

Chapter 1

Stock Market Values: Introduction to Principal Component Analysis

In today's data-rich society, it is common for researchers to collect large data sets. Large data sets can be difficult to interpret and it is likely that several of the variables within large data sets are correlated. Principal component analysis (PCA) uses the variance-covariance matrix of the data to create new variables (components) that consist of uncorrelated, linear combinations of the original variables. PCA is used to simplify the data structure and still account for as much of the total variation in the original data as possible.

1.1 A Geometric Interpretation

To illustrate principal components geometrically, a relatively small data set will be used. Clearly, low-dimensional data typically would not need to be reduced, so the example provided in sections 1.1 and 1.2 may not be the best use of principal components. However this small data set is used to allow for easy visual interpretation and calculations.

The goal of principal components is to transform a large number of original variables to a smaller set of uncorrelated components that capture most of the variability appearing in the data. Open the file labeled **2006 Financial Data**. The first two columns of this data set will be called matrix \mathbf{X} . The first column, Dow (also called vector \mathbf{X}_1) gives the closing Dow stock market values for every business day in 2006. Vector \mathbf{X}_2 gives the closing S&P 500 values for every business day in 2006. Each vector \mathbf{X}_i represents one variable of interest and consists of the n observations for that variable. $n = 251$ which was the number of business days in 2006.

1. Standardize the two data vectors, \mathbf{X}_1 and \mathbf{X}_2 , and store in \mathbf{Z}_1 and \mathbf{Z}_2 :

[In Minitab, select: Calc \Rightarrow Standardize. In Input Column(s) select Dow. Check Subtract mean and divide by std. dev. For Store results in: type Z1 , click OK]. Repeat this process for S&P, storing in Z2. You may choose to use different labels, but for the remainder of this lab, \mathbf{Z}_1 and \mathbf{Z}_2 will refer to the standardized Dow and S&P.

2. **Create time series plots of the standardized Dow and S&P, \mathbf{Z}_1 and \mathbf{Z}_2**

[In the Minitab pull down menu: Graph \Rightarrow Time Series Plot (select Simple) click OK... In series, double click to select Z1 Z2 click OK]

Do you see similar patterns in these graphs (i.e. is the data correlated)? Describe why or why not you would expect this.

3. **Create a scatterplot of the data** [In the Minitab pull down menu: Graph \Rightarrow Scatterplot (select Simple) click OK... Double click to select Z1 as the X variable and Z2 in any of the Y variable, click OK]

Note that you may need to adjust the axes in order to the graphs provided in this lab [Double click the Y axis on the Minitab graph you just created : On the Scale tab (select Minimum and Maximum) input appropriate minimum and maximum values, click OK]

Does the data look highly correlated? With your pencil, draw a line that represents the direction of the most variability in the scatterplot. If done correctly, the line drawn will represent the first **eigenvector**, a vector representing the direction of the variation in the data.

To understand the variability of our data matrix, \mathbf{X} , calculate the correlation matrix.

$$\text{Corr}(\mathbf{X}) = \mathbf{R} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \tag{1.1.1}$$

where $r_{21} = r_{12} = \frac{S_{12}}{S_1 S_2} = \frac{\text{Cov}(\mathbf{X}_1, \mathbf{X}_2)}{\sqrt{\text{Var}(\mathbf{X}_1)} \sqrt{\text{Var}(\mathbf{X}_2)}}$ is the sample correlation between data vectors \mathbf{X}_1 and \mathbf{X}_2 .

4. **Calculate the correlation matrix, \mathbf{R}**

[In Minitab, select: Stat \Rightarrow Basic Statistics \Rightarrow Correlation ... In Variables select Z1 Z2. Check Store matrix (display nothing), click OK]. This automatically stores the correlation matrix as Corr1. To view this matrix it is best to use the session editor.

[In Minitab, click your mouse anywhere on the session window, then select **Editor** \Rightarrow **Enable commands**]

The session window will now show a prompt which accepts written commands. After the `MTB >` prompt type `print Corr1` to view the correlation matrix. Minitab is not case sensitive, so `corr1` or `CORR1` are considered identical names for the same object. Submit the 2 by 2 correlation matrix, **R**.

