Inferring Graph Structures for Stock Market Data

An Exploration of Graphical Models

Suchit Mehrotra & Nathaniel Corder

North Carolina State University

Introduction

Causal inference is of core importance to the financial markets. If one could determine the causal structure underlying the price movement of financial instruments, theoretically they should be able to exploit any inefficiencies that arise. As evidenced by the recent economic uncertainty and volatility in the financial markets, these relationships are almost impossible to infer, which implies even a rudimentary understanding of how prices interact can be of great value to investors.

To understand the interconnectedness of the financial markets on a particular day, we use graphical models to display the underlying structure of price movements. We use 10 index funds which represent the economy as a whole, and seek to determine undirected and causal relationships between them. Furthermore, we investigate whether additional explanatory information is present in the lagged returns of the 10 index funds selected.

Graphical Models

A graph, G, is defined as $G = (V, E)$ where V denotes the number of vertices, or nodes, in the graph, and E the number of connections, called edges, between the nodes. These edges can be directed, undirected, or bi-directed. The number of vertices is the number of random variables in the data set, and the number of edges is determined via the application of statistical techniques. A few directed relationships displayed by a graph are:

(i) **Chain:** $X \to Y \to Z$

(ii) **Chain:** $X \leftarrow Y \leftarrow Z$

(iii) **Fork:** $X \leftarrow Y \to Z$

(iv) **Collider:** $X \to Y \leftarrow Z$

Cases (i), (ii) and (iii), imply $X \perp\!\!\!\perp Z|Y$, whereas case (iv) implies $X \perp\!\!\!\perp Z$ without any conditioning. Using these relationships, the joint distribution of a set of variables can be inferred from a graphical model.

Data

Using the R package *quantmod* we downloaded data for 10 index funds (Table 1), representing the underlying US economy from January 2007 to April 2015. We then calculated the log-returns (defined below) for each fund, and used these to conduct our inference.

$$log(returns) = log\left(\frac{price_t}{price_{t-1}}\right)$$

Table 1: List of Index Funds Used

	Ticker	Industry
1	SPY	S&P
2	XLB	Materials
3	XLE	Energy
4	XLF	Financials
5	XLP	Consumer Staples
6	XLI	Industrials
7	XLU	Utilities
8	XLV	Healthcare
9	XLK	Technology
10	XLY	Consumer Discretionary

Methods & Results

Let $G = (V, E)$ be a graph. In our specific situation, we have data for 10 funds; therefore, we have a graph with 10 nodes: $G = (10, E)$. The number of edges in our graphs will be determined from the methods that are used.

Method 1: Gibbs Sampling with Log-Returns

We used multiple methods to derive the graphical models presented in our results section. Our first goal was to create an undirected graph based on the raw log-returns. The procedure is described below:

- Assume $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Using Gibbs sampling, derive Σ and use this to determine the partial correlation matrix $\boldsymbol{P}$ at each iteration.

$$\boldsymbol{P}_{ij} = \left[\frac{\Sigma^{-1}_{ij}}{\sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}}\right]^{I(i \neq j)}$$

- Create a 95% credible set for each set of partial correlations and if zero is not in the set, create an edge between the two variables.
  - Because this requires 55 tests, we use a Bonferroni correction to adjust for multiple comparisons.

The results of this method are shown in Figure 1:

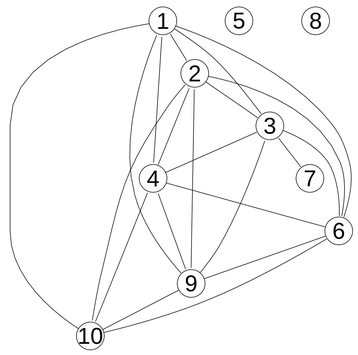

Figure 1: Undirected Graph For Log-Returns with Bonferroni Correction

Method 2: Using the PC Algorithm with Log-Returns

The graph in Figure 1 shows undirected relationships between the the funds in our data. However, we are also interested in causal relationships which can be summarized via a Directed Acyclic Graph (DAG). To do this we use the PC algorithm, the steps for which are outlined below (Shalizi, 2013):

- Start with a full (saturated) graph.
- For each pair of variables A and B, check if $A \perp\!\!\!\perp B$. If they are, remove edge connecting A and B.
- For each pair of variables A and B still connected, check if there exists a variable C, such that $A \perp\!\!\!\perp B|C$. If so, remove edge {A,B}.
- For each pair of variables A and B which remain connected, if there exists a set of variables $\boldsymbol{D} = (C_1, \ldots, C_k)$ such that $A \perp\!\!\!\perp B|\boldsymbol{D}$, remove edge {A,B}.
- Continue this procedure until $k = p - 2$, where p is the number of total variables.

