CHAPTER 11
PRINCIPAL COMPONENTS ANALYSIS

We now have the tools necessary to discuss one of the most important concepts in mathematical statistics: **Principal Components Analysis (PCA)**. PCA involves the analysis of eigenvalues and eigenvectors of the covariance or correlation matrix. Its development relies on the following important facts:

Theorem 11.0.1: Diagonalization of Symmetric Matrices

All $n \times n$ real valued symmetric matrices (like the covariance and correlation matrix) have two very important properties:

1. They have a complete set of n linearly independent eigenvectors, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, corresponding to eigenvalues
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$
2. Furthermore, these eigenvectors can be chosen to be *orthonormal* so that if $\mathbf{V} = [\mathbf{v}_1 | \ldots | \mathbf{v}_n]$ then
$$\mathbf{V}^T\mathbf{V} = \mathbf{I}$$
or equivalently, $\mathbf{V}^{-1} = \mathbf{V}^T$.

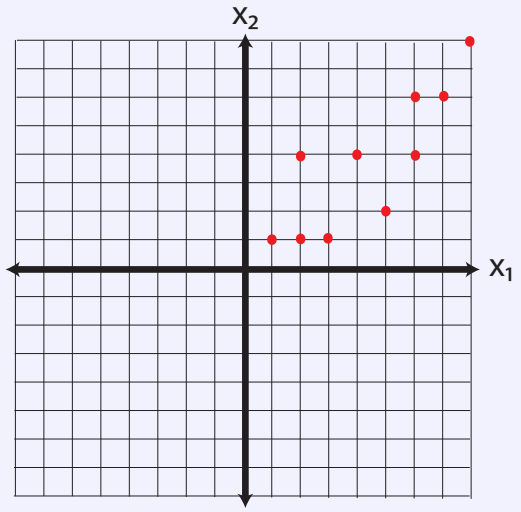
Letting $\mathbf{D}$ be a diagonal matrix with $D_{ii} = \lambda_i$, by the definition of eigenvalues and eigenvectors we have for any symmetric matrix $\mathbf{S}$,
$$\mathbf{SV} = \mathbf{VD}$$
Thus, any symmetric matrix $\mathbf{S}$ can be diagonalized in the following way:
$$\mathbf{V}^T\mathbf{SV} = \mathbf{D}$$

Covariance and Correlation matrices (when there is no perfect multicollinearity in variables) have the additional property that all of their eigenvalues are positive (nonzero). They are *positive definite* matrices.

Now that we know we have a complete set of eigenvectors, it is common to order them according to the magnitude of their corresponding eigenvalues. From here on out, we will use $(\lambda_1, \mathbf{v}_1)$ to represent the *largest* eigenvalue of a matrix and its corresponding eigenvector. When working with a covariance or correlation matrix, this eigenvector associated with the largest eigenvalue is called the **first principal component** and points in the direction for which the variance of the data is maximal. Example 11.0.1 illustrates this point.

Example 11.0.1: Eigenvectors of the Covariance Matrix

Suppose we have a matrix of data for 10 individuals on 2 variables, $\mathbf{x}_1$ and $\mathbf{x}_2$. Plotted on a plane, the data appears as follows:



Our data matrix for these points is:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 4 \\ 3 & 1 \\ 4 & 4 \\ 5 & 2 \\ 6 & 4 \\ 6 & 6 \\ 7 & 6 \\ 8 & 8 \end{pmatrix}$$

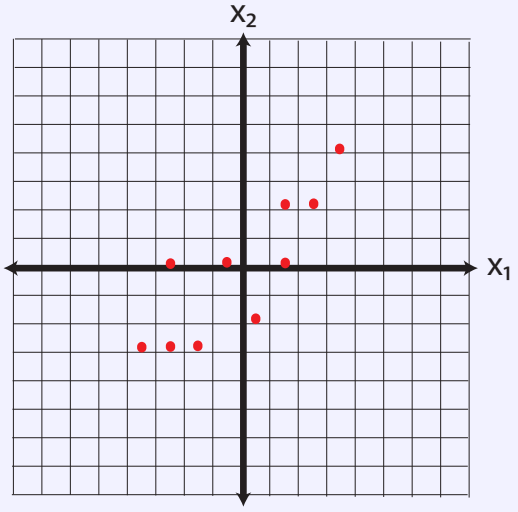
the means of the variables in $\mathbf{X}$ are:

$$\bar{\mathbf{x}} = \begin{pmatrix} 4.4 \\ 3.7 \end{pmatrix}.$$

When thinking about variance directions, our first step should be to center the data so that it has mean zero. Eigenvectors measure the spread of data around the origin. Variance measures spread of data around the mean. Thus, we need to equate the mean with the origin. To center the data, we simply compute

$$\mathbf{X}_c = \mathbf{X} - \mathbf{e}\bar{\mathbf{x}}^T = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 4 \\ 3 & 1 \\ 4 & 4 \\ 5 & 2 \\ 6 & 4 \\ 6 & 6 \\ 7 & 6 \\ 8 & 8 \end{pmatrix} - \begin{pmatrix} 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \\ 4.4 & 3.7 \end{pmatrix} = \begin{pmatrix} -3.4 & -2.7 \\ -2.4 & -2.7 \\ -2.4 & 0.3 \\ -1.4 & -2.7 \\ -0.4 & 0.3 \\ 0.6 & -1.7 \\ 1.6 & 0.3 \\ 1.6 & 2.3 \\ 2.6 & 2.3 \\ 3.6 & 4.3 \end{pmatrix}.$$

Examining the new centered data, we find that we've only translated our data in the plane - we haven't distorted it in any fashion.



