CSIC 5011 Project 1

Anonymous Author(s)
Affiliation
Address
email

Qi Bu
qbu@connect.ust.hk

1 Introduction

In this project I chose the financial data containing closed prices of 452 stocks in 4 years, totally 1258*452 data points. I would like to first test how well PCA is to approximate the stock prices and then use inverse of correlation coefficient as "distances" between every two stocks and see whether we can construct a low-dimensional visualization indicating the relations of the stocks.

2 PCA

The raw data X is a 1258-by-452 matrix, and we may find the corresponding eigenvalues and eigenvectors of $\hat{\Sigma_n}$:

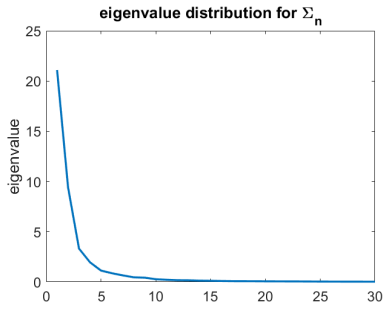

Figure 1: Largest 30 eigenvalues of the matrix $\hat{\Sigma_n}$.

According to the figure, although the matrix $\hat{\Sigma_n}$ is 1258-by-1258, the largest several eigenvalues dominate others. Hence, I picked the top five eigenvalues and corresponding eigenvectors to reconstruct the data. The correlation coefficients between the raw data and the reconstructed ones for each stock is shown in the figure below.

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

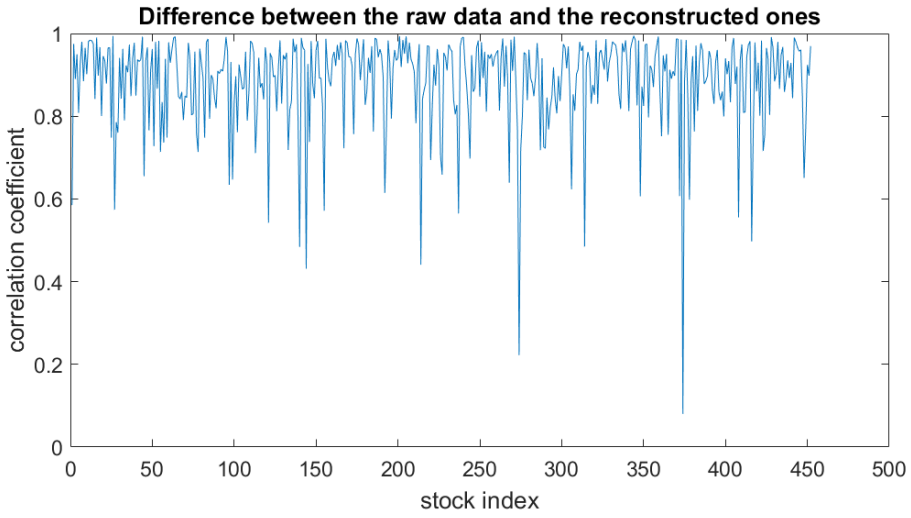

Figure 2: Largest 30 eigenvalues of the matrix $\hat{\Sigma_n}$.

From Figure. 2, the correlation coefficients almost keep high, and we may conclude that most of the reconstructed stock prices are similar to the raw data in terms of linear correlation.

3 MDS

For classical MDS, the input should be a squared distance matrix. However, for financial data it may not be appropriate to use Euclidean distance. Therefore I would like to consider using correlation coefficient between each two time series of stocks. The problem is that the correlation coefficient is between -1 and 1, with values close to -1 or 1 indicating strong correlation. Therefore, I would like to use the inverse of absolute values of correlation coefficients, and then take logarithms as "pseudo distances". In this way, we make the diagnal terms of the distance matrix be zero, and a larger distance implies a weaker correlation.

The result after running the MDS program is shown on the figure below.

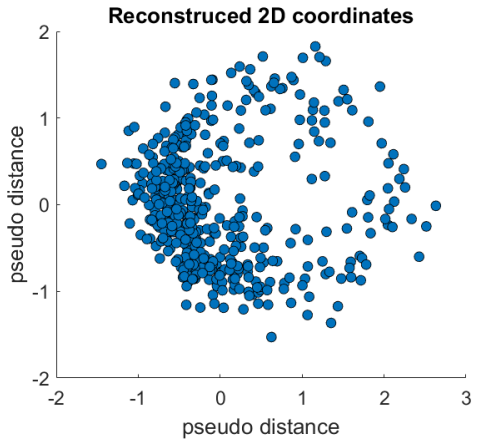

Figure 3: Reconstructed coordinates of stocks after performing MDS.

We may clearly find a "cluster" in this 2D graph. Since the distances here are related to the inverse of correlation coefficients, I think this figure suggests a close relation between different stocks and the change of one stock price may affect other closely correlated stocks.

4 Summary

In this project I used PCA and MDS to analyze the financial data. By using a pseudo distance related to correlation coefficient of stocks, a "cluster" of highly correlated stocks were found.

Appendix

Here is the matlab program used in this project:

```
%-------------------------------------------
% CSIC 5011 Mini Project 1
% Qi Bu 20090594
% Oct. 2017
%-------------------------------------------

load('snp452-data.mat')

%-------------------------------------------
% PCA for financial data 1258*452
%-------------------------------------------

k = 5;    % dimension of the affined space

% we need to preprocess the data
for stk = 1 : size(X, 2)
    X(:, stk) = X(:, stk) ./ max( X(:, stk) );
end
mu_n = mean(X, 2);
Y = X - mu_n * ones(1, size(X, 2));
Sigma = Y * Y' / size(X, 2);

% find the eigenvalues and eigenvectors
[V, D] = eigs(Sigma, 30 );
figure
plot(diag(D), 'LineWidth',2)
ylabel('eigenvalue')
set(gca, 'FontSize', 14)
title('eigenvalue distribution for {\Sigma_n}')
hold off

% find the projections on the top-k eigenvectors and reconstruct the data
X_rec = zeros( size(X) );
for stk_num = 1 : size(X, 2)
    X_rec(:, stk_num) = mu_n;
    coeffs = zeros(k,1);
    for dim = 1 : k
        coeff = V(:, dim)' * Y(:, stk_num);
        coeffs(dim) = coeff;
        X_rec(:, stk_num) = X_rec(:, stk_num) + V(:, dim) * coeff;
    end
end
```

```
% find the difference between the real stock prices and reconstructed ones
CorrS = zeros(size(X, 2), 1);
for stk = 1 : size(X, 2)
    Corr = corrcoef(X(:, stk), X_rec(:, stk));
    CorrS( stk ) = Corr(1, 2);
end
figure
plot(CorrS)
xlabel('stock index')
ylabel('correlation coefficient')
title('Difference between the raw data and the reconstructed ones')
set(gca, 'FontSize', 14)
hold off


%-------------------------------------------
% MDS for financial data 1258*452
%-------------------------------------------

R = abs(corrcoef( X )) .^ (-1);
R = log(R);
pos = cmdscale( R, 2 );
figure
scatter( pos(:, 1), pos(:, 2), 50, 'filled', 'MarkerEdgeColor', 'k' )
xlabel('pseudo distance')
ylabel('pseudo distance')
title('Reconstruced 2D coordinates')
set(gca, 'FontSize', 14)
hold off
```