Chapter 10

Introduction to Principal Component Analysis: Stock Market Values

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. -John Tukey

October. This is one of the peculiarly dangerous months to speculate in stocks. The others are July, January, September, April, November, May, March, June, December, August, and February. -Mark Twain

Advancements in technology have made data collection and computationally intensive statistical techniques much more feasible. Researchers from many disciplines are now routinely using automated calculations to analyze data sets with dozens or hundreds of variables. Principal component analysis (PCA) is an exploratory tool used to simplify a large and complex data set into a smaller, more easily understandable data set. This is different than ANOVA or multiple regression techniques that focus on statistical inference or model building to explain or predict relationships between variables. PCA summarizes complex data sets by by creating new variables that are linear combinations (weighted sums) of the original data. In addition, each of these new variables, called principal components, are constructed to be uncorrelated with all others. This chapter will

1. use key indices of the stock market to describe the process of creating principal components,
2. introduce linear algebra techniques useful for statistics, and
3. use principal component analysis to address a controversy in the climate change literature.

---

1John Tukey (1915 - 2000) had a formal background in chemistry and mathematics. Conducting data analysis during World War II peaked his interest in statistics and he became one of the most influential statisticians of the 20th century.
10.1 Introduction

What do stock prices and global warming have in common? Superficially, the two seem unrelated, but if you take a more abstract, statistician’s-eye view, you’ll find that two important common features stand out. First: Both the stock market and climate change involve time series - stock prices and environmental temperatures are numerical measures that change as time passes. This feature might be important in a different context, but here, it is largely coincidental. Second, and more important for our purposes: Both stock prices and temperatures lend themselves to multiple indirect measurements. Stock prices are measured by the sales records of individual stocks (Google, GM, etc.) and by indices like “the Dow” (Dow Jones Industrial Average), “the S&P” (Standard and Poor’s 500), “the Nasdaq” (National Association of Securities Dealers Automated Quotations), etc. For the study of global warming, temperatures can be measured directly at any of hundreds of weather stations around the world. Historical temperatures can also be measured indirectly through things like ice cores, tree rings, and coral growth patterns.

For both situations - market and climate - we have a set of many measurements that we can regard as “trying to tell us about” a single underlying “true”, albeit imaginary, number representing either asset value or global temperature. Large data sets (i.e. data sets with a large number of variables) can be difficult to interpret and it is likely that several of the variables are correlated. In this chapter we will discuss using principal component analysis to draw important information from large data sets. Principal component analysis (PCA) uses the variability (i.e. spread) of each variable and the correlation between variables to create a simplified set of new variables (components) that consist of uncorrelated, linear combinations of the original variables. The goal of PCA is to reduce the number of variables in a data set and still account for as much of the total variation in the original data as possible.

In this chapter, we will introduce PCA by combining stock market variables into an overall stock market variable. In the corresponding project, we will transform several indirect temperature variables into one principal component that can be used to measure change in global surface temperatures.

10.2 A Visual Interpretation of PCA

Scatterplots

The stock market example provided in the following sections contain only a few variables. Low-dimensional data (i.e. data with only a few variables) typically would not need to be reduced. However this small data set was specifically chosen to simplify the visual interpretation and calculations. Figure 10.1 shows a three-dimensional scatterplot of the Dow, S&P, and Nasdaq values. Multiple two-dimensional scatterplots as shown in Figure 10.2 are also
often created.

Figure 10.1: Three-Dimensional Scatterplot of the 2006 Dow, S&P, and Nasdaq Closing Values.

Figure 10.2: Matrix Plot of the 2006 Dow, S&P, and Nasdaq Closing Values.

Figures 10.1 and 10.2 show that each of the stock market indices are highly correlated. Since these variables all share a similar pattern, a single variable may be able to encapsulate most
of the information contained in these three variables. Principal component analysis is used to create linear combinations of the original data that capture as much information in the original data as possible.

Each of the two-dimensional scatterplots shown in Figure 10.2 can be thought of as a projection (or a shadow) of the multivariate data onto a two-dimensional space. In other words, the graph in Figure 10.1 can be rotated to look like each of the two-dimensional scatterplots. While the graphs shown in Figures 10.1 and 10.2 are useful, there may be another rotation of the data that would more clearly show patterns within the data. As the number of variables (i.e. the number of dimensions) increase, it becomes more and more challenging to ensure that the graphs are rotated in a way that best allows the researcher to visualize meaningful features within the data. Later sections will show that principal component analysis creates linear combinations that represent rotations of the original data onto a new axis that emphasizes patterns in the data.

[Note: Only one principal component is needed in the stock market example to summarize most of the information contained in the original data. However, in many studies more than one principal component is useful. If there are \( k \) variables in a data set, it is possible to create up to \( k \) principal components.]

[Key Concept] Principal component analysis is used to create linear combinations of the original data that may summarize much of the information contained in the original data set. These linear combinations represent a rotation onto a new axis that best reveal patterns or structures within the data.

### Time Series Plots

The stock market data set also contains dates for each business year in 2006. The following activity uses time series plots to visually compare two stock market variables, the Dow and the S&P, to the first principal component. Later sections will demonstrate how to calculate principal components.

**Activity: Visualizing the Data**

1. Open the stock market data file called `2006Stocks`. Create time series plots of the Dow and S&P. Both time series should be on the same graph.

   Because the Dow and S&P 500 use very different scales, the first step in simplifying the data is to standardize each variable.

2. Standardize the Dow and S&P columns, (i.e. for each element in the Dow column, subtract the Dow mean and divide by the Dow standard deviation. Save the standardized Dow values in a new column labeled \( Z_1 \). Repeat this process for the S&P 500 column and store the standardized data in \( Z_2 \). You may choose to use different labels,
but for the remainder of this chapter, $Z_1$ and $Z_2$ will refer to the standardized Dow and S&P, respectively.

(a) Create time series plots of the standardized Dow and S&P, $Z_1$ and $Z_2$. Both time series should be on the same graph.

(b) Do you see similar patterns in $Z_1$ and $Z_2$ (i.e. are the variables correlated)? Describe why you would or would not expect stock market indices to be correlated.

(c) Explain why the time series plot using $Z_1$ and $Z_2$ is more useful to compare patterns in the stock market than the time series plot created in question 1.

(3) Later sections will show that the first principal component is calculated as $PC1 = Y_1 = 0.707Z_1 + 0.707Z_2$.

(a) Use software to calculate PC1 and submit a time series plot of $Z_1, Z_2$ and PC1. All three series should be on one plot.

(b) Describe the relationship between PC1 and the standardized stock market data.

Figure 10.3 provides a time series plot similar to the one calculated in question 3a, except this graph is based on three standardized stock market indices; the 2006 Dow, S&P, and Nasdaq ($Z_3$) closing values. Notice that the first principal component based on these three variables provides similar information (shows a similar pattern) as the original data. Instead of using all three variables, one term ($PC1 = Y_1 = 0.582Z_1 + 0.608Z_2 + 0.540Z_3$) can be used as a simplified overall measure of patterns in the 2006 stock market values.

Notice in Figure 10.3 that PC1 is not simply an average of all three stock market variables, but emphasizes the key patterns in the data. For example, when all three stock values are increasing, PC1 increases at a faster rate than the other values. Similarly if all three values are simultaneously decreasing, PC1 decreases more quickly than the other terms.

The pattern in the stock market example is quite easy to see without PC1 since there are only three variables and they are highly correlated. When there are more variables, principal components are much more useful. The following activity shows that various linear combinations can be created to emphasize different characteristics of the data.

