Understanding the Dimension and Structure of S&P 500 Stocks

Students: Thi Van Nguyen, Nicholas DeMarco, Xiyong Yan

Instructor : Professor Xingye Qiao

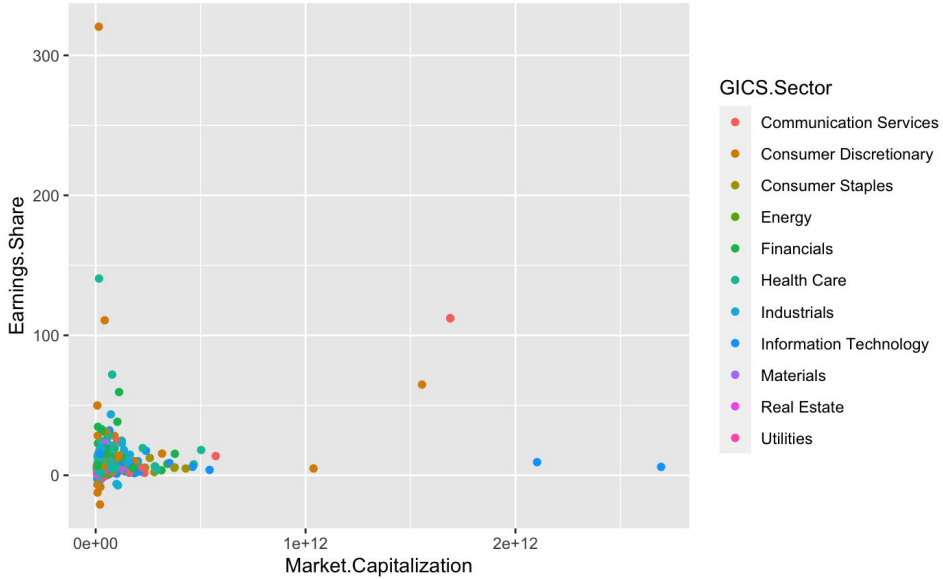
Table of Contents

Introduction

- Stocks are known to be extremely volatile which can play a role in determining whether an investor should purchase shares.
- In addition to this, most investment strategies priotierize grouping stocks based upon their similarities in order to provide a strategic method for future shareholders.
- The main goal of our project is to analyze the variance-covariance and grouping structure of stocks within the select group of stocks within the S&P 500 to provide investors with sufficient information for those interested in purchasing shares.
- We shall also provide an analysis of the relationship among the groups of variables within the dataset.

Understanding the Dataset

- The dataset used for this study is the yahooQF() S&P 500 stock open sourced in R.
- The original dataset offers over 50 variables associated with one given company.
- Research into the measures which apply to a company's earnings per share have allowed us to select 9 additional variables in addition which are known to have some effect (GICS Sector, Market Capitalization, 52 Week Change in Low Price, 52 Week High Price Change, Book Price, PE Ratio, Book Value, EPS Net Year Estimate, Dividend Yield).
- For purpose of visualization, we work with the four largest sectors (Financials, Health Care, Industrials, and Information Technology).



