Visualization and Forecasting on Finance Data

Yifan Hao* Yueying Hu† Yakun Li‡ Yonglin Liu§
Department of Mathematics,
The Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong.

Abstract

In this report, we conduct analyses on a collection of stock market index of 500 companies. First, we use Principal Component Analysis (PCA), Sparse PCA (SPCA), Multi-Dimension Scaling (MDS), Mapping (ISOMAP), and t-distributed Stochastic Neighbour Embedding (t-SNE) for data reduction. Data with 1000+ features can be visualized on 2D surfacce. Next, we try to do forecasting. Several prediction methods, namely, PCA, SPCA and target-PCA (t-PCA), are used and the forecasting results are explained in detail.

1 Introduction

The Standard and Poor's 500, or simply the SNP'500, is a stock market index tracking the stock performance of 500 of the largest companies listed on stock exchanges in the United States. It is one of the most commonly followed equity indices.

In order to interpret such dataset, many methods can be used to drastically reduce its dimensionality in an efficient way, such that the redundant noise can be removed while most of the information can be preserved. In this project, Principal Component Analysis (PCA), Sparse PCA (SPCA) and Multi-Dimension Scaling (MDS), which belong to linear dimensionality reduction methods, are applied to reduce the dimension of SNP'500 data. Manifold learning methods like Isometric Mapping (ISOMAP), and t-distributed Stochastic Neighbour Embedding (t-SNE) are also utilized. Because data in two or three dimensions can be plotted to show its inherent structure, we choose to visualize the high dimensional SNP'500 data on two-dimensional surface.

Finally, based on the performance during data reduction, PCA, SPCA together

*yhaoah@connect.ust.hk
†yhucn@connect.ust.hk
‡ylinv@connect.ust.hk
§yliuks@connect.ust.hk

with t-PCA are used to predict 9 different types of stocks. We combine different data reduction with different machined learning methods, evaluate the performance of forecasting model and explore the reason why linear model has higher accuracy.

2 Dataset

2.1 Data Description

The finance dataset SNP'500 contains 452×1258 matrix, which represents the closed price of stocks from 452 American company in 1258 consecutive workdays. In another word, each row represents the stock price change of one company, each column represents one workday, and the entry of the matrix $\mathbf{X}_{ij}$ represents the stock price of company (i) is in a certain workday (j).
Grouping the finance data by the class information of stock, we can divide the dataset into 10 different classes, including Industrials, Financials, Health Care, Consumer Discretionary, Information Technology, Utilities, Materials, Consumer Staples, Telecommunications Services, and Energy. The histogram of detailed distribution of stock classses can then be obtain, shown in Figure 1. We can see that most of the companies are from Financials, Consumer Discretionary, and Information Technology, while the group of Utilities is comparatively negligible. This may be due to the economic mechanism of capitalism of USA.

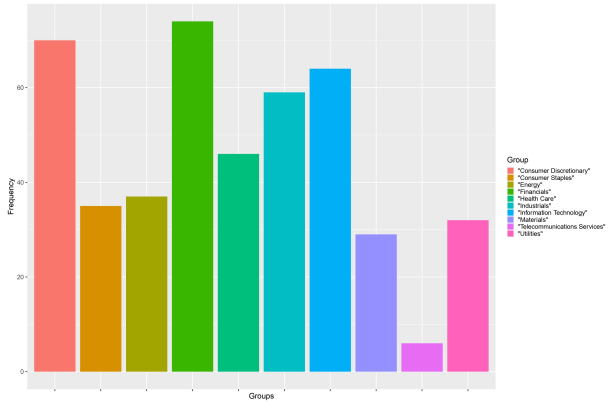


Figure 1: The detailed distribution of different stock classes

If we take the average value of each group, the time series of the average stock price from 10 classes are shown in Figure 2. We can find that stock of different industry has its unique growth curve, and both the staring and the ending prices are not the same. These stocks reflect the economic structure of the United States in some aspect, and can be used to forecast the direction of the economy.

