Learning Stock Networks with Robust PCA

José Vinícius de Miranda Cardoso, Yixin Men, Shunkang Zhang

CSIC5011 – Topological and Geometric Data Reduction and Visualization

Objectives

Our goals in this project is to

- illustrate the potential of robust principal component analysis in revealing correlations between stock prices time-series data.

Introduction

Principal Component Analysis (PCA) is a fundamental technique to understand high-dimensional big datasets. In short, PCA computes a low-dimensional subspace onto which a data matrix can be optimally projected (in an ℓ_2-norm sense).

However, in practice, the subspace computed by PCA is often sensitive to outliers or to non-Gaussianity in the distribution of the data matrix, which is often the case in financial time-series data. To overcome these shortcomings, Candes [1] proposed a framework called Robust PCA. In this project we present a comparative analysis between PCA and Robust PCA when applied to estimate financial networks from time-series stock data. We consider a financial network as an undirected, weighted graph denoted as a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$, where $\mathcal{V} = \{1, 2, \ldots, p\}$ is the vertex (or node) set, $\mathcal{E}$ is the edge set that is a subset of the set of all possible unordered pairs of p nodes such that $(i, j) \in \mathcal{E}$ iff nodes i and j are connected. We denote the number of elements in $\mathcal{E}$ by $|\mathcal{E}|$. $\boldsymbol{W} \in \mathbb{R}_{+}^{p \times p}$ is the symmetric weighted adjacency matrix that satisfies $W_{ii} = 0$, $W_{ij} > 0$ iff $(i, j) \in \mathcal{E}$, $W_{ij} = 0$ otherwise.

Data Explanation

We use time-series price data from stocks that belong to the S&P500 Index. Denote $p_i(t)$ the price of the i-th stock at the t-th day. Stock price data are known to be non-stationary, approximately following a Brownian motion process. To remove the non-stationarity, we first compute the log-returns of each time-series, which are defined as

$$\log(1 + r_i(t)) \triangleq \log(p_i(t)) - \log(p_i(t-1)), \quad (1)$$

The quantity $\log(1 + r_i(t))$ is ultimately what we use in the tasks that will follow.

PCA and Robust PCA

Classical Principal Component Analysis (PCA) seeks the best rank-k estimate (in an ℓ_2-norm sense) of a matrix M by solving

$$\begin{aligned} &\underset{L}{\text{minimize}} \quad \|M - L\|_2, \\ &\text{subject to } \operatorname{rank}(L) \leq k. \end{aligned} \quad (2)$$

Due to the squared error used in PCA, it cannot provide meaningful results when the data matrix M contains outliers or when its features are not Gaussian distributed.

To overcome those limitations, Robust PCA (RPCA) has been proposed by [1]. The assumption underlying RPCA is that the data matrix M can be decomposed as $M = L + S$, where L and S are the low-rank and sparse components, respectively. RPCA can be formulated as the following convex optimization problem

$$\begin{aligned} &\underset{L,S}{\text{minimize}} \quad \|L\|_* + \lambda \|S\|_1, \\ &\text{subject to } L + S = M, \end{aligned} \quad (3)$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ are the nuclear and ℓ_1 norms used to impose low-rankness and sparsity, respectively.

Results

Figure 1 illustrates 5-year worth of price data of a subset of stocks from the S&P500 index.

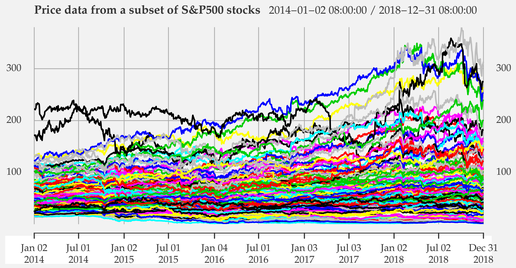

Figure: Stock prices illustration.

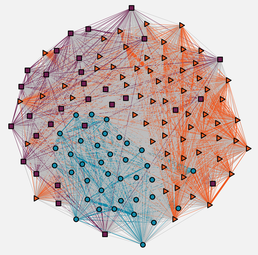

Figure: Graphical network obtained naively from the sample correlation matrix.

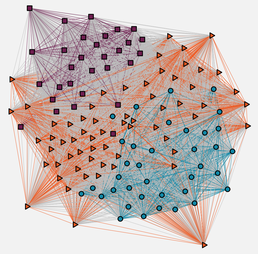

Figure: Graphical network obtained with PCA.

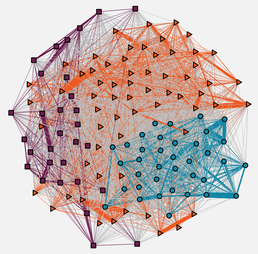

Figure: Graphical network obtained with Robust PCA.

Discussion

To estimate financial network graphs, we select 130 stocks from three sectors (Industrials, Consumer Staples, and Energy), from the period of 2014 to 2018. Stocks within sectors should show a more correlated behaviour. Figure 2 shows a naive graph estimated directly from the sample correlation matrix of the stocks time-series. From Figure 2, it is difficult to distinguish which stocks belong to each sector. Each color represent one sector. Figure 3 shows the graph estimated using classical PCA where we set the rank of the matrix to be equals 3 (i.e., the number of known sectors). Figure 4 shows the same graph, but estimated with Robust PCA. As we can notice, Robust PCA presents an improvement on the estimation of the financial networks, as revealed by a clear graph.

Conclusion

In this project, we use both PCA and robust PCA to estimate the correlation among stocks data. As for robust PCA, we add sparse constraint to reduce the noise information involved in the estimated covariance matrix. Based on the shown graph, it is easy to see the improvement by using robust PCA for the noisy stock data. What's more, estimating financial networks still remains a challenge and an active research topic [2].

References

1. E. J. Candes, X. Li, Y. Ma, J. Wright. Robust Principal Component Analysis?. Pre-print arXiv:0912.3599, 2009.
2. G. Marti, F. Nielsen, M. Binkowski, P. Donnat. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Pre-print arXiv: 1703.00485, 2019.