**Introduction**

Canonical corelation analysis is used to identify and measure the associations among two sets of variables. Canonical correlation is appropriate in the same situations where multiple regression would be, but where there are multiple inter-correlated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

Canonical correlation analysis focuses on the correlation between a linear combination of variables in one set and linear combination of the variables in another set. The idea is to determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and so on. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations.

The canonical correlations measure the strength of association between the two set of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pairs of canonical variables.

**Theory:**

We shall be interested in measures of association between two groups of variables. The first group of p variables is represented by the (px1) random vector X(1). The second group of q variables is represented by the (qx1) random vector X(2). For the random vectors X(2) and X(1), let

E(X(2))= Cov(X(2))=

E(X(1))= Cov(X(1))=

Cov(X(1), X(1))=

Linear combination provide simple summary measures of a set of variables. Set

U = a’ X(1)

V= b’ X(2)

For some pair of coefficient vectors a and b. Then, we know

Var(U) = a’ Cov(X(1))a= a’a

Var(V)= b’Cov(X(2))b=b’b

Cov(U,V)= a’COV(X(1), X(2))b= a’b

We shall seek coefficient vectors a and b such that

Corr(U,V) =

Is as large as possible.

We define the following:

The first pair of canonical variables, or first canonical variate pair, is the pair of linear combinations, U1 and V1, having unit variances, which maximize the correlation.

The second pair of canonical variables, or second canonical variate pair, is the pair of linear combinations U1, V2 having unit variances, which maximize the correlation among all choices and is uncorrelated to the first canonical variables.

At the kth step,

The kth pair of canonical variables, is the pair of linear combinations Uk, Vk having unit variances, which maximize the correlation among all choices uncorrelated with the previous k-1 canonical pairs.

**Data**

We downloaded a dataset consisting of closing stock market values for 10 companies, namely, Boing, NASDAQ, Cocacola, DELL, General Electrics, General Motors, IBM, Mc Donalds,Microsoft and Pepsi for everday starting from 5/4/1989 to 3/4/2009 generating 5000 data for each companies.

We plan to generate best canonical pairs by playing with the number of variables available. Best canonical pair is the one with the highest correlation coefficient for the first canonical pairs.

We first analyzed data to study correlation among the variables.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | BA |  | NASDAQ | COCA | DELL | GE | GM | IBM | MD | MSFT | PEPSI | | BA | 1 |  | 0.8358 | 0.812 | 0.8084 | 0.8622 | 0.348 | 0.8199 | 0.8805 | 0.8304 | 0.9089 | | NASDAQ | 0.8358 |  | 1 | 0.8846 | 0.947 | 0.9723 | 0.5809 | 0.8863 | 0.9111 | 0.9587 | 0.8975 | | COCA | 0.812 |  | 0.8846 | 1 | 0.9288 | 0.9142 | 0.5947 | 0.8773 | 0.8176 | 0.9073 | 0.8533 | | DELL | 0.8084 |  | 0.947 | 0.9288 | 1 | 0.9774 | 0.557 | 0.9215 | 0.8626 | 0.9851 | 0.9294 | | GE | 0.8622 |  | 0.9723 | 0.9142 | 0.9774 | 1 | 0.594 | 0.9256 | 0.8869 | 0.9749 | 0.9374 | | GM | 0.348 |  | 0.5809 | 0.5947 | 0.557 | 0.594 | 1 | 0.443 | 0.3116 | 0.5307 | 0.3579 | | IBM | 0.8199 |  | 0.8863 | 0.8773 | 0.9215 | 0.9256 | 0.443 | 1 | 0.8242 | 0.8951 | 0.8788 | | MD | 0.8805 |  | 0.9111 | 0.8176 | 0.8626 | 0.8869 | 0.3116 | 0.8242 | 1 | 0.9063 | 0.9233 | | MSFT | 0.8304 |  | 0.9587 | 0.9073 | 0.9851 | 0.9749 | 0.5307 | 0.8951 | 0.9063 | 1 | 0.9413 | | PEPSI | 0.9089 |  | 0.8975 | 0.8533 | 0.9294 | 0.9374 | 0.3579 | 0.8788 | 0.9233 | 0.9413 | 1 | |  |  |  |  |  |  |  |  |  |  |
| Table 1. Correlation Matrix |  |  |  |  |  |  |  |  |  |  |

The highest correlation seen in the correlation structure is .9851 between stocks of DELL and MICROSOFT, which is intuitive as they both are related to technology sector. We can see that Microsoft and NASDAQ also has high correlation.