**Activity: Interpreting Linear Combinations**

(4) Each of the following linear combinations creates a new variable $C_i$. For each equation listed below, create one time series plot similar to Figure 10.3 that includes the standardized Dow, S&P, and Nasdaq closing values ($Z_1, Z_2, Z_3$) and $C_i$. Then explain what characteristic of the data is emphasized by that linear combination.

(a) $C_1 = 1Z_1 + 0Z_2 + 0Z_3$
Figure 10.3: Time series plot of the standardized Dow, S&P, and Nasdaq closing values ($Z_1$, $Z_2$, and $Z_3$) and the first principal component PC1.

(b) $C_2 = 1Z_1 - 1Z_2 + 0Z_3$

(c) $C_3 = \frac{1}{3}Z_1 + \frac{1}{3}Z_2 + \frac{1}{3}Z_3$

(d) $C_4 = 0Z_1 + 0Z_2 - 1Z_3$

10.3 Calculating Principal Components for Two Variables

Vector Notation

When researchers are working with data sets that have many variables, they often use techniques based on matrix algebra. While this chapter does not require a prior knowledge of matrix algebra, we will introduce some terms often used in this type of analysis.
The first five rows of the data file labeled 2006Stocks are shown below. In this file there are four variables: Dow, S&P, Nasdaq, and Date. The first column, Dow (also called vector $X_1$) gives the closing Dow stock market values for every business day in 2006. The second column, vector $X_2$, gives the closing S&P 500 values for every business day in 2006.

<table>
<thead>
<tr>
<th>Dow</th>
<th>S&amp;P</th>
<th>Nasdaq</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>10847.4</td>
<td>1268.8</td>
<td>2243.74</td>
<td>1/3/2006</td>
</tr>
<tr>
<td>10880.2</td>
<td>1273.46</td>
<td>2263.46</td>
<td>1/4/2006</td>
</tr>
<tr>
<td>10882.2</td>
<td>1273.48</td>
<td>2276.87</td>
<td>1/5/2006</td>
</tr>
<tr>
<td>10959.3</td>
<td>1285.45</td>
<td>2305.62</td>
<td>1/6/2006</td>
</tr>
<tr>
<td>11011.9</td>
<td>1290.15</td>
<td>2318.69</td>
<td>1/9/2006</td>
</tr>
</tbody>
</table>

To keep the calculations simple, this section will only use the first two columns of data. Thus, throughout this section the first two columns of this data set (not including labels) will be called matrix $X$. Each column will be treated as a vector $X_i$ that represents one variable of interest and consists of the $n$ observations for that variable. In this example $n = 251$ which was the number of business days in 2006 and $k = 2$ is the number of variables in our data set. Matrix $X$ consists of 251 rows and 2 columns and thus is called a “251 by 2” matrix.

Creating a Correlation Matrix

The strong linear relationship between the Dow and S&P in figure 10.4 shows that the two variables are highly correlated. Often the letter $r$ is used to represent the sample correlation between two variables. Subscripts are often added to represent the correlation between specific pairs of variables for data sets with more than two variables. For example, $r_{12}$ is the sample correlation between data vectors $X_1$ and $X_2$.

[Note: A matrix is a rectangular array of numbers; rows of a matrix are row vectors and columns of a matrix are column vectors. In this chapter each row represents a business day and each column will represent a variable of interest. The extended activities provide more details on notation and mathematical calculations involving matrices.]

When you have several variables it is beneficial to summarize their relationships with a correlation matrix $R$. This matrix is always square, with the number of rows and the
number of columns equal to the number of variables in the data set. In addition, this matrix
will always have ones on the diagonal and will be symmetric above and below the diagonal.

\[
\text{Corr}(X) = R = \begin{bmatrix}
    r_{11} & r_{12} \\
    r_{21} & r_{22}
\end{bmatrix} = \begin{bmatrix}
    1 & r_{12} \\
    r_{12} & 1
\end{bmatrix}
\]

(10.3.1)

**Activity: Calculating the Correlation Matrix**

(5) Calculate the 2 by 2 correlation matrix for the Dow and S&P variables.

(6) Calculate the 2 by 2 correlation matrix for the Dow and Nasdaq variables.

[Key Concept] A common first step in exploring a data set with many variables is
to check for dependencies among the variables. One way to do this is to calculate
a correlation matrix of all variables to identify any strong linear relationships.

The next section will discuss how a correlation matrix, \( R \), can be used to create a single
vector that will best explain the “direction” of the data (i.e. the pattern best representing
the relationships between all the variables).
Finding the Direction of Largest Variability

The direction of most variability (most spread of the data) is easy to visualize in two dimensions. Figure 10.5 is a plot of the standardized Dow and S&P data, $Z_1$ and $Z_2$. An oval has been drawn around the data. The direction of the most variability (where the oval is the most spread out) is shown as a vector, $v_1$.

Figure 10.5: Standardized daily closing values for the Dow and S&P 500 indices. The first eigenvector $v_1$, is drawn from the origin (0, 0) to the point (0.707, 0.707) and represents the direction of the most variation in the data.

The vector, $v_1$, is called the first eigenvector of the correlation matrix, $R$. While it can be difficult to visualize multidimensional data, the first eigenvector of the correlation matrix can always be considered the direction of the most spread in a multidimensional cluster of data points.

Figure 10.5 also displays a second eigenvector, $v_2$, that is perpendicular to $v_1$ and, expresses the direction of the second largest amount of variability. $v_2$ is drawn from the origin (0, 0) to the point (0.707, −0.707).

[Note: The definition of an eigenvector (sometimes called the characteristic vector) of the correlation matrix, $R$, is any vector, $v$, that is parallel to $Rv$. $Rv$ and $v$ are parallel if one is a constant multiple of the other. Thus the process of finding an eigenvector consists of finding any $v$ that satisfies the equation $Rv = \lambda v$ where $\lambda$ is a constant. In essence this...}
means that multiplying the vector \( \mathbf{v} \) by the matrix \( \mathbf{R} \) may stretch or shrink the vector, but the direction of vector \( \mathbf{v} \) is the same as \( \mathbf{Rv} \).

A technique using matrix algebra can be used to find the eigenvectors of the correlation matrix, \( \mathbf{R} \). Computer software is typically used to calculate eigenvectors. However, since this example consists of a very small data set, the next activity allows you to find eigenvectors by hand and by using software (the extended activities provide more advanced mathematical details).

**Activity: Calculating Eigenvectors and Eigenvalues**

Calculating the first eigenvector of the correlation matrix \( \mathbf{R} \), involves solving the equation \( \mathbf{Rv} = \lambda \mathbf{v} \). Using matrix algebra, it is possible to show:

\[
\mathbf{Rv} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} + rv_{12} \\ rv_{11} + v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} = \lambda_1 \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \lambda_1 \mathbf{v}.
\]

(7) *Algebra Required:* The above equation can be written as a system of two equations:

\[
\begin{align*}
&v_{11} + rv_{12} = \lambda_1 v_{11} \\
&rv_{11} + v_{12} = \lambda_1 v_{12}
\end{align*}
\]

where \( r = r_{12} = 0.97135 \).

Find the first eigenvector, \( \mathbf{v}_1 \), by adding the two equations and simplifying to show that \( \lambda_1 = 1 + r = 1.97135 \). Then substitute \( \lambda_1 = 1.97135 \) into either equation to show \( v_{11} = v_{12} \).