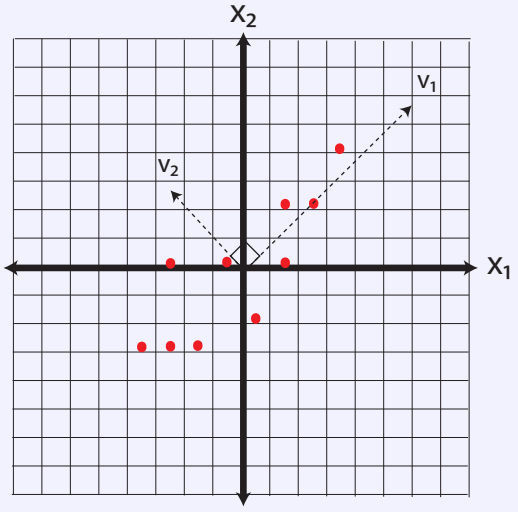
Thus the covariance matrix is:

$$\mathbf{\Sigma} = \frac{1}{9}(\mathbf{X}_c^T \mathbf{X}_c) = \begin{pmatrix} 5.6 & 4.8 \\ 4.8 & 6.0111 \end{pmatrix}$$

The eigenvalue and eigenvector pairs of $\mathbf{\Sigma}$ are (rounded to 2 decimal places) as follows:

$$(\lambda_1, \mathbf{v}_1) = \left(10.6100, \begin{bmatrix} 0.69 \\ 0.72 \end{bmatrix}\right) \text{ and } (\lambda_2, \mathbf{v}_2) = \left(1.0012, \begin{bmatrix} -0.72 \\ 0.69 \end{bmatrix}\right)$$

Let's plot the eigenvector directions on the same graph:



The eigenvector $\mathbf{v}_1$ is called the **first principal component**. It is the direction along which the variance of the data is maximal. The eigenvector $\mathbf{v}_2$ is the **second principal component**. In general, the second principal component is the direction, orthogonal to the first, along which the variance of the data is maximal (in two dimensions, there is only one direction possible.)

Why is this important? Let's consider what we've just done. We started with two variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, which appeared to be correlated. We then derived *new variables*, $\mathbf{v}_1$ and $\mathbf{v}_2$, which are linear combinations of the original variables:

$$\mathbf{v}_1 = 0.69\mathbf{x}_1 + 0.72\mathbf{x}_2 \tag{11.1}$$
$$\mathbf{v}_2 = -0.72\mathbf{x}_1 + 0.69\mathbf{x}_2 \tag{11.2}$$

These new variables are completely uncorrelated. To see this, let's represent our data according to the new variables - i.e. let's change the basis from $\mathcal{B}_1 = [\mathbf{x}_1, \mathbf{x}_2]$ to $\mathcal{B}_2 = [\mathbf{v}_1, \mathbf{v}_2]$.

Example 11.0.2: The Principal Component Basis

Let's express our data in the basis defined by the principal components. We want to find coordinates (in a 2×10 matrix $\mathbf{A}$) such that our original (centered) data can be expressed in terms of principal components. This is done by solving for $\mathbf{A}$ in the following equation (see Chapter 9 and note that the *rows* of $\mathbf{X}$ define the points rather than the columns):

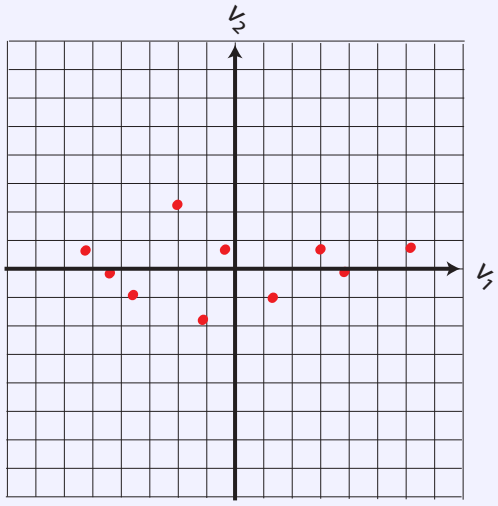
$$\mathbf{X}_c = \mathbf{A}\mathbf{V}^T \tag{11.3}$$

$$\begin{pmatrix} -3.4 & -2.7 \\ -2.4 & -2.7 \\ -2.4 & 0.3 \\ -1.4 & -2.7 \\ -0.4 & 0.3 \\ 0.6 & -1.7 \\ 1.6 & 0.3 \\ 1.6 & 2.3 \\ 2.6 & 2.3 \\ 3.6 & 4.3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \\ a_{71} & a_{72} \\ a_{81} & a_{82} \\ a_{91} & a_{92} \\ a_{10,1} & a_{10,2} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix} \tag{11.4}$$