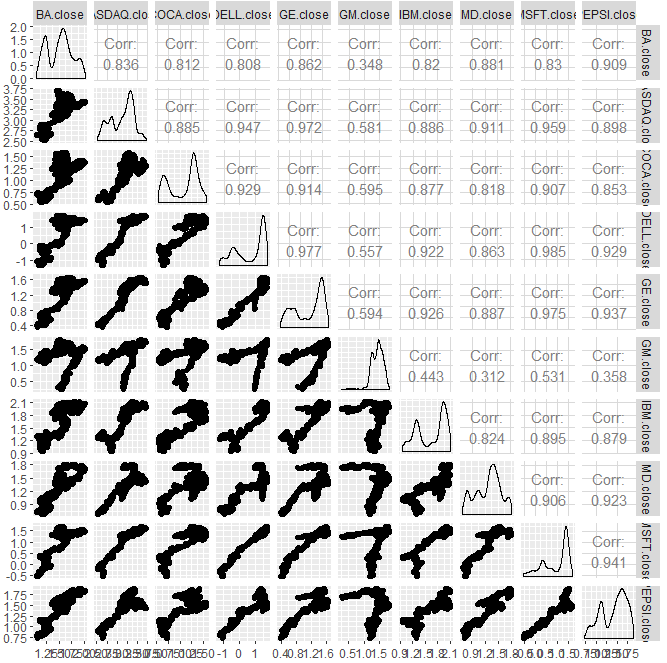
.

Fig. Matrix of plot of given data

We also plotted the scatter plots for all the variables of two data sets to see the underlying structure. We can develop some understanding as to how to group the data from this plot.

We can see that correlation between the variables in a set is pretty high for the first set consisting of data Boing, NASDAQ, DELL, IBM, Microsoft.

Our first intuitive grouping for canonical correlation Analysis (CCA) is to group all the technical companies together and non-technical companies together.

We grouped Boeing, NASDAQ,DELL,IBM, Microsoft together and rest of the other companies together. Lets see the observations.

Correlation between first canonical elements was seen to be 0.73, which is a good start.

Then, we group those data that see good correlation structure from Fig 1.

1st Set: Mc Donalds, Dell, GE, Nasdaq

2nd set: Boeing, Coc, GM,IBM,Microsoft,Pepsi

In this case, the correlation coeff was 0.801, which is quite good.

We tried with a few other combinations of the data to come up with the best correlation coefficient for the first coefficient pair.

**Canonical Correlation Coefficient for various combinations:**

1st Set: Boeing, Nasdaq, DELL, GE, GM,Microsoft,IBM

2nd Set: CocaCola, MD, Pepsi

Corelation: 0.58 (for 1st canonical pair)

1st Set: Boeing, DELL,Microsoft

2nd Set: CocaCola, Pepsi, GE, GM, Nasdaq, ,IBM,MD

Correlation: 0.677 (for 1st canonical pair)

1st Set: Boeing, Nasdaq, DELL, GE, GM,Microsoft,IBM,MD

2nd Set: CocaCola, Pepsi

Correlation: 0.677 (for 1st canonical pair)

**Best Result:**

1st set: Being, Nasdaq,Cocacola,Dell, IBM,MD

2nd Set: GE,GM,Microsoft,Pepsi

The correlation coeff between the first canonical pair was 0.828 in this case.

Under more analysis, we may be able to find out some data set and combination that would result in correlation coefficient to be 1 for the first canonical correlation analysis.

The correlation coefficients obtained for the best case were:

|  |  |  |  |
| --- | --- | --- | --- |
| 0.8282 | 0.408 | 0.290 | 0.142 |

As we can see, the correlation coefficient for the first canonical variates is very large implying that the linear combinations we obtain the first canonical variables were highly correlated to each other.

The coefficients of the variables in this case were:

|  |  |
| --- | --- |
| McDonalds | -0.12 |
| Boeing | -0.065 |
| DELL | -0.028 |
| Cocacola | -0.049 |
| IBM | -0.042 |
| NASDAQ | -0.229 |

|  |  |
| --- | --- |
| Pepsi | -0.015 |
| Microsoft | -0.155 |
| GE | -0.15 |
| GM | -0.0760 |

We obtained the highest correlation coefficient when we divided the portfolio as above.

Conclusion:

In this lab, we were able to perform canonical correlation analysis on the stock data, and were able to generate highest correlation at one of the possible combinations. We could try all the options to obtain 1 as the coefficient.