The scatterplot created in question 3) is the standardized version of Figure 1.1. Both graphs show that the two variables that are highly correlated:

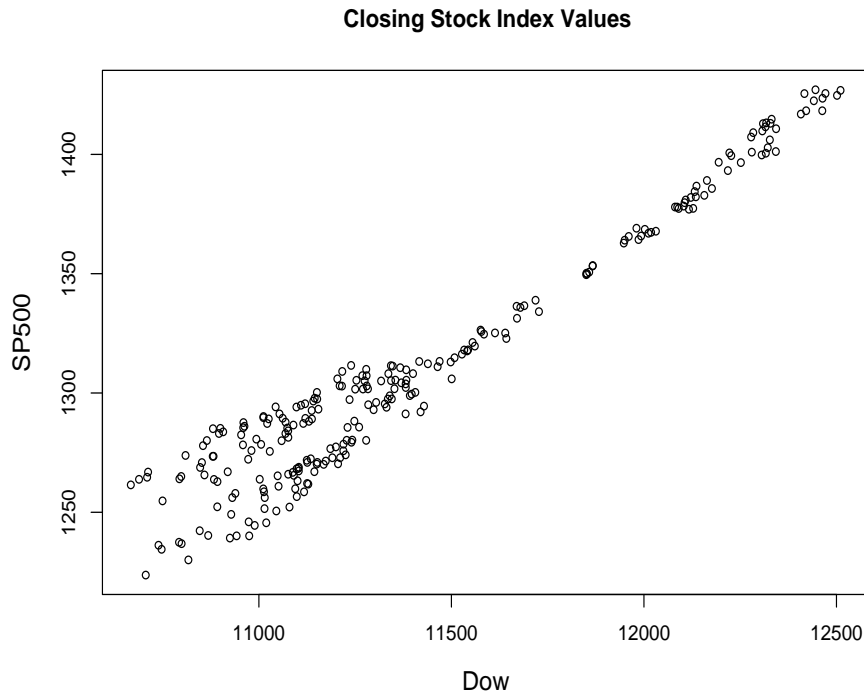


Figure 1.1: 2006 daily closing values for the Dow and S&P 500 indices.

5. Calculate the eigenvectors of **R**

[In Minitab, select **Calc** \Rightarrow **Matrices** \Rightarrow **Eigen Analysis**. In **Analyze Matrix** input `Corr1`, then type a column named `EValues` to place the eigenvalues and type `EVectors` as the Matrix of eigenvectors, click **OK**] Submit the eigenvalues and eigenvectors that have been calculated.

Figure 1.2 is a plot of the standardized data, \mathbf{Z}_1 and \mathbf{Z}_2 . The first eigenvector, \mathbf{v}_1 , is drawn from the origin $(0, 0)$ to the point $(.707, .707)$. \mathbf{v}_1 is in the direction representing the largest

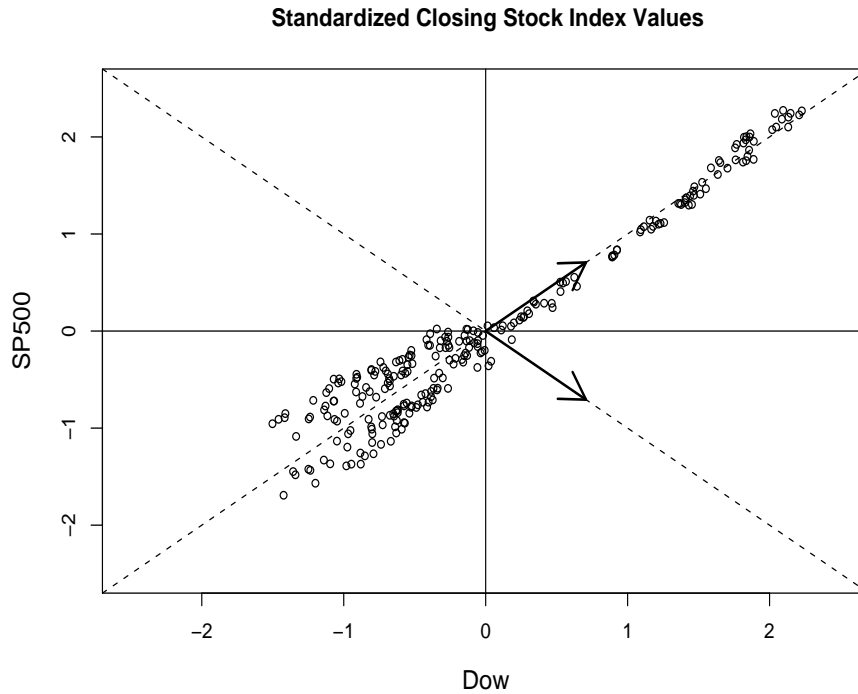


Figure 1.2: Standardized daily closing values for the Dow and S&P 500 indices with the eigenvectors of the correlation matrix shown.

amount of variability. \mathbf{v}_2 is perpendicular to \mathbf{v}_1 and, expresses the direction of the second largest amount of variability. Eigenvectors better illuminate the variation of the data set and can be used to rotate the data in the direction of the most variability.