(8) *Algebra Required:* Find the second eigenvector by solving the equation \( \mathbf{Rv} = \lambda \mathbf{v} \). This can also be written as a system of two equations:

\[
\begin{align*}
&v_{21} + 0.97135v_{22} = \lambda_1 v_{21} \\
&0.97135v_{21} + v_{22} = \lambda_1 v_{22}
\end{align*}
\]

The second eigenvector is found by subtracting \( 0.97135v_{21} + v_{22} = \lambda_2 v_{22} \) from \( v_{21} + 0.97135v_{21} = \lambda_2 v_{21} \) to solve for \( \lambda_2 \) and then showing that \( v_{21} = -v_{22} \).

(9) Use software to calculate the eigenvectors and eigenvalues of \( \mathbf{R} \) with software.

In the previous questions where calculations were done by hand, the solution to \( \mathbf{v}_1 \) was not exact. The first eigenvector, \( \mathbf{v}_1 \), can be any vector with two elements \( (v_{11} \text{ and } v_{12}) \) where \( v_{11} = v_{12} \). Figure 10.5 provides an intuitive explanation. Any vector starting at the point \((0, 0)\) and going to any point \((x, x)\) would represent the direction of most variability.

The specific values of each eigenvector are calculated so that the sum of the squared elements equal one. This normalization is done so that the total variability of the principal components is equal to the total variability of the original (or standardized) data. In this example,

\[
\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} \\ \sqrt{1/2} \end{bmatrix}.
\]
10.3. **CALCULATING PRINCIPAL COMPONENTS**

Thus the sum of the squared elements, \( v_{11}^2 + v_{12}^2 = 0.707^2 + 0.707^2 = 1 \). Similarly,

\[
v_2 = \begin{bmatrix}
v_{21} \\ v_{22}
\end{bmatrix} = \begin{bmatrix}
0.707 \\ -0.707
\end{bmatrix} = \begin{bmatrix}
\sqrt{1/2} \\ -\sqrt{1/2}
\end{bmatrix}
\]

and \( v_{21}^2 + v_{22}^2 = 0.707^2 + -0.707^2 = 1 \).

**[Key Concept]** The elements of each eigenvector are normalized so that their squared values sum to one. This normalization is done so that the total variability of the principal components is equal to the total variability of the original (or standardized) data.

The sign of the eigenvectors is arbitrary. In Figure 10.5 the vector, \( v_1 \), could start at the origin, \((0, 0)\) and go to \((0.707, 0.707)\) or \( v_1 \) could start at \((0, 0)\) and go to \((-0.707, -0.707)\). In either case, \( v_1 \) is in the direction of the most variability in the data.

\[
v_1 = \begin{bmatrix}
v_{11} \\ v_{12}
\end{bmatrix} = \begin{bmatrix}
0.707 \\ 0.707
\end{bmatrix} \quad \text{or} \quad v_1 = \begin{bmatrix}
-0.707 \\ -0.707
\end{bmatrix}.
\]

\[
v_2 = \begin{bmatrix}
v_{21} \\ v_{22}
\end{bmatrix} = \begin{bmatrix}
0.707 \\ -0.707
\end{bmatrix} \quad \text{or} \quad v_2 = \begin{bmatrix}
-0.707 \\ 0.707
\end{bmatrix}.
\]

**[Note:** For a data set with \( k \) variables and \( n \) observations, the correlation matrix will be a symmetric matrix with \( k \) rows and \( k \) columns. As long as none of the \( k \) variables are perfectly correlated (i.e., no variable is an exact linear combination of the other variables), the correlation matrix can be used to calculate \( k \) uncorrelated eigenvectors and corresponding eigenvalues.]

The previous questions also calculated constant values \( \lambda_1 \) and \( \lambda_2 \) called **eigenvalues**. Matrix algebra can be used to show that each eigenvalue is equal to the variance of the corresponding principal component. The eigenvectors are sorted according to the size of their eigenvalues. The eigenvector corresponding to the largest eigenvalue is considered the “first” eigenvector, \( v_1 \). The second largest eigenvalue determines the “second” eigenvector, \( v_2 \), and so on.

**Creating Principal Components**

To calculate principal components, multiply the standardized data by each eigenvector using the following formula:

\[
Y_i = Zv_i = v_{i1}Z_1 + v_{i2}Z_2 + \ldots + v_{ik}Z_k \quad i = 1, \ldots, k
\]

(10.3.2)

In our example there are only two columns of data, thus \( k = 2 \). This will create two new variables, the first principle component, \( PC1 = Y_1 = v_{11}Z_1 + v_{12}Z_2 = 0.707Z_1 + 0.707Z_2 \), and the second principle component \( PC2 = Y_2 = v_{21}Z_1 + v_{22}Z_2 = 0.707Z_1 + -0.707Z_2 \).
Since the eigenvectors are ordered according to the size or their corresponding eigenvalues, \( \mathbf{Y}_1 \) will be the linear combination of the variables with the most spread. \( \mathbf{Y}_2 \) will be uncorrelated to \( \mathbf{Y}_1 \) and will include the rest of the total variability.

Figure 10.6 is a scatterplot of PC1 and PC2. In essence principle components provide weights that rotate the data in figure 10.5 onto a new axis. Now the spread of the data is represented along PC1, the new horizontal axis. The second component, PC2, is along the vertical axis explains the rest of the variability in the data.

![Principal Components of the Stock Index Values](image)

Figure 10.6: Principal components of the daily closing values for the Dow and S&P 500 indices.

[Key Concept] Principle components are found by using eigenvectors of the correlation matrix to weight the original variables. These weights rotate the data along a new axis representing the direction of the largest amount of the variation in the data.

Activity: Calculating Principal Components

(10) Use software to calculate the principle components PC1 = \( \mathbf{Y}_1 = v_{11} \mathbf{Z}_1 + v_{12} \mathbf{Z}_2 \), and PC2 = \( \mathbf{Y}_2 = v_{21} \mathbf{Z}_1 + v_{22} \mathbf{Z}_2 \).

(a) What is the correlation between PC1 and PC2?
(b) Calculate the variance of PC1 and PC2 and explain how the variances of the principal components are related to the eigenvalues. Note that the sum of these two variances is equal to the variance of $Z_1$ plus the variance of $Z_2$. The eigenvalues were normalized so that the sum of the variances of all the principle components will be equal to the sum of the variances of the original standardized data.

(11) Create a time series plot. Submit a time series plot of $Z_1$, $Z_2$, PC1, and PC2. All four series should be on one plot. Your final plot should be similar to figure 10.7. Explain how PC1 and PC2 are related to $Z_1$ and $Z_2$.

![Time Series Plot](image)

Figure 10.7: Principal components of the daily closing values for the Dow and S&P 500 indices.

Notice in figure 10.7 that PC1 has much more variability than PC2. PC1 follows a similar pattern as $Z_1$ and $Z_2$. PC1 also increases at a faster rate when both $Z_1$ and $Z_2$ are increasing. In addition when PC1 increases at a faster rate than $Z_1$ and $Z_2$, PC2 somewhat compensates for this by decreasing at those times. Since PC1 explains most of the variation in this example, there is no need to use PC2. Thus PC1 can be used to reduce the two dimensional data set to just one dimension.
10.4 Determining the Number of Principle Components to Use

In earlier problems eigenvectors and eigenvalues for the correlation matrix were found simultaneously. Corresponding to each eigenvector, \( v_i \), \( i = 1, 2, ..., k \), there is an eigenvalue, \( \lambda_i \). As discussed in question 10b, the eigenvalue, \( \lambda_i \), is equal to the variance of the corresponding principal component. In other words, \( \lambda_1 \) is equal to the variance of PC1.

