diagonal elements shift the whole spectrum along the real axis. For small lags the spectra of the market mode removed data and the partial lagged correlation matrix show different shapes. C^{τ}_{res} typically has one large real eigenvalue or one pair of complex eigenvalues separated from the bulk spectrum and with a large positive real value. For greater lags the shapes will be similar and, interestingly, there is always a fraction of eigenvalues outside the noise spectrum, e.g. $\tau = 14$.

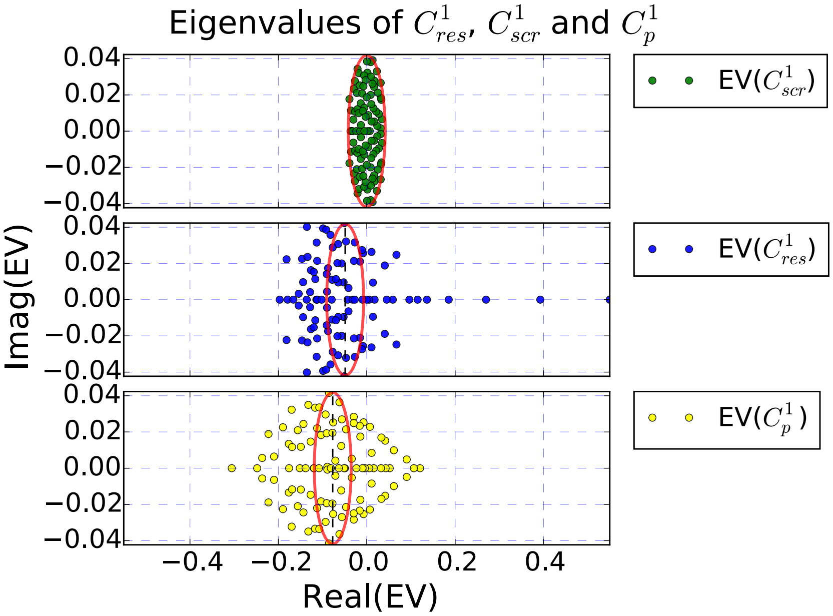

Figure 23: Eigenvalues of C^1_{scr}, C^1_{res} and C^1_p ($\tau = 1$). The red solid line indicates the theoretical support for eigenvalues of an asymmetric random gaussian matrix. The scrambled data shows very good agreement with the theoretical result not only for $\tau = 1$, but for all considered lags.

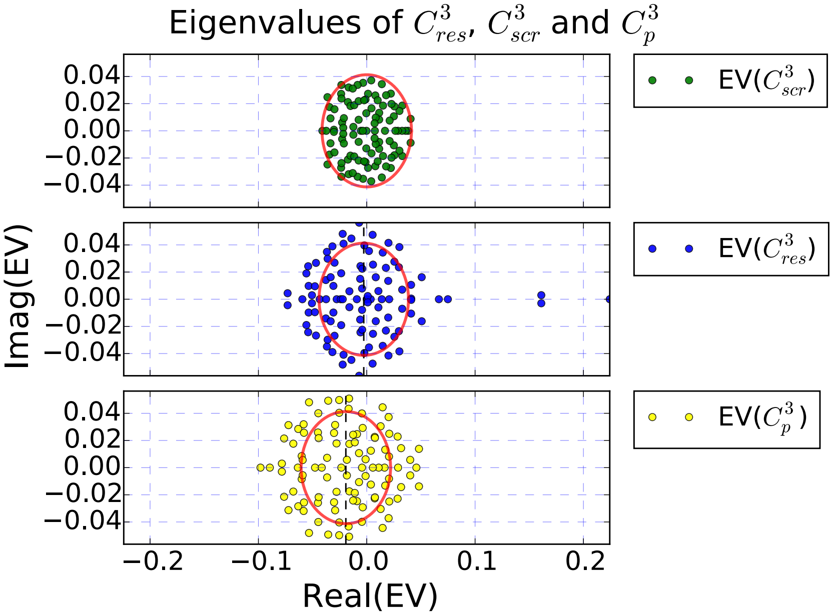

Figure 24: Eigenvalues of C^3_{scr}, C^3_{res} and C^3_p ($\tau = 3$). The red solid line indicates the theoretical support for eigenvalues of an asymmetric random gaussian matrix. The scrambled data shows very agreement with that limit.