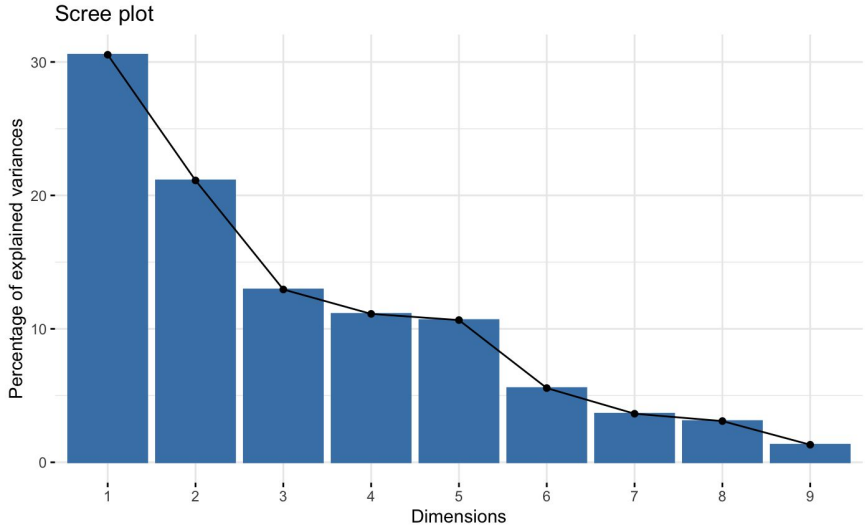
Methodology

- Principal Component Analysis (PCA)
- K-means
- Gaussian Mixture Model (GMM)
- Canonical Correlation Analysis (CCA)

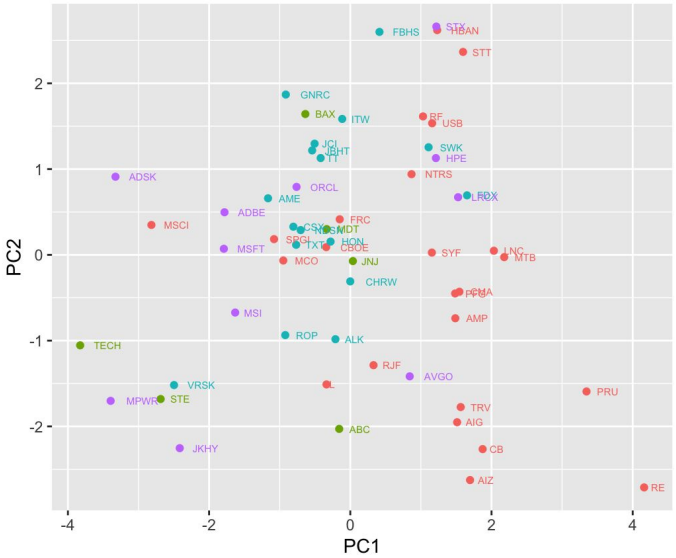
Results and Discussion

Variance-Covariance Structure via PCA

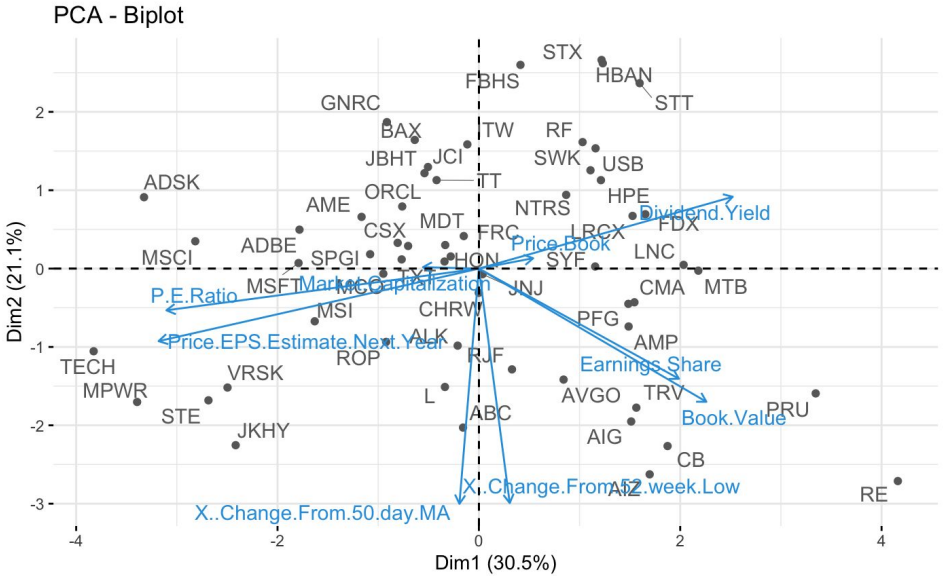
- Using the scree plot below, we have determined that two principal components will be sufficient.
- The elbow begins with the third principal component however, we shall use two for visualization purposes and then consider three.