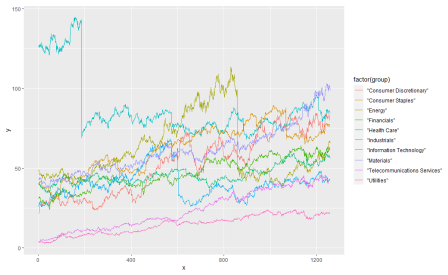


Figure 2: Time series of the stock price

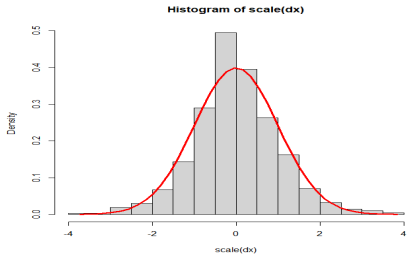


Figure 3: Normalization Result

2.2 Data Preprocessing

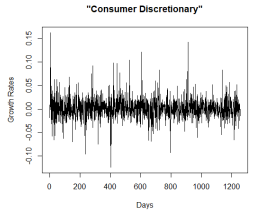
As for the data pre-pocessing, here we calculate the increase rate of stock price by the following formula:

$$Rate = \frac{P_{t+1} - P_t}{P_t}$$

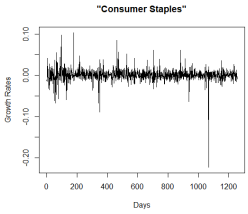
where P_t represents the closed price of previous day and P_{t+1} represents the closed price of current day.
Firstly, we standardize the increase rate by setting the mean value of each feature to zero and the variance of each feature to one. We use this normalized dataset in the following data reduction and classification. Figure 3 shows the increase rate distribution of one stock after data normalization, and we can find that it fits the normal distribution very well.
In addition, if we follow the classification of Figure 1 and ignore the negligible group of Utilities, Figure 4 shows the time series of increase rate from 10 different classes. As we can see from the figure, the moments of extreme increasing rates of stock prices in different fields are distinct, and the peaks also vary. These changes may be caused by industry differences and some significant events of the corresponding field at certain days.

3 Data Reduction and Visualization

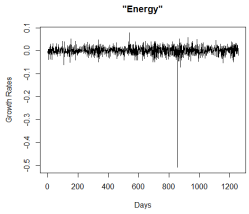
Visualizing high-dimensional datasets can be daunting, as they lack the intuitive structure of two- or three-dimensional datasets. To better visualize such data, it's necessary to reduce the dimensions. Linear dimensionality reduction methods, such as PCA, MDS, can make it to some extent. But linear frameworks assume the data lies in a low dimensional linear subspace, which often misses the non-linear structure. This is where manifold learning comes into play, as it can generalize linear frameworks to nonlinear ones. In this section, we'll use SPCA, MDS, ISOMAP, and t-SNE to reduce the dimensions of our data and visualize the results in Figure 5, where different classes are color-coded. We'll briefly discuss these methods and showcase their corresponding visualization outcomes.