The eigenvectors are sorted according to the size of their eigenvalues. The eigenvector corresponding to the largest eigenvalue is considered the “first” eigenvector, \mathbf{v}_1 . The second largest eigenvalue determines the “second” eigenvector, \mathbf{v}_2 , and so on.

Note that the sign of the eigenvectors is arbitrary. In this example, it makes no difference whether:

$$\mathbf{v}_1 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \quad \text{or} \quad \mathbf{v}_1 = \begin{bmatrix} -0.707 \\ -0.707 \end{bmatrix} .$$

The dashed lines in Figure 1.2 show both the positive and negative directions of the two eigenvectors. An eigenvector (and hence a principal component) can be multiplied by -1 without impacting the interpretation of the results in any way.

To calculate the principal components, multiply the standardized data by each eigenvector

using the following formula:

$$\mathbf{Y}_i = \mathbf{Z}\mathbf{v}_i = v_{i1}\mathbf{Z}_1 + v_{i2}\mathbf{Z}_2 + \dots + v_{ik}\mathbf{Z}_k \quad i = 1, \dots, k \quad (1.1.2)$$

In our example there are only 2 columns of data, thus $k = 2$, but these calculations are typically conducted for much larger data sets (see Johnson, R., and Wichern, D., *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982, p. 363).

This will create two new variables, Y_1 and Y_2 . Since the eigenvectors are ordered according to the size of their corresponding eigenvalues, Y_1 will be the linear combination of the data with the most spread. Y_2 will be uncorrelated to Y_1 and will include the rest of the total variability.

Figure 1.3 represents the impacts of the data that is now rotated corresponding to the eigenvectors \mathbf{v}_1 and \mathbf{v}_2 forming the new axes.

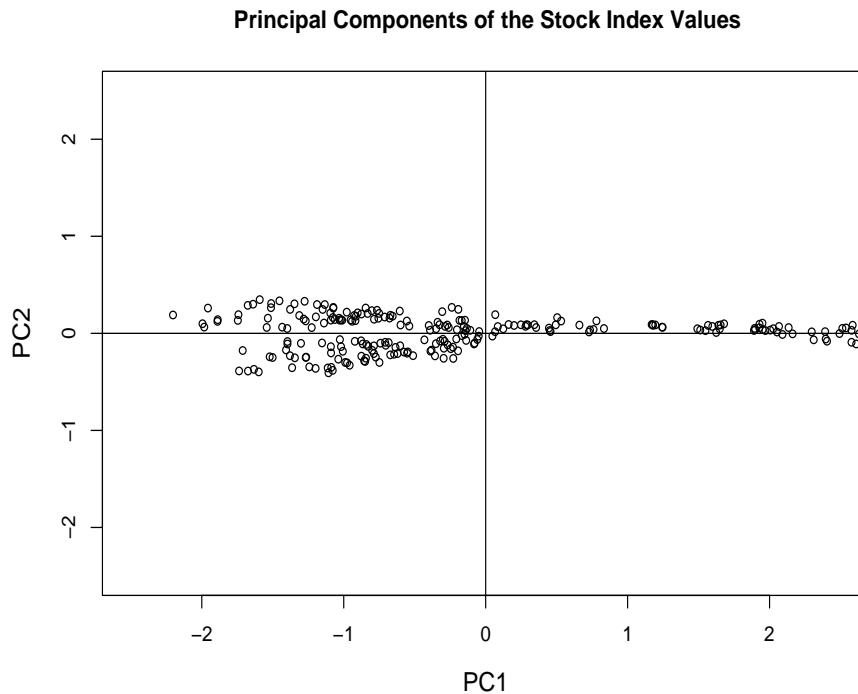


Figure 1.3: Principal components of the daily closing values for the Dow and S&P 500 indices.

6. Calculate the Principle Components and Create a Timeseries Plot

[In Minitab, select Calc \Rightarrow Calculator. In Store result in Variable input PC1. In Expression input $Z1*(.707) + Z2*(.707)$, click OK] Repeat the process using \mathbf{v}_2 to create PC2.

Submit a time series plot of Z1, Z2, PC1, and PC2. [In Minitab, select Graph \Rightarrow Time Series Plot. Select Simple. In Series input Z1 Z2 PC1 PC2. In Multiple Graphs select Overlaid on the same graph] Explain how PC1 and PC2 are related to Z1 and Z2.

7. Determine the Number of Principle Components to Use

Divide the first eigenvalue by the sum of all eigenvalues (only 2 in this example). The result is the proportion of the total variation that is explained by the first principle component. What percentage of the variation is explained by the first principle component? What percentage of the variation is explained by the second principle component? Since PC1 explains a majority of the variation, there is no need to use PC2. Thus PC1 can be used to reduce the two dimensional data set to just one dimension.