Visualization of Two Principal Components



Sector
- Financials
- Health Care
- Industrials
- Information Technology

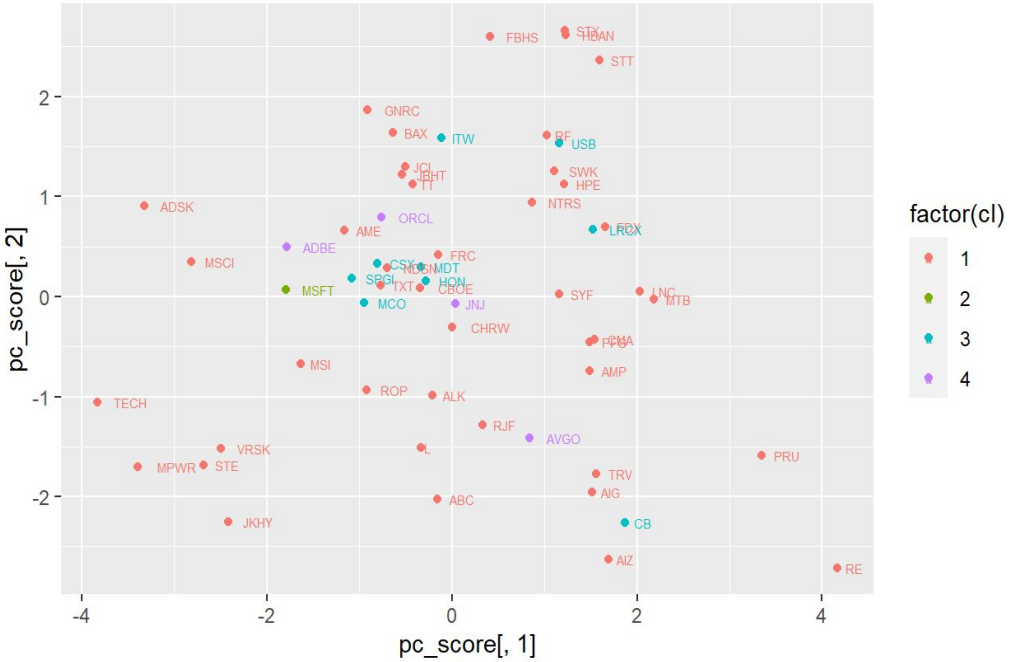


Using Three Principal Components

	PC1	PC2	PC3
Market.Capitalization	-0.09268162	0.003637526	-0.3109707
X..Change.From.52.week.Low	0.05112164	-0.600521760	0.3086315
X..Change.From.50.day.MA	-0.03222309	-0.602416872	0.3257740
Earnings.Share	0.33166894	-0.282128836	-0.5385151
Price.Book	0.09007860	0.026595080	-0.3235114
P.E.Ratio	-0.51800746	-0.106865588	-0.1719984
Book.Value	0.37714650	-0.341111165	-0.3797474
Price.EPS.Estimate.Next.Year	-0.53097211	-0.185833672	-0.1746106
Dividend.Yield	0.42087685	0.183954025	0.3208996

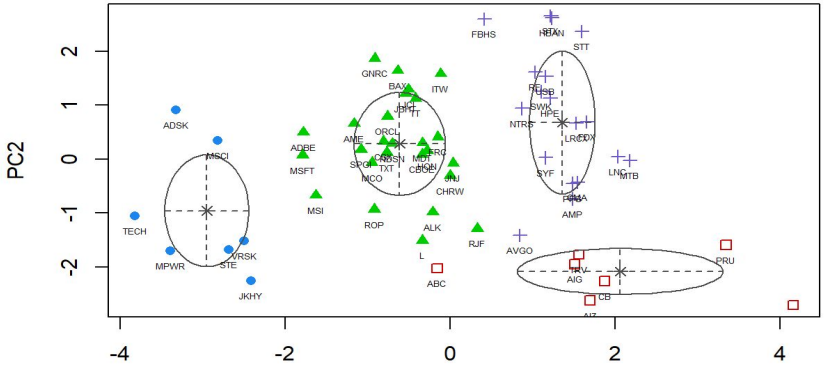
Clustering via K-Means

- We expected there to be four groups however, there aren't four clear clusters formed using K-Means.
- This may be due to using only the first two principal components which have a combined proportion of 52% of the variation explained by the variances.