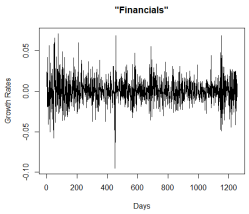
(a) Consumer Discretionary



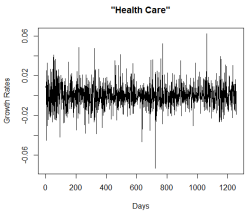
(b) Consumer Staples



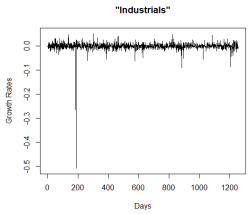
(c) Energy



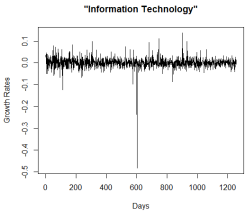
(d) Financials



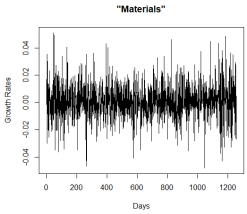
(e) Health Care



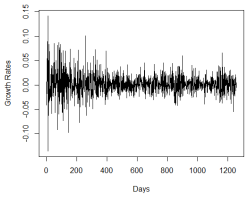
(f) Industrials



(g) Information Technology



(h) Materials

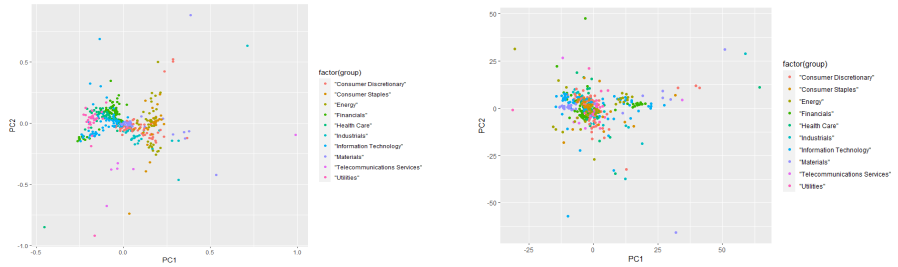


(i) Telecommunications

Figure 4: Time series of the increasing rate of different classes

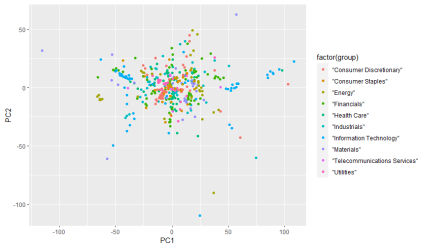
3.1 PCA

Principal component analysis (PCA) [1] is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables, principal components, that successively maximize variance, which changes the original problem into solving an eigenvalue/eigenvector problem.
With the linear framework, PCA has a quite good performance on data reduction and visualization. Thus, it is chosen to do forecasting and will be explained in detail in the next section. And the reduction results shall be omitted here for brevity.

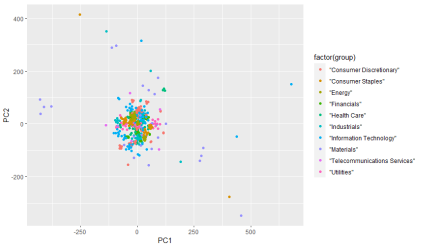


(a) SPCA

(b) MDS



(c) ISOMAP



(d) t-SNE

Figure 5: Project data on four different methods.

3.2 SPCA

A particular disadvantage of ordinary PCA is that the principal components are usually linear combinations of all input variables. Sparse principal component analysis (sparse PCA) [2] extends PCA by introducing the LASSO method (adopting an l_1 penalized regression approach) to generate modified principal components with sparse loadings, which could improve the interpretability, enhance model's predictive power and reduce operational costs.

Figure 5 (a) shows the visualization result by SPCA. From the figure, we notice that some stocks from the same class group together and the cluster separate well with other classes. Like, And the pink cluster (Utilities) and purple cluster (Materials) can be differentiated in the cloud of points. It indicates that SPCA has comparable performance.

3.3 MDS

Multidimensional scaling (MDS) [3] is a means of visualizing the level of similarity of individual cases of a dataset. It projects the high dimensional points into a low subspace by using the pairwise similarity of samples, where the Euclidean distance of each pair of samples is as consistent as possible while in the original high-dimensional space.

Figure 5 (b) shows the visualization result by MDS. The dots of different colors are highly overlapping and it is difficult to distinguish which class they belong

to. This is possibly because the data does not lie well in the manifold defined by MDS.

3.4 ISOMAP

Isometric Mapping (ISOMAP) [4] is an extension of the classical MDS method by changing the Euclidean distance into geodesic distance, which is used for nonlinear dimensionality reduction. It seeks a lower-dimensional embedding which maintains geodesic distances between all points.
Figure 5 (c) shows the visualization result by ISOMAP. Compared to the SPCA result from Figure 5 (a), the ISOMAP seems to be sparse. However, some dots from the same class form the cluster and can be differentiated well from other clusters, especially the the pink cluster (Utilities).

3.5 t-SNE

T-distributed stochastic neighbor embedding (t-SNE) [5] is one of a family of stochastic neighbor embedding methods. This algorithm computes the probability that pairs of datapoints in the high-dimensional space are related, and then chooses low-dimensional embeddings which produce a similar t-distribution. It increases the distance between the clusters with large distances and solves the crowding problem.
Now, we focus on Figure 5 (d), the visualization result by t-SNE. Compared to the SPCA result from Figure 5 (a), different classes are highly overlapping and hard to be identified. This is mainly because many information was lost when the data with 1000+ features is reduced to 2-dimensions.

