Dimension Reduction and Visualization of Financial Data

Supun Kamburugamuve, Pulasthi Wickramasinghe, Saliya Ekanayake, Milinda Pathirage, Maya Smith, Anthony Scott and Geoffrey Fox

School Of Informatics and Computing, Indiana University

Introduction

In this poster we present the results of applying Multidimensional Scaling (MDS) to visualize stock data in 3 Dimensional space. Years of historical daily stock data is segmented using a sliding window approach to create a continuous visualization of the stock data through time. The distance between the stocks for a time period is calculated primarily using the Pearson Correlation. The continuous stock data in 3D is visualized through a large scale web based visualization tool developed.

Data

Original Data

- Obtained daily stock values using The Center for Research in Security Prices (CSRP)[1] database through the Wharton Research Data Services (WRDS) web interface
- 2004 to 2015 Daily Stock prices in the form of a CSV file
- We use the information
  - ID, Date, Symbol, Factor to Adjust Volume, Factor to Adjust Price, Price, Outstanding Stocks

Data Cleaning

- Negative price values : There were 517208 records with negative values out of 18921341 which is 2.73% of total values.
- Missing Values: Drop stock with more than 5% missing values
  - There are 1138 total stocks with more that 5% of missing values per year from 2004 to 2015
  - In average 103 stocks were dropped per year window.
- Split data: 2456 stock splits in the 2004 to 2015 period

Data Segmentation

- Sliding Window data segmentation
- 1 Year window with different shifts
  - When the shift is large, cannot match subsequent 3D projections
  - We use 1 Year window with 7 Day shifts
- From 2004 to 2015 with 1 Year and 7 Days shifts; there are 523 data segments with about 6500 stocks per segment.

Distance Calculation

- Distances d_{ij} between two stocks is calculated using Pearson Correlation with different choices for "length" l_i of each stock
- Change of stocks into account by using the concept of length.
- A constant stock was added as a fiducially point of reference

$$d_{ij} = \sqrt{l_i + l_j - 2 \times l_i \times l_j \times corr(s_i, s_j)}$$

$$l = 1 \; or \; l = 10 \times |\log_2(change \; of \; stock)|$$

Weight Calculation

- Market capitalization of a stock is taken as the weight

Data Processing Workflow

The data is cleaned and converted to the format that is accepted by the MDS algorithm. Then MDS algorithm is applied to produce the 3D points. Each MDS Projection is ambiguous to rotations, reflections and translation. This was addressed by a least squared fit to find the best transformation between all consecutive pairs of projections.

1. Source: CRSP, Center for Research in Security Prices. Graduate School of Business, The University of Chicago 2015. Used with permission. All rights reserved. www.crsp.uchicago.edu

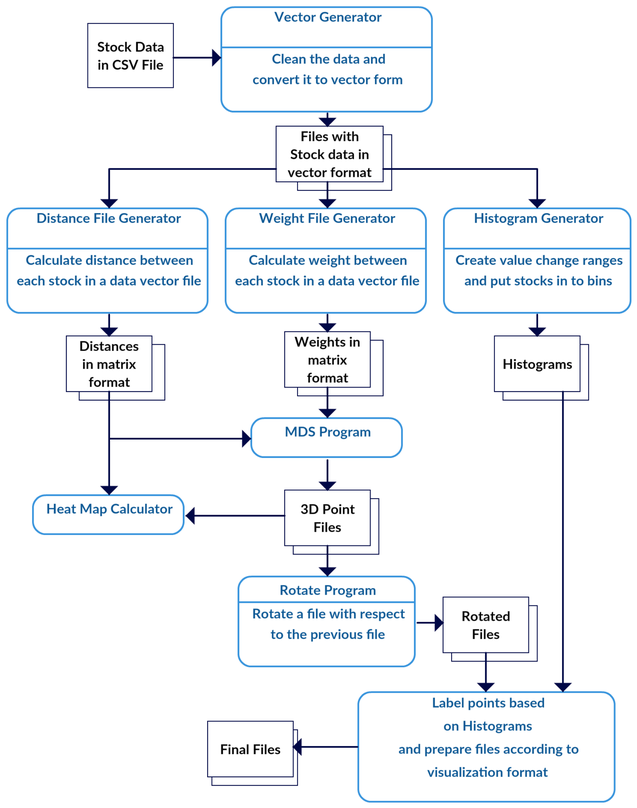

Fig 1. Data analysis pipeline

Because consequent segments of stock data are different we don't expect a perfect matching between the consequent plots.

We applied labels to the 3D points based on the classification of stocks to sectors as well as by using a histogram to put stocks in to different categories using the change of stock prices over a given period. Fig. 2 Shows distribution of distances for a sample data segment and how well the original distances are matched with the 3D projected distances. Fig. 3 Shows a histogram of the errors (chisq values) of using least squared fits to match the consecutive MDS projections.