GMM using Four Clusters

- Most of companies in red group are belong Financials sector.
- Most of companies in light blue and green group are belong industrial and information technology sector.

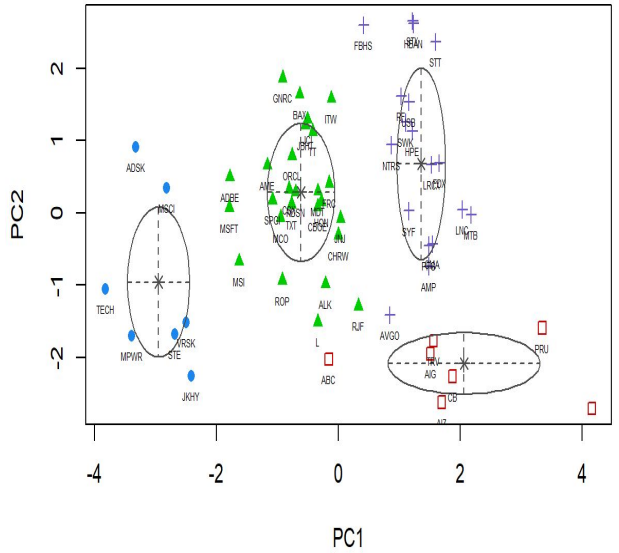


	A	B	C	D	E	F	G	H	I	J	K
1	Tickers	Company	GICS.Sector	Market Capitalizat	% Change From 52-week	% Change F	% Chang	Earnings	Ex-Divid	Price/Bo	P/E Rati
16	AIG	AIG	Financials	51519250432	0.43466547	-0.01540832	0.045519	10.82	1.65E+09	0.79904	5.90573
52	AIZ	Assurant	Financials	10925355008	0.313081	-0.00838043	0.091161	22.826	1.65E+09	1.922753	8.294051
106	CB	Chubb	Financials	89772703744	0.36031586	-0.03233025	0.01677	19.27	1.65E+09	1.513794	10.99689
380	PRU	Prudential	Financials	44606480384	0.25383556	-0.04604735	0.033389	19.505	1.64E+09	0.720632	6.075366
448	TRV	Travelers	Financials	44444278784	0.28233176	-0.01468238	0.044947	14.49	1.65E+09	1.546542	12.78261
499											

	A	B	C	D	E	F	G
1	Tickers	Company	GICS.Sector	Marl	% Change Fı	% Change F	Earnings
189	FDX	FedEx	Industrials	5E+10	0.021253057	-0.07939925	19.091
227	HPE	Hewlett Packard Ent	Information Technology	2E+10	0.20939186	-0.05310106	2.795
231	HON	Honeywell	Industrials	1E+11	0.10119258	0.01064479	7.91
275	LRCX	Lam Research	Information Technology	7E+10	0.04334809	-0.11885258	32.084
333	NDSN	Nordson	Industrials	1E+10	0.09893514	-0.03733647	8.47
349	ORCL	Oracle	Information Technology	2E+11	0.11376901	-0.01456604	2.664
415	SWK	Stanley Black & Deck	Industrials	2E+10	0.025984505	-0.07492348	10.158
434	TXT	Textron	Industrials	1E+10	0.19504537	-0.03690417	3.296