Principal components are ordered according to their variances. The first principal component is the linear combination that encapsulates most of the variability. In other words, the first principal component represents a rotation of the data along the axis representing the largest spread in the multidimensional cluster of data points. The second principal component is the linear combination that explains the most of the remaining variability while being uncorrelated (i.e. perpendicular) to the first principal component. If there was a third principal component, it would explains most of the remaining variability while being uncorrelated to the first two principal components. This pattern continues for all consecutive principal components.

[Key Concept]Principal components are ordered according to their variances. Thus the principal component with the largest corresponding eigenvalue is called the first principal component. The principal component with the second largest variance is called the second principal component. This process continues for all principal components calculated from our data set of interest.

The proportion of the total variation that is explained by the first principle component can be found by dividing the first eigenvalue by the sum of all eigenvalues (only 2 in this example). In this example, the percentage of the variation is explained by the first principle is:

\[
\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.9714}{1.9714 + 0.0286} = 0.986
\]

Similarly, we can show that only 1.4% of the variation is explained by the second principle component. Since PC1 explains over 98% of the variation in this example, there is no need to use PC2. Thus PC1 can be used to reduce the two dimensional data set to just one dimension.

10.5 A Three-Dimensional Example

The file 2006Stocks contains the daily closing stock prices of the Dow, S&P 500, and Nasdaq financial indices for all business days in 2006. In this data set \( n = 251 \), which is the number of business days in 2006 and \( k = 3 \) represents the three variables of interest. This section briefly conducts a principal component analysis on three variables in order to show how this technique can be extended to more than just two variables. The same steps will be used as
in the previous section however the sizes of the matrices and number of vectors will increase.

(1) Calculate the correlation matrix, $R$. The correlation between each pair of variables is calculated. For example the correlation between the Dow and Nasdaq, $r_{13} = 0.683$.

\[
\text{Corr}(X) = R = \begin{bmatrix}
1 & r_{12} & r_{13} \\
 r_{21} & 1 & r_{23} \\
 r_{31} & r_{32} & 1 \\
\end{bmatrix}
= \begin{bmatrix}
1 & 0.971 & 0.683 \\
0.971 & 1 & 0.809 \\
0.683 & 0.809 & 1 \\
\end{bmatrix}
\]

(10.5.1)

The correlation matrix shows us that all three variables are positively correlated. In addition, the strongest linear relationship is between the Dow and S&P while the Dow and Nasdaq have the weakest linear relationship. Figures 10.1 and 10.2 agree with the information found with the correlation matrix.

(2) Calculate the $k = 3$ eigenvectors and eigenvalues of $R$. The first eigenvector provides a vector in the direction of the largest variability in the cluster of 3-dimensional data points. Each eigenvector and eigenvalue are calculated by solving the following equation:

\[
Rv_1 = \begin{bmatrix}
1 & 0.971 & 0.683 \\
0.971 & 1 & 0.809 \\
0.683 & 0.809 & 1 \\
\end{bmatrix}
\begin{bmatrix}
v_{11} \\
v_{12} \\
v_{13}
\end{bmatrix}
= \lambda_1
\begin{bmatrix}
v_{11} \\
v_{12} \\
v_{13}
\end{bmatrix}
= \lambda_1 v_1.
\]

The order of the eigenvectors are identified by their corresponding eigenvalues. The first eigenvector has the largest eigenvalue, the second eigenvector has the second largest eigenvalue, and so on.

(3) Use the eigenvectors to transform the standardized data into principal components. These principal components represent a rotation onto a new axis that best summarizes the information in the original data. The first principal component in this example is:

\[
PC1 = \mathbf{Z}v_i = \mathbf{v}_{i1}\mathbf{Z}_1 + \mathbf{v}_{i2}\mathbf{Z}_2 + \mathbf{v}_{i3}\mathbf{Z}_3
\]

\[= 0.582\mathbf{Z}_1 + 0.608\mathbf{Z}_2 + 0.540\mathbf{Z}_3\]

Where $\mathbf{Z}$ represents a 251 by 3 matrix containing vectors $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$, the standardized Dow, S&P, and Nasdaq values respectively.

**Activity: Calculating Principal Components for Three Dimensions**

(12) Standardize the Nasdaq column into a new column, $\mathbf{Z}_3$ and conduct a principal component analysis on all three variables.
(a) Calculate the eigenvectors and eigenvalues.
(b) Calculate the principal components.

(13) Create a 3-D plot of the data. Draw the first two eigenvectors onto the graph by hand or with computer software.

(14) Visualize the principal components in a time series plot. Remember that it might be more appropriate to plot a negative principal component (e.g. $-\text{PC1}$ instead of $\text{PC1}$) since the sign of each eigenvector is arbitrary. Submit this graph and explain how the time series pattern of PC1, PC2, and PC3 relate to $Z_1$, $Z_2$, and $Z_3$.

(15) What percentage of the variability is explained by the first principal component? What percentage of the variability is explained by the first two principal components?
While working with three variables is fairly manageable, the principal component analysis has shown that the first principal component can be used to reduce the data set to only one variable and still summarize over 88% of the variability.

**Conclusions**

Principal component analysis is used to re-express a large and complex data set so that only the first few variables (dimensions) account for as much of the variability as possible. In addition to reducing the number of variables, principal component analysis also creates uncorrelated variables.

It is important to recognize that PCA is scale sensitive. Notice that the Dow has a much larger magnitude and range than the S&P and Nasdaq. The extended activities show that if the data were not standardized before PCA was conducted (i.e. if the covariance matrix was used instead of the correlation matrix), the Dow would have much more influence than the other variables.

In this example, the first principal component assigns nearly equal weight to each stock market index. Thus the first principal component can roughly be thought of as an overall average stock market index value. When the when the stock market values than average, the first principal component will also be larger than average. When all the stock market values are smaller than average, the first principal component will be smaller than average. Caution should be used in interpreting the principal components. If PCA is repeated on a similar new data set, the coefficients (i.e. the linear combinations) would change. The extended activities discuss how to interpret principal components and how to determine the number of principal components to retain. PCA is most effective if just a few linear combinations of a large data set explain most of the variability.

In this example, three variables were combined into one overall stock market value. Instead of looking at three variables, we are now able to use just one variable to understand patterns in the stock market. The first eigenvalue tells us that using just one principal component instead of all three variables still explains 88.3% of the variability.
Extended Activities: A Closer Look at Statistical Models

10.6 Determining the Number of Components to Retain

In data sets with many variables, it is often appropriate to retain more than just one principal component. There is no exact test to determine how many principal components to retain, but there are several general guidelines based on understanding the variance in the data.

The variance of each standardized variable is equal to one (these ones are represented in the diagonal of the correlation matrix). Thus, the sum of the variances in the standardized data set is equal to \( k \), the number of variables in our data.

In addition, in principal component analysis each eigenvector is normalized. Thus the sum of all the variances of the principal components equals the sum of all the variances represented in our correlation matrix. Thus for any data set with \( k \) variables:

\[
    k = \text{variance}(Z_1) + \text{variance}(Z_2) + \ldots + \text{variance}(Z_k) \\
    = \text{variance}(PC_1) + \text{variance}(PC_2) + \ldots + \text{variance}(PC_k) \\
    = \lambda_1 + \lambda_2 + \ldots + \lambda_k
\]

where \( Z_i, i = 1, 2, \ldots, k \) represents a standardized variable and \( PC_i \) represents the \( i^{th} \) principal component.