4 Forecasting

Here we consider the following latent factor model on the N covariates and the target y_{t+h}, which can be solved by PCA methods:

$$X_{it} = \boldsymbol{\lambda}_i^\top \boldsymbol{f}_t + e_{it}, \tag{1}$$

$$y_{t+h} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{f}_t + \boldsymbol{b}^\top \boldsymbol{w}_t + \epsilon_{t+h}. \tag{2}$$

For $t = 1, 2, \cdots, T$ and $i = 1, 2, \cdots, N$, $\boldsymbol{f}_t$ is an r-dimensional vector of the latent factors, $\boldsymbol{\lambda}_i$ is the factor loading corresponding to the i-th covariate, and $\boldsymbol{w}_t$ is an l-dimensional vector of lagged terms. Write $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ and $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_T)$. Denote by $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)^\top$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)^\top$ the factor matrix and the factor loading matrix, respectively.

And in some cases, we do not need to take all covariates into account while estimating factor $\boldsymbol{f}$. Due to this, it is more efficient to use data reduction methods to select a vital subgroup in high-dimensional $\boldsymbol{X}_t$, which is also named

Table 1: MSE of Different Models

Class \ Model	Baseline	PCA	SPCA	t-SPCA
Industrials	0.1359‰	0.1354‰	0.1374‰	0.1352‰
Financial	0.1903‰	0.1889‰	0.1887‰	0.1886‰
Health Care	0.1309‰	0.1304‰	0.1298‰	0.1295‰
Consumer Discretionary	0.3924‰	0.3926‰	0.3947‰	0.3914‰
Information Technology	0.3323‰	0.3312‰	0.3327‰	0.3310‰
Utilities	0.2858‰	0.2873‰	0.2889‰	0.2865‰
Materials	0.1857‰	0.1874‰	0.1925‰	0.1874‰
Consumer Staples	0.1858‰	0.2023‰	0.2035‰	0.2022‰
Energy	0.8682‰	0.8678‰	0.9390‰	0.8686‰

Class \ Model	Baseline	t-MDS	t-ISOMAP	t-t-SNE
Industrials	0.1359‰	0.1350‰	0.1350‰	0.1358‰
Financial	0.1903‰	0.1885‰	0.1882‰	0.1883‰
Health Care	0.1309‰	0.1297‰	0.1296‰	0.1296‰
Consumer Discretionary	0.3924‰	0.3925‰	0.3921‰	0.3906‰
Information Technology	0.3323‰	0.3303‰	0.3309‰	0.3302‰
Utilities	0.2858‰	0.2867‰	0.2869‰	0.2865‰
Materials	0.1857‰	0.1870‰	0.1876‰	0.1876‰
Consumer Staples	0.1858‰	0.2022‰	0.2022‰	0.2022‰
Energy	0.8682‰	0.8681‰	0.8680‰	0.8680‰

as t-PCA. Here we use SPCA, MDS, ISOMAP and t-SNE to take this procedure. And these four models are called t-SPCA, t-MDS, t-ISOMAP and t-t-SNE, respectively. Besides, we also explore the forecasting performance of general PCA [6] and scaled-PCA [7].

Due to the limited sample size of Telecommunications Services (TS), which is only 6, we remove this class from normalized dataset. Thus, we only predict 9 different types of based on the reduction dataset. The dataset is split into training set (60%) and test set (40%). The mean square error (MSE) of the six models together with the baseline are summarized in Table 1. It shows that PCA and SPCA method can generally improve the performance, while t-PCA of four different models maintain similar accuracy with less data dimension. However, in some cases, baseline presents the best prediction, which is mainly because the linear forecasting is not suitable with nonlinear features generated by manifold learning. Besides, the loss of information during data reduction is another reason.

5 Conclusion

For data reduction and visualization, PCA, SPCA, MDA, MDA, ISOMAP and t-SNE methods are carried out to reduce the features of data, project the data distribution on low dimension embedding. Much information is lost when data with 1000+ features is visualized on two-dimensional surface. And here PCA and SPCA can distinguish different classes much more clearly.
For data forecasting, other than PCA and SPCA, we introduce another method t-PCA, since it is more efficient to use data reduction methods to select a vital subgroup in high-dimensional for forecasting. The manifold learning results consist well with our analysis.

6 Individual Contribution

- Report and analysis: Yueying Hu; Yakun Li.
- Code and model setting: Yifan Hao; Yonglin Liu.

References

[1] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[2] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[3] AL Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39, 1992.

[4] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[6] Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[7] Dashan Huang, Fuwei Jiang, Kunpeng Li, Guoshi Tong, and Guofu Zhou. Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695, 2022.