It can be seen that the PC algorithm is a backward selection procedure from a full graph. The results from using the PC algorithm with log-returns is shown in Figure 2

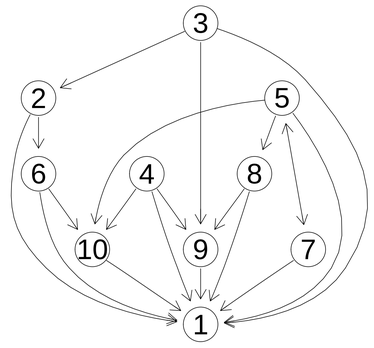

Figure 2: Inferred Causal Graph for Log-Returns

Method 3: Conditioning on Past Returns

Figure 1 and 2 show graphs for raw log-returns of the index funds of interest. However, we are also interested in determining relationships between funds after conditioning on past returns. That is, given yesterdays returns what additional information can be obtained for returns today. To do this we regressed each fund's returns on the previous day's returns of all other funds, and the previous five day's returns for the fund itself. A mathematical representation of this is given below:

$$Y_{it} = \beta_0 + \boldsymbol{\beta}\boldsymbol{Y}_{t-1} + \gamma_1 Y_{i,t-2} + \cdots + \gamma_4 Y_{i,t-5} + e_{it} \qquad (1)$$

We then took the residuals generated by these models, and estimated their covariance matrix using Gibbs sampling. The information contained in the residuals was information not explained by our regression, and their covariance matrix on a particular day, t, is representative of the relationships of stock movements on day t conditioned on past returns.

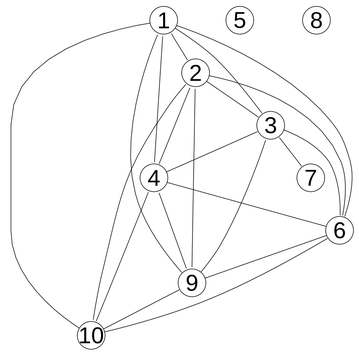

Figure 3: Undirected Graph For Residuals with Bonferroni Correction

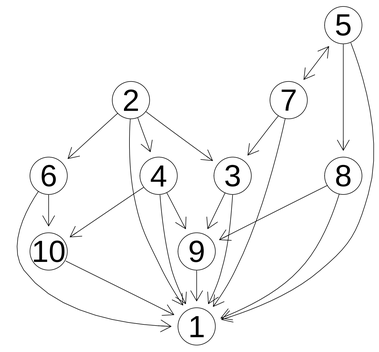

Figure 4: Inferred Causal Graph for Residuals

Figures 3 and 4 show graphical models generated by using Gibbs Sampling and the PC algorithm on the residuals of the regression in Equation 1, respectively. Figure 1 and 3 are exactly the same, implying that the past stock returns provide no additional information regarding the relationships of today's returns. However, using the PC algorithm produces a different graph and changes some of the causal relationships inferred in Figure 2.

Conclusions

- From Figure 1 and 3 it seems that the consumer staples and healthcare index funds are independent of the movement of the overall market.
- The DAG's in Figure 2 and 4 correctly capture the directional relationship of the impact of index funds on the S&P 500.
- The DAG's show that conditioned on the returns of the S&P 500, none of the returns for the index funds are independent.

Caveats

- The Bonferroni correction for the undirected graphs is too conservative. Decreasing the width of the credible set for the partial correlations leads to much more complicated models.
- An assumption of the PC algorithm is that all relevant variables are included in the Graph. This is not necessarily the case, as unmeasured variables such as economic growth and monetary policy are not included.
- The normality assumption is not necessarily met. Stock returns tend to have more extreme returns than the normal distribution allows.

References

[1] Peter Hoff. *A first course in Bayesian statistical methods*. Springer, New York London, 2009.

[2] Sren Hjsgaard. *Graphical models with R*. Springer, New York, 2012.

[3] Radhakrishnan Nagarajan. *Bayesian networks in R with applications in systems biology*. Springer, New York, NY, 2013.

[4] Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Pittsburgh, PA, 2013.