Thus, when each variable is standardized, the percentage of the variation explained by the \( i^{th} \) principle component is found by dividing the \( i^{th} \) eigenvalue by \( k \):

\[
    \frac{\lambda_i}{k}, \quad i = 1, 2, \ldots, k
\]

(10.6.1)

[Note: In some cases researchers may choose to not standardize each variable. This is equivalent to using the variance-covariance matrix instead of the correlation matrix. Thus sum of the variances is no longer equal to \( k \). Even without standardization, it can be shown that the sum of the variances of each variable is equal to the sum of the variances of the principal components calculated with eigenvectors of the variance-covariance matrix. Later sections describe why standardization is typically advised.]

Extended Activity: Fisher’s Iris Data
Ronald Fisher published an analysis on the size of iris sepals and petals. This data, collected over several years by Edgar Anderson, was used to show that these measurements could be used to differentiate between species of irises. The dataset Versicolor contains four measurements (sepal length, sepal width, petal length, and petal width) on 50 versicolor irises. The petals are colorful and flashy while the sepal is under the petals and tends to be less colorful. Conduct a principal component analysis on the Versicolor data. Be sure to use the correlation matrix instead of the covariance matrix.

(a) List the four eigenvalues. How many eigenvalues are greater than one? Kaiser recommended retaining the principal components with corresponding eigenvalues greater than one. However, this should be a general rule of thumb, not an absolute rule.

(b) What is the percentage of the variation explained by the first principle component?

(c) What is the percentage of the variation explained by the first two principle components?

(d) Calculate a scree plot for this data.

The scree plot is a simple graphic used to display the eigenvalues for each successive principal component. The general rule is to look for a change in slope. The principal components corresponding to a steep curve at the beginning of the plot should be kept. Components where the slope appears flat are not retained. In figure 10.8, it is clear that the first eigenvalue corresponds to a very large proportion of the overall variance \( \frac{2.926}{4} = 0.732 \). Thus, only the first principal component should be used. Not all scree plots are this clear. Often the scree plot looks more like a smooth curve and the determination of how many components to include is more challenging.

Others suggest that the number of components should be determined by the cumulative percent variation. However, the acceptable percentage of variation that is explained depends on the context of the problem. Only 80% may be needed for descriptive purposes. However, a larger percentage may be needed if further analysis is to be done.

**Caution:** While most of the variability is explained by the first few principal components, principal components with smaller eigenvalues may also be important. For example a few principal component are often calculated and then used in a regression analysis. Researchers should be aware that a meaningful linear combination, or single explanatory variable, may be omitted that would be highly correlated with the response variable.

**Caution:** Eigenvalues very close to zero should not automatically be ignored. When eigenvalues are equal to zero, one or more variables are redundant (perfect linear combinations

---


### 10.7 Interpreting Principal Components

The four eigenvectors for the correlation matrix for the *Versicolor* iris data set are given below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal length</td>
<td>0.482</td>
<td>0.611</td>
<td>-0.491</td>
<td>0.392</td>
</tr>
<tr>
<td>Sepal width</td>
<td>0.465</td>
<td>-0.673</td>
<td>-0.540</td>
<td>-0.199</td>
</tr>
<tr>
<td>Petal length</td>
<td>0.535</td>
<td>0.307</td>
<td>0.340</td>
<td>-0.710</td>
</tr>
<tr>
<td>Petal width</td>
<td>0.515</td>
<td>-0.283</td>
<td>0.593</td>
<td>0.550</td>
</tr>
</tbody>
</table>

The first eigenvector shows roughly equivalent coefficients. Thus the first principal component can be considered as an overall size measurement. When the iris has larger sepal and petal values than average, the first principal component will also be larger than average. When all the measurements are smaller than average, the first principal component will be smaller than average. If an iris has some original measurements larger than average and some smaller (or if all the iris’s measurements are near average) the first principal component will be

---

Figure 10.8: Scree Plot for the versicolor iris data. There is a clear change in slope after the second component, suggesting that only one component is needed.

of other variables). Rounding error may make the eigenvalues very small instead of zero. Ignoring redundant variables may cause problems in interpreting future analysis.
close to average. The first eigenvector shows us that this one principal component accounts for over 73% of the variability in the data.

Some researcher may also choose to retain PC2. The second eigenvector also has a nice intuitive interpretation. The second principal component will be large when the iris is wider than average, but has shorter lengths. In addition, the sepal has more influence than the petal in the value of PC2.

The third eigenvector shows that the third principal component is a measure of the difference between sepals and petals. PC3 will be large if the petals are larger than average while the sepals are smaller than average. However, the eigenvalue indicated that this principal component will only account for about 1% of the variation in the original data.

Principal component loadings and loading plots are often calculated to better understand how principal components are characterized by the original variables. Figure 10.9 provides a loading plot for the first two components. This plot simply helps to visualize the first and second eigenvectors. The x-axis represents the weights for PC1 and the y-axis represents the weights for PC2. Figure 10.9 shows that all weights (all elements of the eigenvector) on the first principal component are positive, while PC2 has positive weights for lengths but negative weights for widths.

![Loading Plot for the first two principal components for the versicolor iris data.](image)

**Figure 10.9:** Loading Plot for the first two principal components for the versicolor iris data. The Sepal width vector starts at the point (0, 0) and ends at (0.465, −0.673), corresponding to the weights provided in the first two eigenvalues.

**Principal component loadings** for the first two principal components are shown below. These loading are the correlation between the original data and the principal components.
10.8. EXTENDED ACTIVITIES: A CLOSER LOOK AT PCA

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal length</td>
<td>0.826 = corr(PC1, Z1)</td>
<td>0.451 = corr(PC2, Z1)</td>
</tr>
<tr>
<td>Sepal width</td>
<td>0.795 = corr(PC1, Z2)</td>
<td>-0.497 = corr(PC2, Z2)</td>
</tr>
<tr>
<td>Petal length</td>
<td>0.914 = corr(PC1, Z3)</td>
<td>0.227 = corr(PC2, Z3)</td>
</tr>
<tr>
<td>Petal width</td>
<td>0.882 = corr(PC1, Z4)</td>
<td>-0.209 = corr(PC2, Z4)</td>
</tr>
</tbody>
</table>

Principal component loadings show similar patterns as the eigenvectors, except they have the added benefit that they can be easily interpreted as correlations. The squared correlation (similar to $R^2$ in regression) represents the percent of variance in that variable explained by the principal component.

[Note: The correlations calculated as principal component loading are often called unrotated factor loadings. Factor analysis and PCA are very similar, however factor analysis makes different model assumption about underlying model structures. PCA attempts to summarize all variability, while factors analysis uses only the variability that is common with the other variables. While some researchers treat these methods as interchangeable, PCA is useful for data reduction, while factor analysis is useful in detecting structure within a data set.]

## 10.8 Incorporating Principal Components into Other Statistical Methods (Requires Multiple Regression)

You may have worked with a large number of variables before, such as in conducting multiple regression analysis. Unlike multiple regression, in PCA there is no response variable. Principal components is often a first step in a series of statistical analyses. PCA transform a large number of original variables to a smaller set of uncorrelated components that capture most of the variability in the data. Thus, PCA simplifies a more complex data set in order to make further analysis easier. In addition to reducing the number of variables, PCA is beneficial in multiple regression analysis because each component is uncorrelated with the others.

### Extended Activity: Using Principal Components in Regression

(17) The dataset cars contains the make, model, equipment, mileage and Kelley Blue Book suggested retail price of several used 2005 General Motor cars. Kelley Blue Book (www.kbb.com) has been an accurate resource for pricing cars for over 80 years and was used as a source to collect this data. In this activity you will create a regression model that will describe the association of several explanatory variables (car characteristics) to the retail value of a car.