The data analysis workflow is implemented using MPI and Java and runs in a 32 node cluster with 24 CPU cores and 48 GB of memory in each node. Since we are using MPI we used a file based approach for data management.

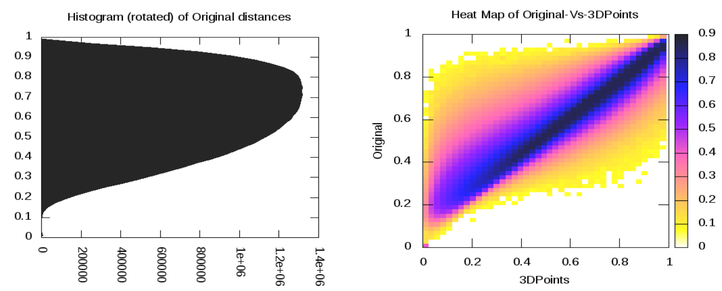

Fig 2. Distance Histogram and Heat map of a MDS run

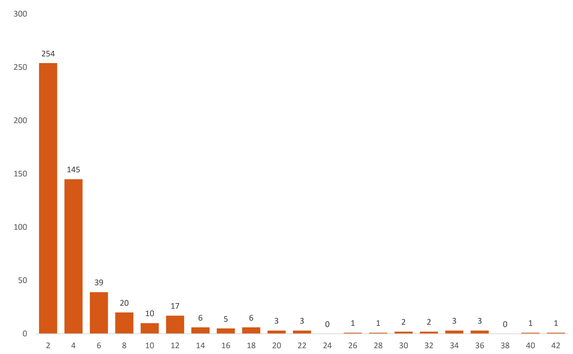

Fig 3. Histogram of errors in matching consecutive 3D projections

WebPlotViz for Data Visualization

- A Web Browser based 3D Visualization tool developed at Digital Science Center of Indiana University
- Uses three.js for rendering 3D plots in the browser
- Can be used to visualize sequence of time series 3D data
- Data stored in NoSQL databases and allows scalable plot data processing in the back end
- http://spidal-gw.dsc.soic.indiana.edu/

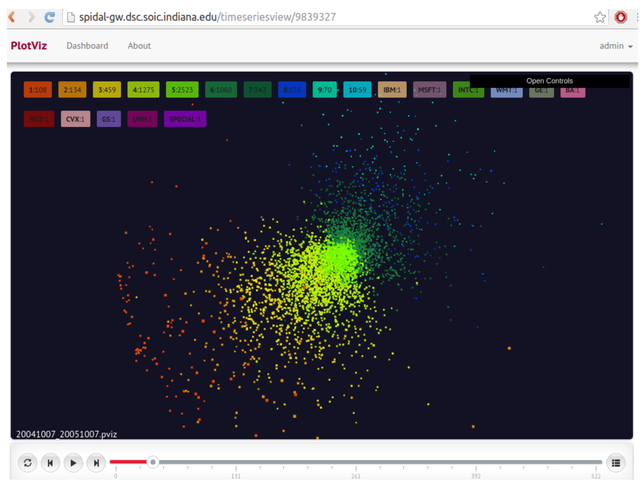

Fig 4. PlotViz data visualization

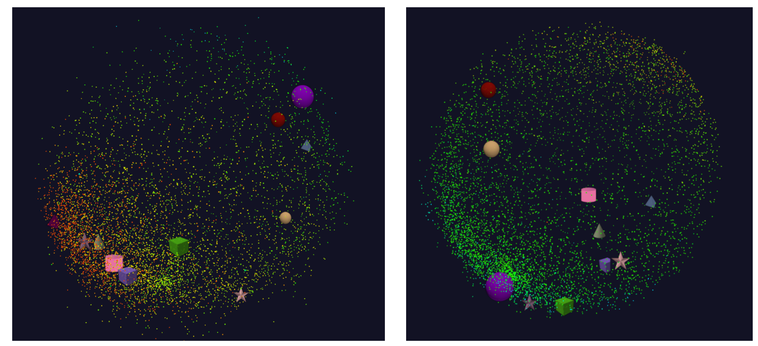

Fig 5. Stock data in 3D

Conclusions & Future Work

We showcased the results of applying Multidimensional Scaling to financial data to visualize the changes of stocks over time. Our methods are scalable both in number of stocks and number of data segments considered. We are extending our approach in to different distance metrics as well as data segmentations to get more deeper understanding about the financial data. We are planning to extend the system using big data technologies to efficiently manage large amounts of data required at various steps of the workflow.

Acknowledgement

This work is funded by NSF 1443054 SPIDAL grant and we would like to extend our gratitude to FutureSystems team for their support with the computational infrastructure.

References

1. Ruan, Yang. "Scalable and robust clustering and visualization for large-scale bioinformatics data." PhD diss., Indiana University, 2014.