1	Tickers	Company	GICS.Sector	Marl	% Change Fi	% Change F	Earnings	Price/Bo	P/E Ratio	Book Va	Price/EP
16	AIG	AIG	Financials	5E+10	0.43466547	0.04551853	10.82	0.79904	5.90573	79.971	10.30645
105	CB	Chubb	Financials	9E+10	0.36031586	0.016770445	19.27	1.513794	10.99689	139.986	12.94502
191	FDX	FedEx	Industrials	5E+10	0.021253057	-0.07939925	19.091	2.147945	10.6469	94.63	9.045839
229	HPE	Hewlett Packard Ent	Information Technology	2E+10	0.20939186	-0.05310106	2.795	1.005376	5.620751	15.626	7.076577
277	LRCX	Lam Research	Information Technology	7E+10	0.04334809	-0.11885258	32.084	10.19279	14.66619	46.165	12.29553
379	PRU	Prudential	Financials	4E+10	0.25383556	0.033389345	19.505	0.720632	6.075366	164.439	9.032012
421	SWK	Stanley Black & Decl	Industrials	2E+10	0.025984505	-0.07492348	10.158	2.086919	13.79898	67.166	10.44486
447	TRV	Travelers	Financials	4E+10	0.28233176	0.044947024	14.49	1.546542	12.78261	119.764	12.80913

1	Tickers	Company	GICS.Sector	Price/EP	Dividend Yi
189	FDX	FedEx	Industrials	9.045839	0.01458151
227	HPE	Hewlett Packard Enterprise	Information Technology	7.076577	0.031209363
274	SWK	Stanley Black & Decker	Industrials	10.44486	0.02123717
275	LRCX	Lam Research	Information Technology	12.29553	0.012276394
317	MPWR	Monolithic Power Systems	Information Technology	37.59224	0.005862524
323	MSCI	MSCI	Financials	35.96765	0.007540916
417	STE	Steris	Health Care	29.53602	0.006790477
458	VRSK	Verisk	Industrials	32.64452	0.00543835



- The percentage change of the companies in red group is higher than blue group.

- The Price/EPS of the companies in light blue group is lower than the heavy blue group.Vice versa with dividend yield of the two groups.