(a) In the Cars data set, liters and cylinders are highly correlated as they both are a measure of engine size. Instead of using either the liter or cylinder variable,
create the first principal component obtained from a PCA of just these two variables. Use this first principal component, plus mileage, make, and model in a regression analysis to predict retail price.

(b) Run a regression analysis with liter, cylinder, mileage, make, and model in a regression analysis to predict retail price. In the previous question, using the first principle component from the PCA of liter and cylinder eliminated multicollinearity, but how did it impact the R-sq value? Did using PCA instead of all of the variables cause us to miss a key explanatory variable? Which model would you suggest as better?

PCA is often used to avoid multicollinearity among several explanatory variables in multiple regression. However, caution should be used. If only the first few principal components are included as variables in a model, an explanatory variable that is highly correlated with the response could easily be left out.

10.9 Comparing Regression and Principal Components

The direction of the first eigenvector, $v_1$, can be considered as the line of closest fit to all the data points. In other words, the line representing the most variation in the data is also the line closest to all points in the data set. This is different than a regression line, where distance is only measured in one direction.

A small data set based on tarantulas weights and heart rates is shown in figure 10.10 to emphasize the difference between regression and PCA. Carrel and Heathcoat collected data on the weight and resting heart rate of several species of spiders. They used the resting heart rate of spiders to calculate their metabolic rates and showed that some species of spiders have lower metabolic rates than others.

The lines created by multiple linear regression and the first principal component both minimize the sum of squared distances between each data point and a line. In multiple linear regression the distance between the points and the regression line are only measured in terms of the response variable. The distance between the response and the predicted response is based on one dimension.

PCA focuses on finding a line closest to all the points in multiple dimensions. The line of interest in PC1 is represented by the first eigenvector, $v_1$. Algebra can be used to show that the shortest distance between a point to a line is always perpendicular to the line. This distance is often measured in multiple dimensions. If the two dimensional data set in figure

---

Figure 10.10: Comparing Regression and PC1 for Standardized Tarantula Weights and Heart Rates. The least squares regression line minimizes the difference (the vertical distance) between the observed weight and the predicted weight. PC1 minimizes the difference between each point and the line in two-dimensional space.

10.10 could only be seen in one dimension (a single line), the most informative line (the line closest to all points) would be along the direction of the first eigenvector, $v_1$.

### 10.10 The Impacts of Standardizing Each Variable

In the previous examples the correlation matrix was used. However there are times when it is appropriate to use the covariance matrix instead of the correlation matrix (i.e the unstandardized data is used instead of the standardized data). Principal components are very sensitive to differences in scale.

[Note: The variance-covariance matrix of $X$ (often called the covariance matrix) is a $p$ by $p$ symmetric matrix and is typically shown as:

$$
\text{Var}(X) = \begin{bmatrix}
S_1^2 & S_{12} & \cdots & S_{1p} \\
S_{12} & S_2^2 & \cdots & S_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
S_{1p} & S_{2p} & \cdots & S_p^2
\end{bmatrix}.
$$

(10.10.1)

where the diagonal contains the variances of each of the data vectors (e.g. $\text{var}(X_1) = S_1^2$ and $\text{var}(X_2) = S_2^2$, etc...) and $S_{ij}$ represents the covariance of vectors for any two vectors, $X_i$]
and \( X_j \). Appendix D uses matrix algebra to show the correlation matrix is the covariance matrix of the standardized data. In the diagonal of the correlation matrix \( 1 = \frac{s_{1j}^2}{s_{11}s_{11}} \) and \( r_{ij} = \frac{s_{ij}}{s_is_j} \).

**Extended Activity: Using the Covariance Matrix in PCA**

(18) Calculate the three eigenvectors and eigenvalues of the unstandardized 2006Stock data (i.e., use the covariance matrix instead of the correlation matrix).

(a) How do the eigenvector weights change when the data is not standardized?
(b) Provide an interpretation of the first eigenvalue. In other words, does the first eigenvalue using the unstandardized data indicate PC1 will explain more or less of the variability than when the data was standardized?
(c) Calculate the covariance matrix of \( X \) and compare the variances of the original data \( \text{var}(X_1), \text{var}(X_3), \) and \( \text{var}(X_3) \) to the variances of the principal components \( \text{var}(PC1), \text{var}(PC2), \) and \( \text{var}(PC3) \).

(19) Create a time series plot of the original data \( (X_1, X_2, \) and \( X_3) \) and the first principal component \( (PC1) \). All four series should be on one plot. Describe the relationship between PC1 and the stock market data.

(20) Assume that the Nasdaq values are measured on a different scale. The Nasdaq is actually a composite measure of over 5000 different stocks. Multiply the Nasdaq values by 5000 and label the new variable NewNasdaq.

(a) Conduct PCA using the covariance matrix on the Dow, S&P, and NewNasdaq. How does changing the scale influence the principal components?
(b) Conduct PCA using the correlation matrix on the Dow, S&P, and NewNasdaq. How does changing the scale influence the principal components?

Using the correlation matrix ensures that each variable is centered at zero and has a variance equal to one. The eigenvalues and hence the eigenvectors clearly depend on whether the correlation or the covariance matrix is used. The variables with the largest variances will have the most influence on the principal components.

[Note: The dependence of the principal components on the size of the variance can be explained geometrically. Figure 10.10 shows that PCA minimizes that distance between each point and the first principal component. The minimum distance is measured from a point to PC1 by a line perpendicular to PC1. If the original points are plotted instead of the standardized points, the perpendicular lines (and thus the sum of squared distances) also change.]
[Key Concept] Principal components are heavily influenced by the magnitude of the variances. If the covariance matrix is used the principal components will tend to follow the direction of the variables that have the largest variability. If the correlation matrix is used principal components represent the strongest patterns of intercorrelation within the data.

There are situations where it may be appropriate to use the covariance matrix. For example, assume several students are given a 20 question survey that attempts to measure their attitudes towards statistics. For each question students select one of five responses (1= “Strongly Disagree” to 5 = “Strongly Agree”). All questions (all 20 variables) are original measure on the same scale. However there may be a few questions that are confusing or unclear and most students give the question a value of 3. The researcher may want these unclear questions to have less influence on the principal components than questions where students provided stronger responses (i.e. students tended to respond with a 1 or 5). Using the covariance matrix would put less emphasis on the questions that had small variances (i.e. all students answered the question similarly) and place more emphasis on the questions with more distinct responses (i.e. larger variances). Caution should be used, however, since even small differences in variances can dramatically influence the principal components and the corresponding loadings and the interpretations are typically very different.

[Key Concept] The covariance matrix can be used in PCA if all the original variables are measured on a similar scale and there is reason to believe that the magnitude of the variance should influence our interpretation of the data.

10.11 Calculating Eigenvectors and Eigenvalues Using Matrix Algebra  (Matrix Algebra Required)

This section provides detailed mathematical calculations of principal component analysis. Appendix D provides a brief review of matrix calculations, including the calculation of variance-covariance and correlation matrices.

A scalar quantity $\lambda$ is an eigenvalue (also called a characteristic root) of a matrix $A$ if there is a nonzero vector $v$ that satisfies the equation $Av = \lambda v$. Such a vector $v$ is called an eigenvector of $A$ corresponding to $\lambda$.

For invertible $k \times k$ matrices, the $k$ eigenvalues are distinct and the $k$ eigenvectors are orthogonal, which will make them more applicable (and interpretable) for statistical methods. Covariance and correlation matrices for data collected on $k$ variables are invertible square matrices as long as each random variable is not a linear combination of the other variables.