1.2 A 3 Dimensional Example

Open the Minitab file “2006 Financial Data.” This file contains the daily closing stock prices of the Dow, S&P 500, and Nasdaq financial indexes for 2006. In this data set $n = 251$ which is the number of business days in 2006 and $k = 3$ representing the three variables of interest.

8. Standardize Nasdaq into a column Z3

9. Create time series plots of this data

[In the Minitab pull down menu: Graph \Rightarrow Time Series Plot (select Multiple) click OK... In series, double click to select Z1 Z2 Z3 click OK]

Describe any patterns that you see. With a pencil sketch a line on the time series plot displaying the general pattern that you would expect the first principal component to show.

10. Create a 3-D plot of the data

[In the Minitab pull down menu: Graph \Rightarrow 3DScatterplot (select Simple) click OK... Double click to select Z1 Z2 Z3 in any of the X, Y, and Z variables, click OK]

Use the arrows on the 3D Graph Tools to rotate the graph. Rotate in each direction until the 3D graph looks like a 2D graph with S&P on the Y axis and Dow is on the X axis. This graph should look like Figure 1.1.

11. Calculate the Eigenvectors and Eigenvalues

[In Minitab, select Stat \Rightarrow Multivariate \Rightarrow Principal Components.... In Variables, select the Z1 Z2 Z3 In Graphs select Scree Plot] Notice that if you select Type of Matrix to be Correlation the variables (Z1 Z2 Z3) or (Dow S&P Nasdaq) will give identical results.

12. Calculate the Principle Components

In Minitab use the `Calc` \Rightarrow `Calculator` as in question 5) and equation (1.1.2)) to calculate PC1, PC2, and PC3. Note that in question 5) $k = 2$ columns, but now $k = 3$.

13. Visualize the First Two Principal Components in a 3 Dimensional Plot

First add the points corresponding to the eigenvectors to your scatterplot:

[In Minitab, add the eigenvectors to the bottom of the Z1 Z2 Z3 columns. In other words, write -0.540 in the 252th row of Z1, -0.582 in the 252th row of Z2, -0.608 in the 252th row of Z3]

Place the second eigenvector in the 253rd row of the Z1 Z2 Z3 columns and place (0,0,0) on the 254th row of the Z1 Z2 Z3 columns.

Second, create a new column that will create a code so that the "eigenvector points" are distinct from all the other data

Create a new column in the Minitab worksheet called `CODE`, write the word `original` in the first 251 rows of a new `CODE` column. Write `First` on the 252th row and write `Second` on the 253rd row and `Origin` on the 254th row. Be sure to click and drag to copy the word `original`, do not type it 251 times.

Third, draw a three dimensional graph.

[In the Minitab pull down menu: select `Graph` \Rightarrow `3DScatterplot` (select `With Groups`). Click OK. Select Z1 Z2 Z3 in any of the X, Y, and Z variables... select `CODE` for the Categorical variable. Click OK]

Use the arrows on the 3D Graph Tools to rotate the graph enough to convince yourself that PC1 (the vector from the `Origin` to the `First` point) is in the direction of the most variation.

Continue to rotate the graph to convince yourself that PC2 (the vector moving from the `Origin` to the `Second` point) is perpendicular to the PC1 vector. Submit this rotated graph on an attached page. *This graph does not need to be perfectly rotated.*

14. Determine How Many Principal Components to Use

In the correlation matrix, the diagonal elements sum to k . The eigenvalues will also sum to k . Thus the proportion of variation explained by the first principal component can be found by dividing the first eigenvalue by k .

The Scree plot is a simple graphic used to display the proportion of variation explained by each principal component. There is no exact technique to determine how many components to use, however, in this case, it is clear that the first eigenvalue has a very large proportion of the overall variance. In this case, only the first principal component should be used. Submit the percent of the variation that can be explained by the first principal component.

15. **Visualize the Principal Components in a Time Series Plot**

In this example, PCA is used to reduce the 251 by 3 matrix \mathbf{X} to a 251 by 1 vector consisting of the first principal component that can still explain 88.3% of the variability.

Create a time series plot. First delete the points added in Question 13 so that each column has 251 entries. [Graph \Rightarrow Time Series Plot \Rightarrow Multiple] of Z1 Z2 Z3 PC1 PC2 PC3. Remember that it might be more appropriate to plot $-\text{PC1}$ instead of PC1.

Submit this graph and explain how the time series pattern of PC1, PC2, and PC3 relate to Z1, Z2, and Z3.