Implementing CCA

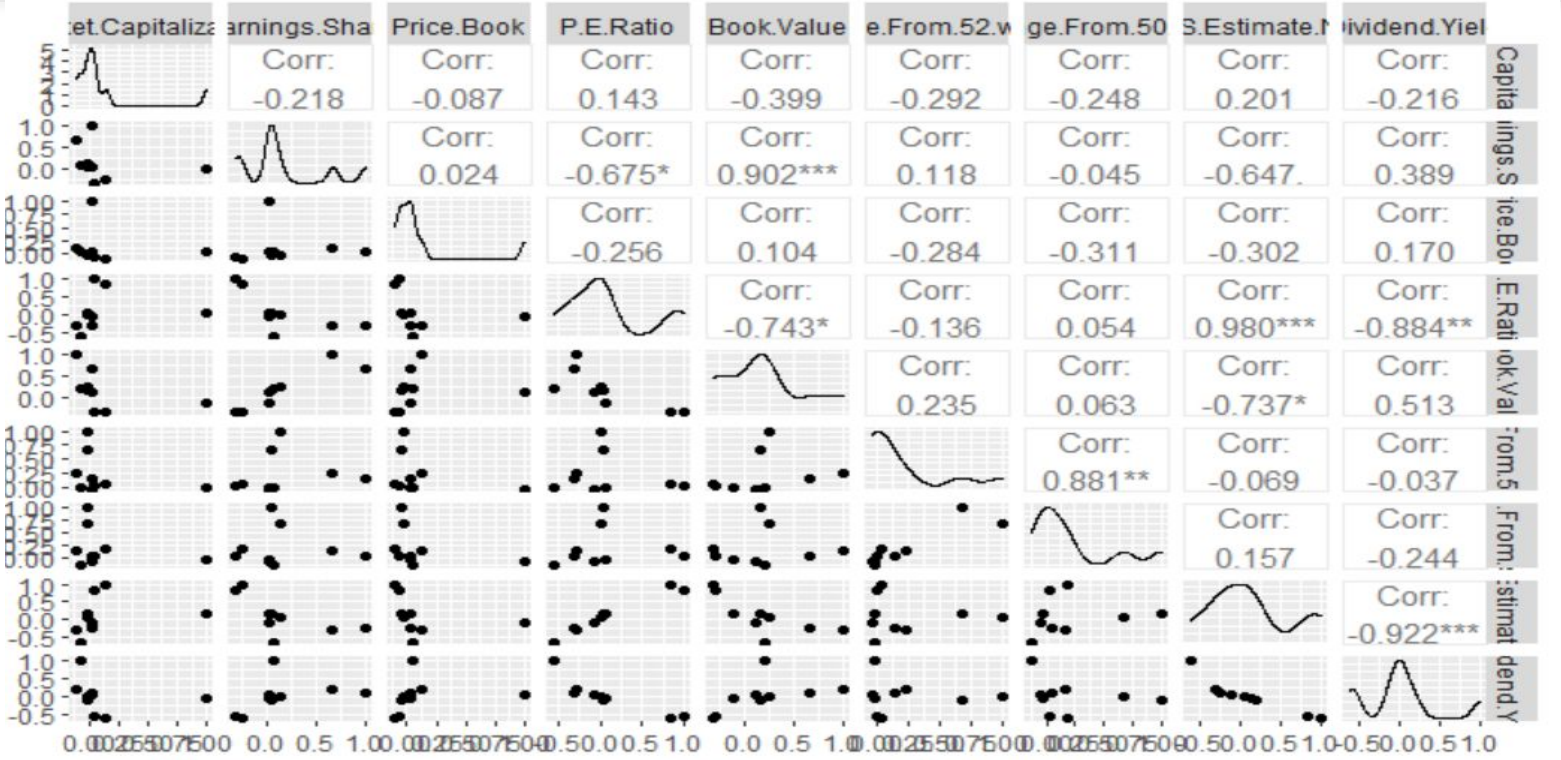
- First of all we find the correlation among all variables. Use the correlation coefficients to put high correlation variables into the same groups.

- In this pattern, we shuffle around all variables and put them in two groups, then perform CCA. We try to find out how we divide the data, so that we will obtain the minimal canonical correlations in first dimension. The following table shows the correlation between all variables.

	Market.Capitalization	Earnings.Share	Price.Book	P.E.Ratio	Book.Value	X..Change.From.52.week
Market.Capitalization	1.000000000	0.02282059	0.03253220	0.049278760	-0.1038002	-0.00959
Earnings.Share	0.022820589	1.00000000	0.03099675	-0.303939563	0.6687681	0.14396
Price.Book	0.032532197	0.03099675	1.00000000	-0.065022853	0.1225171	-0.02760
P.E.Ratio	0.049278760	-0.30393956	-0.06502285	1.000000000	-0.3001582	0.00715
Book.Value	-0.103800239	0.66876814	0.12251714	-0.300158249	1.0000000	0.24544
X..Change.From.52.week.Low	-0.009599575	0.14396601	-0.02760033	0.007159557	0.2454450	1.00000
X..Change.From.50.day.MA	-0.022503275	0.05497081	-0.04105439	0.033259991	0.1708513	0.67506
Price.EPS.Estimate.Next.Year	0.143855460	-0.25107341	-0.10098329	0.843173736	-0.3149054	0.04097
Dividend.Yield	-0.075445551	0.07309201	0.04001033	-0.553666508	0.2075151	-0.00500

	X..Change.From.50.day.MA	Price.EPS.Estimate.Next.Year	Dividend.Yield
Market.Capitalization	-0.02250327	0.14385546	-0.075445551
Earnings.Share	0.05497081	-0.25107341	0.073092012
Price.Book	-0.04105439	-0.10098329	0.040010334
P.E.Ratio	0.03325999	0.84317374	-0.553666508
Book.Value	0.17085125	-0.31490545	0.207515100
X..Change.From.52.week.Low	0.67506556	0.04097307	-0.005001721
X..Change.From.50.day.MA	1.00000000	0.18879129	-0.129359672
Price.EPS.Estimate.Next.Year	0.18879129	1.00000000	-0.625630017
Dividend.Yield	-0.12935967	-0.62563002	1.000000000

Visualizing the Correlations



Canonical Correlations

- Using price, earnings per share, dividend yield, and PE ratio as the response and the remaining variables as the predictors, we obtain the following canonical correlations.