Calculating the Eigenvalues and Eigenvectors for a $2 \times 2$ Matrix
Let \( A \) be any \( 2 \times 2 \) matrix \( A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \).

The two eigenvalues and their corresponding eigenvectors of \( A \) must satisfy:

\[ A \mathbf{v} = \lambda \mathbf{v} \]

or equivalently,

\[ A \mathbf{x} - \lambda \mathbf{x} = [A - \mathbf{I} \lambda] \mathbf{x} = \mathbf{0} \]

where \( \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \) is the \( 2 \times 2 \) identity matrix, \( \mathbf{v} \) is any \( 2 \times 1 \) column vector, and \( \mathbf{0} \) is the \( 2 \times 1 \) vector of zeros.

To solve this equation for \( \lambda \), we write out the details:

\[
\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \lambda = 0
\]

\[
\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.
\]

This matrix equation can be written as a system of two equations:

\[
(a - \lambda)v + bv_2 = 0 \\
cv_1 + (d - \lambda)v_2 = 0.
\]

This system of equations can be solved using techniques similar to those in question 7. However, instead of using the equation \( A \mathbf{v} = \lambda \mathbf{v} \) to solve for the eigenvalues, an equivalent equation using the determinant \( |A - (I)\lambda| = 0 \) is often used.

The determinant of any \( 2 \times 2 \) matrix \( A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \), written \( |A| = ad - bc \).

The solution to \( |A - (I)\lambda| = (a - \lambda)(d - \lambda) - bc = 0 \) is given as:

\[
\lambda = \frac{1}{2} \left[ (a + d) \pm \sqrt{(a - d)^2 + 4bc} \right].
\]

Thus, the two eigenvalues are for any \( 2 \times 2 \) matrix \( A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \) are:

\[
\lambda_1 = \frac{1}{2} \left[ (a + d) + \sqrt{(a - d)^2 + 4bc} \right] \quad \text{and} \quad \lambda_2 = \frac{1}{2} \left[ (a + d) - \sqrt{(a - d)^2 + 4bc} \right]
\]
Once the eigenvalues are found, the corresponding eigenvectors can be calculated. The eigenvectors of $\mathbf{A}$, $\mathbf{v}_1 = \begin{bmatrix} v_{11} & v_{12} \end{bmatrix}^T$ and $\mathbf{v}_2 = \begin{bmatrix} v_{21} & v_{22} \end{bmatrix}^T$ satisfy the equations

$$[\mathbf{A} - \mathbf{I}\lambda_1] \mathbf{v}_1 = 0 \quad \text{and} \quad [\mathbf{A} - \mathbf{I}\lambda_2] \mathbf{v}_2 = 0$$

Typically, each eigenvector is normalized so that it has length one. That is, we set the eigenvector, $\mathbf{v}_i$, corresponding to $\lambda_i$ such that $\mathbf{v}_i^T \mathbf{v}_i = v_{i1}^2 + v_{i2}^2 = 1$.

**Eigenvalues and Eigenvectors for a $2 \times 2$ Correlation Matrix**

For any two variables from a random sample of $n$ observational units: $\mathbf{X}_1 = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \end{bmatrix}^T$ and $\mathbf{X}_2 = \begin{bmatrix} X_{21} & X_{22} & \cdots & X_{2n} \end{bmatrix}^T$, two standardized vectors can be calculated:

$$\mathbf{Z}_1 = \frac{\mathbf{X}_1 - \overline{X}_1}{S_1} \quad \text{and} \quad \mathbf{Z}_2 = \frac{\mathbf{X}_2 - \overline{X}_2}{S_2}.$$  

Then, it can be shown that $\mathbf{R} = \text{cov}(\mathbf{Z}_1, \mathbf{Z}_2) = \text{corr}(\mathbf{Z}_1, \mathbf{Z}_2) = \text{corr}(\mathbf{X}_1, \mathbf{X}_2) = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ where “cov” represents the variance-covariance matrix, “corr” represents the correlation matrix, and $r$ is the sample correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$.

**Extended Activity: Calculating Eigenvectors and Eigenvalues**

(21) Show that the eigenvalues for any $2 \times 2$ correlation matrix are $\lambda_1 = 1 + r$ and $\lambda_2 = 1 - r$.

(22) Show that the eigenvectors for any $2 \times 2$ correlation matrix are $\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T$ and similarly, $\mathbf{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}^T$.

**Calculating Principal Components**

Principal components are linear combinations of the the standardized data vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$:

$$\mathbf{Y}_1 = PC_1 = \mathbf{v}_1^T \mathbf{Z} = v_{11} \mathbf{Z}_1 + v_{12} \mathbf{Z}_2 \quad \text{and} \quad \mathbf{Y}_2 = PC_2 = \mathbf{v}_2^T \mathbf{Z} = v_{21} \mathbf{Z}_1 + v_{22} \mathbf{Z}_2.$$  

These new variables are just linear combinations of the original data, using the eigenvectors as the weights.
If the correlation is positive ($r > 0$), then the largest eigenvalue is $\lambda_1 = 1 + r$ and the first principal component $Y_1 = (Z_1 + Z_2)/\sqrt{2}$. 

If the correlation is negative, the larger eigenvalue is $\lambda_2 = 1 - r$ and the “first” principal component $Y_2 = (Z_1 - Z_2)/\sqrt{2}$.

**Extended Activity: Understanding the Variability of Principal Components**

(23) Show that $\text{var}(Y_1) = \lambda_1$.

(24) Show that $\lambda_1 + \lambda_2 + \ldots + \lambda_k = k$ where $k$ is the number of variables in the original data set.

(25) Show that $Y_1$ is that linear combination of the original standardized data that has the most variance. That is, show that $\text{var}(w_1Z_1 + w_2Z_2)$ is maximized (subject to the constraint $w_1^2 + w_2^2 = 1$) by the weights $w_1 = 1/\sqrt{2} = v_{11}$ and $w_2 = 1/\sqrt{2} = v_{12}$.
Summary

Principal component analysis uses linear combinations of large and complex data sets to create a few new components that summarizes as much of the variability as possible.

The steps to calculate principal components for a data set consisting of \( k \) variables:

1. Calculate the correlation matrix, \( R \).
2. Calculate the \( k \) eigenvectors and eigenvalues of \( R \).
3. Use the eigenvectors to create principal components, uncorrelated linear combinations of the standardized data.
4. Order the principal components by the size of their corresponding eigenvalue and determine how many principal components to retain.

Principal component analysis is one of many exploratory tools that are useful for providing a general understanding and possible meaningful linear combinations. PCA is often calculated from random variables sampled from a larger population. A new sample from the same population would result in different coefficients (i.e. the linear combinations would change). Principal components are typically created to be used as variables in future data analysis methods and is most effective if just a few linear combinations of a large data set explain most of the variability.

Scree plots look at eigenvalues in order to determine how many principal components should be retained. Eigenvectors that indicate that the corresponding principal component explains much of the variability in the dataset, it is retained. Some researchers suggest retaining all components with eigenvalues greater than one while others retain enough principal components to explain 70% - 90% of the variability within the original data.

Principal component analysis is a non-parametric technique, since no assumptions are made about the distribution of the data. It is important to recognize that PCA is scale sensitive. Standardizing the data (which is equivalent to using the correlation matrix to calculate eigenvalues and eigenvectors), is often suggested. If the data is not standardized (covariance matrix is used), the principal components will follow variables with larger variances instead of representing the correlation structure of the entire dataset.