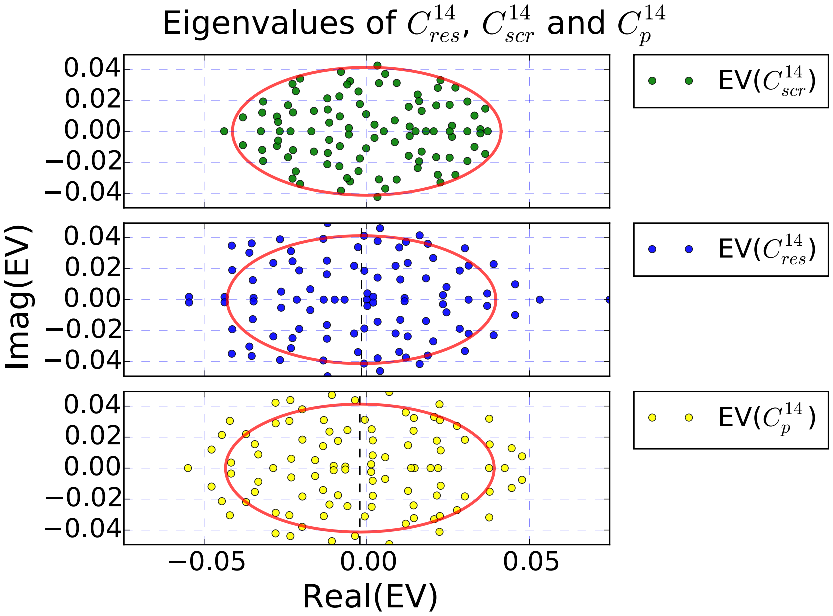

Figure 25: Eigenvalues of C^{14}_{scr}, C^{14}_{res} and C^{14}_{p} ($\tau = 14$). The red solid line indicates the theoretical support for eigenvalues of an asymmetric random gaussian matrix. The scrambled data shows very agreement with that limit.

A further analysis will include an investigation of the associated eigenvectors and the structure of the time-lagged correlation matrices. I believe that the deformation of the eigenvalue spectrum for C^1_p along the real axis is related to the strong symmetric cross-correlations. Generally speaking, the symmetric portion is related to real eigenvalues, whereas the antisymmetric portion of a matrix relates to pure complex pairs of eigenvalues. So far I couldn't find any reference on that and neither do I have a mathematical proof for that. It is purely hypothetical and has to be investigated in the future. In summary, these results are more a proof of concept for identifying siginicant information about the correlations.

9 Conclusion and Outlook

The concept of partial correlation including all other stocks at once has been succesfully applied to the NYSE data from 2001-2003. Surprisingly, the third party correlation did dampen the raw correlation between two stocks. I expected the opposite to be the case. Furthermore, I was able to show that stocks with strong influence on all other stocks within one unit time (five minutes) are also more likely to be influenced by all other stocks. A typical exponential decay time of seven minutes seems to be a property of both, autocorrelations and cross-correlations. In the last part it was shown that even for great lags τ eigenvalues of the time-lagged correlation matrix lie outside of the noise region. It would be interesting to investigate the eigenvalue spectrum further in order to understand the relation between the shape of the matrix, in terms of symmetric and antisymmetric parts, and the associated shape of the eigenvalue spectrum. Future work could include the creation of a model based on the observed partial correlations outside the noise region. Also previous work of cluster identification for synchronous correlations could be repeated for the partial correlations. Also time-lagged partial correlations could be used to compare with work done on partial correlations with only one or two conditions [1, 2]. Finally, I think it would be very useful to do a singular value decomposition to see what are the strongest mutual changes between lags.

Acknowledgements

I want to thank Chester Curme for introducting the topic to me, providing the data, pointing out relevant papers and for all the helpful discussions.

References

[1] D. Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. N. Mantegna, and E. Ben-Jacob, Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market, PloS one **5** (2010) no. 12, e15032.

[2] D. Y. Kenett, X. Huang, I. Vodenska, S. Havlin, and H. E. Stanley, Partial correlation analysis: Applications for financial markets, ArXiv e-prints (Feb., 2014) , `arXiv:1402.1405 [q-fin.ST]`.

[3] Y. Shapira, D. Y. Kenett, and E. Ben-Jacob, The index cohesive effect on stock market correlations, The European Physical Journal B-Condensed Matter and Complex Systems **72** (2009) no. 4, 657–669.

[4] G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali, Partial correlation for functional brain interactivity investigation in functional {MRI}, NeuroImage **32** (2006) no. 1, 228 – 237. `http://www.sciencedirect.com/science/article/pii/S1053811906000103`.

[5] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes, Discovery of meaningful associations in genomic data using partial correlation coefficients, Bioinformatics **20** (2004) no. 18, 3565–3574.

[6] P. Fransson and G. Marrelec, The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis, NeuroImage **42** (2008) no. 3, 1178 – 1184. `http://www.sciencedirect.com/science/article/pii/S1053811908007283`.

[7] C. Biely and S. Thurner, Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series, ArXiv e-prints (Sept., 2006) , `arXiv:physics/0609053 [soc-ph]`.

[8] C. Curme, M. Tumminello, R. N. Mantegna, H. E. Stanley, and D. Y. Kenett, Emergence of statistically validated financial intraday lead-lag relationships, ArXiv e-prints (Jan., 2014) , `arXiv:1401.0462 [q-fin.ST]`.

[9] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley, Time-lag cross-correlations in collective phenomena, EPL (Europhysics Letters) **90** (2010) no. 6, 68001.

[10] K. Baba, R. Shibata, and M. Sibuya, Partial correlation and conditional correlation as measures of conditional independence, Aust. N. Z. J. Stat. **46(4)** (2004) 657–664.

[11] J.-P. Bouchaud, L. Laloux, M. Augusta Miceli, and M. Potters, Large dimension forecasting models and random singular value spectra, ArXiv Physics e-prints (Dec., 2005) , `arXiv:physics/0512090 [data-an]`.

[12] E. P. Wigner, On the distribution of the roots of certain symmetric matrices, Annals of Mathematics (1958) 325–327.

[13] H. Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein, Spectrum of large random asymmetric matrices, Physical review letters **60** (1988) no. 19, 1895–1898.

[14] K. B. K. Mayya and R. E. Amritkar, Analysis of delay correlation matrices, ArXiv e-prints (Jan., 2006) , `arXiv:cond-mat/0601279 [cond_mat]`.