Dimension 1	Dimension 2	Dimension 3
0.4691065	0.3379210	0.1335079

- Using next year estimated EPS, and dividend yield, we obtain the following instead using two dimensions.

Dimension 1	Dimension 2
0.8778952	0.1814300

Adjusting the Responses

- Now using change in low price from 52 weeks and change in high price from 50 day as the response.

Dimension 1	Dimension 2
0.4126308	0.2469436

- Now we use market capitalization and price.

Dimension 1	Dimension 2
0.29259990	0.094227760

Understanding the Canonical Coefficients

- The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients.
- For the variable Market.Capitalization, a one unit increase in Market.Capitalization leads to a .85037 decrease in the first canonical variate of set 1 when all of the other variables are held constant.

```
$xcoef
                              [,1]      [,2]
Market.Capitalization -0.8503753 0.5271824
Price.Book             0.5545679 0.8327748

$ycoef
                                     [,1]        [,2]
Earnings.Share                -0.6072701 -0.14105865
P.E.Ratio                      0.8351087 -1.12797979
Book.Value                     0.8203251  0.77766167
X..Change.From.52.week.Low    -0.2712409  0.08107826
X..Change.From.50.day.MA       0.2695405 -0.86730942
Price.EPS.Estimate.Next.Year  -1.3661284  1.04391043
Dividend.Yield                -0.1895540 -0.30277179
```

The Canonical Loadings

- The below correlations are between observed variables and canonical variables which are known as the canonical loadings.
- These canonical variates are actually a type of latent variable.

```
$corr.X.xscores
                           [,1]      [,2]
Market.Capitalization -0.8323340 0.5542744
Price.Book             0.5269033 0.8499252

$corr.Y.xscores
                                    [,1]         [,2]
Earnings.Share                -0.002216261  0.037843927
P.E.Ratio                     -0.077965030 -0.028170502
Book.Value                     0.156213235  0.047307530
X..Change.From.52.week.Low    -0.007143017 -0.028045589
X..Change.From.50.day.MA      -0.003631217 -0.046052391
Price.EPS.Estimate.Next.Year  -0.178333225 -0.008258275
Dividend.Yield                 0.086345483 -0.006453966
```

```
$corr.X.yscores
                           [,1]       [,2]
Market.Capitalization -0.2435408 0.05222803
Price.Book             0.1541719 0.08008655

$corr.Y.yscores
                                    [,1]        [,2]
Earnings.Share                -0.007574375  0.40162185
P.E.Ratio                     -0.266456108 -0.29896181
Book.Value                     0.533880006  0.50205513
X..Change.From.52.week.Low    -0.024412233 -0.29763616
X..Change.From.50.day.MA      -0.012410177 -0.48873486
Price.EPS.Estimate.Next.Year  -0.609478085 -0.08764164
Dividend.Yield                 0.295097448 -0.06849325
```

Conclusion and Further Analysis

- The proportion of variation in the dataset due to the variance can be explained by the first three principle components.
- Using GMM, we can see that there are four distinct clusters where groups 2 and 3 are less volatile than 1 and 4 because the points are more concentrated in their respective confidence ellipse. We recommend for investors to purchase stocks in the less volatile groups in order to minimize the risk.
- From CCA, we can see that two canonical dimensions are statistically significant.
- We hope to examine how the concept of grouping stocks can relate to determining the index as done with index funds.
- Examine how PCA using correlation of SVM differs from using the variances.


thank
you