Eigenvectors and the correlation between principal components and original variables (called loadings) can be used to understand patterns within the data. Caution should be used when interpreting principal components. PCA is typically based on the correlation matrix and correlations only measure linear relationships. Thus, PCA may miss non-linear patterns within a data set. PCA is also influenced by outliers. Transformations on the original data can be attempted to address nonlinear patterns, outliers or skewness however caution should be used in interpreting the results. More advanced techniques can also be used to address these issues, but they are beyond the scope of this text.
Homework

(26) **Course Grades**

Data Source: Grades

In the author’s traditional introductory statistics course, the following weights are applied to determine the overall grade for the course.

- Homework 10%
- Labs 15%
- Project 20%
- Exam 1: 15%
- Exam 2: 15%
- Final Exam: 20%
- Quiz: 5%

(a) Use the weightings assigned by the author to give an overall course score to each of the 120 students appearing in the Grades data set. The top 15% of the students should get an A, the next 25% will be assigned a B, the next 35% will be assigned a C, the next 15% will be assigned a D and the final 10% will be assigned an F.

(b) Use the first principal component from a PCA analysis of the Grades data instead of the assigned weights to create an overall score for each student in the class. What percent of the total variability in grade components is explained by this score? If the first principal component were used instead of the predefined weights, how many students will get a different grade in the course? You must take out the final column of data, the totals created in part (a), to do PCA. Compare the weightings from the PCA versus the instructor’s weightings. Explain how changing the weightings influenced students’ final grades.

(c) *Improper PCA*: Repeat the previous question, but do not standardize the data. Explain why homework now has a much higher weight and is the primary variable in a student’s grade calculation. Hint: Which variable has the highest variability in the unstandardized data set?

(d) The author discovered that several students found homework solutions for the text on the Internet. Thus, several students had very high homework scores, but still had poor grades in the course. Explain why using unstandardized PCA to determine students grades would not be appropriate.

(27) **2008 Stock Market**

Data Source: 2008Stock

2008 was a very volatile year for the stock market. The data set 2008Stock contains several market index values, such as the closing S&P 500, Dow and NASDAQ values for every business day in 2008.
Conduct PCA on the **2008Stock** data set to create one overall market index. Which year, 2006 or 2008, did the first principal component explain more of the variability? Look at the weightings for the 2008 first principal component and interpret their meaning.

(28) **Turtle Shells**  
Data Source: Turtles

Jolicoeur and Mosimann measured the length, width and height of 48 painted turtle shells and conducted PCA on the data. This work has been influential in the field of allometry, the study of the relative growth of a part of an organism in relation to an entire organism.5

(a) Conduct principle component analysis on the entire data set (ignore gender). Can a majority of the variability in the data be explained with just one or two principal components?

(b) Calculate and interpret the principal component loadings.

(c) Create a scatterplot of PC1 versus PC2 where each point is grouped by gender. Was PCA useful in differentiating the genders? Other multivariate techniques are designed to classify or cluster data into clear groups.

(d) Conduct principle component analysis on the entire data set (ignore gender) using only the male data. Repeat the process using only female data. Compare the loadings for both sets of analysis, Would you expect the results to be similar? Why or why not?

(29) **Intelligence**  
Data Source: Intelligence

Measuring intelligence is another example where a principal component analysis is often used. There is not just one test that can be used to accurately measure a person’s intelligence, however there several tests that are known to be related to intelligence. Often intelligence tests can be classified into three groups: spatial reasoning, vocabulary, and memory. The data set **Intelligence** contains test results of 40 people who have each taken several intelligence tests. PCA is used to combine these multiple measurements into an overall measure of intelligence. Psychologists call the first component the general intelligence factor, or g-score.

(a) Use PCA on the **Intelligence** data set to find one overall measure of intelligence. How much of the total variability in the several measures of intelligence is explained by this g-score?

---

(b) Intelligence tests are often standardized into an IQ score so that the mean score is 100 with a standard deviation of 15. Transform your data and find the IQ of the 10th person in the Intelligence data set.

Principal components analysis was used to offer a “scientific” basis for racist, classist, and sexist theories of “intelligence.” See, for example, Stephen Jay Gould’s *The Mismeasure of Man* (1981).

There is little doubt that the US Congressional Immigration Act of 1924 was racially motivated to limit the entrance of several unwanted groups of people into the United States. Kamin (1974, p. 16) states that intelligence tests were used to verify that “83% of the Jews, 80% of the Hungarians, 79% of the Italians, and 87% of the Russians were ‘feeble-minded’.” Snyderman and Herrnstein’s “Intelligence Tests and the Immigration Act of 1924” (in *American Psychologist*, 38, 1983) questions whether or not psychologists had any influence on the act. They claim that intelligence tests were misused as “evidence” of why some people should not be allowed to immigrate to the U.S.

(c) Explain why intelligence tests (such as vocabulary tests) may have biased results when measuring the intelligence of newly arriving immigrants on Ellis Island.

Caution: While vocabulary may be a very appropriate measure of intelligence in some cultures, the same test can be completely invalid if applied to a new population. The use of PCA to construct one easy-to-understand variable from a large complex data set is powerful and useful, but like almost any powerful tool, it can be put to morally repugnant uses. Even the most carefully constructed and complex statistical analysis has incorrect conclusions if the data are not collected properly.

(d) Writing: Write a 2-3 page congressional report. Describe to the congressmen, who have not had a statistics class, how a g-score is derived. Then explain why, even though the concept of g-score is accepted by most psychologists, the test may not accurately describe an overall intelligence measure for all populations.

(30) Crime

Data Source: Crime

The FBI’s Uniform Crime Reporting (UCR) Program gathers crime data from more than 17,000 law enforcement agencies throughout the United States. The Crime data set contains crime rates per capita for several years. Crimes are listed by type of crime (murder, burglary, larceny, etc.), the state in which the crime occurred, and the type of area where the crime occurred (rural, metropolitan, or non-metropolitan cities).

(a) Conduct PCA on the crime variables in this data set to find one or two overall crime “variables.” What percentage of the variation can be explained by the top two principal components? How would you interpret these two new variables in terms of crime characteristics?
(b) Some may suggest to simply total the crime rates, in essence giving a weight of 1 to every variable. Give an example where a “Total” weight may be more appropriate than using PCA. Give an example where PCA may be more appropriate than a “Total” crime value.

(31) **Corn Suitability Rating**  
Data Source: Corn  
Iowa is one of the leading states in corn production. Each year approximately 12 to 13 million acres of corn are planted in Iowa. To evaluate the overall quality of soil, the Iowa Corn Suitability Rating (CSR) for each location is developed based on the kind of soil, slope of the land, erosion potential, water infiltration, and ease of machine operation in the area. The CSR can affect the value of an acre of Iowa farmland by thousands of dollars since it is a measure of the inherent productivity of the land.  

CSRs are listed by county in the Iowa Soil Properties and Interpretation Database at: [http://extension.agron.iastate.edu/soils/soilsurv.html](http://extension.agron.iastate.edu/soils/soilsurv.html).  

Use PCA on the **Corn** data set to calculate an overall measure of the productivity of Iowa cropland. Which variables seem to have the largest impact on this rating?

(32) **Healthy Behaviors**  
Data Source: Health  
The NHIS is a multistage probability sample survey that is conducted continuously throughout the year by the Centers for Disease Control and Prevention’s National Center for Health Statistics (NCHS). One aspect of this survey is examining the prevalence of several health risk behaviors. The table **Health** contains data collected from 2004-2006 on sleep behaviors, amount of drinking, dietary habits, amount of physical activity during leisure time, and smoking.  

Conduct principle component analysis on this data set. Can a majority of the variability in the data be explained with just one or two principal components? If so, how would you interpret these two new variables in terms of health risk behaviors?