Conveniently, our new basis is orthonormal meaning that $\mathbf{V}$ is an orthogonal matrix, so

$$\mathbf{A} = \mathbf{X}\mathbf{V}.$$

The new data coordinates reflect a simple rotation of the data around the origin:



Visually, we can see that the new variables are uncorrelated. You may wish to confirm this by calculating the covariance. In fact, we can do this in a general sense. If $\mathbf{A} = \mathbf{X}_c\mathbf{V}$ is our new data, then the covariance matrix is diagonal:

$$\begin{aligned}
\mathbf{\Sigma}_A &= \frac{1}{n-1}\mathbf{A}^T\mathbf{A} \\
&= \frac{1}{n-1}(\mathbf{X}_c\mathbf{V})^T(\mathbf{X}_c\mathbf{V}) \\
&= \frac{1}{n-1}\mathbf{V}^T((\mathbf{X}_c^T\mathbf{X}_c)\mathbf{V} \\
&= \frac{1}{n-1}\mathbf{V}^T((n-1)\mathbf{\Sigma}_X)\mathbf{V} \\
&= \mathbf{V}^T(\mathbf{\Sigma}_X)\mathbf{V} \\
&= \mathbf{V}^T(\mathbf{V}\mathbf{D}\mathbf{V}^T)\mathbf{V} \\
&= \mathbf{D}
\end{aligned}$$

Where $\mathbf{\Sigma}_X = \mathbf{V}\mathbf{D}\mathbf{V}^T$ comes from the diagonalization in Theorem 11.0.1. By changing our variables to principal components, we have managed to **"hide" the correlation** between $\mathbf{x}_1$ and $\mathbf{x}_2$ while keeping the spacial relationships between data points in tact. Transformation *back* to variables $\mathbf{x}_1$ and $\mathbf{x}_2$ is easily done by using the linear relationships in Equations 11.1 and 11.2.

11.1 Comparison with Least Squares

In least squares regression, our objective is to maximize the amount of variance explained in our target variable. It may look as though the first principal component from Example 11.0.1 points in the direction of the regression line. This is not the case however. The first principal component points in the direction of a line which minimizes the sum of squared *orthogonal* distances between the points and the line. Regressing $\mathbf{x}_2$ on $\mathbf{x}_1$, on the other hand, provides a line which minimizes the sum of squared *vertical* distances between points and the line. This is illustrated in Figure 11.1.

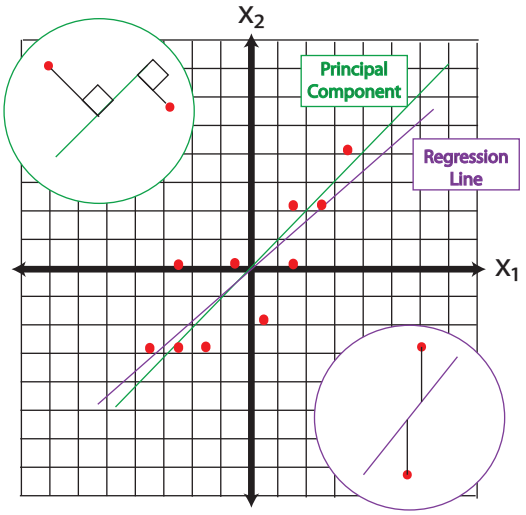


Figure 11.1: Principal Components vs. Regression Lines

The first principal component about the mean of a set of points can be represented by that line which most closely approaches the data points. In contrast, linear least squares tries to minimize the distance in the y direction only. Thus, although the two use a similar error metric, linear least squares is a method that treats one dimension of the data preferentially, while PCA treats all dimensions equally.

11.2 Covariance or Correlation Matrix?

Principal components analysis can involve eigenvectors of either the covariance matrix or the correlation matrix. When we perform this analysis on the covariance matrix, the geometric interpretation is simply centering the data and then determining the direction of maximal variance. When we perform

this analysis on the correlation matrix, the interpretation is *standardizing* the data and then determining the direction of maximal variance. The correlation matrix is simply a scaled form of the covariance matrix. In general, these two methods give different results, especially when the scales of the variables are different.

The covariance matrix is the default for R. The correlation matrix is the default in SAS. The covariance matrix method is invoked by the option:

```
proc princomp data=X cov;
var x1--x10;
run;
```

Choosing between the covariance and correlation matrix can sometimes pose problems. The rule of thumb is that the correlation matrix should be used when the scales of the variables vary greatly. In this case, the variables with the highest variance will dominate the first principal component. The argument against automatically using correlation matrices is that it is quite a brutal way of standardizing your data.

11.3 Applications of Principal Components

Principal components have a number of applications across many areas of statistics. In the next sections, we will explore their usefulness in the context of dimension reduction. In Chapter 14 we will look at how PCA is used to solve the issue of multicollinearity in biased regression.

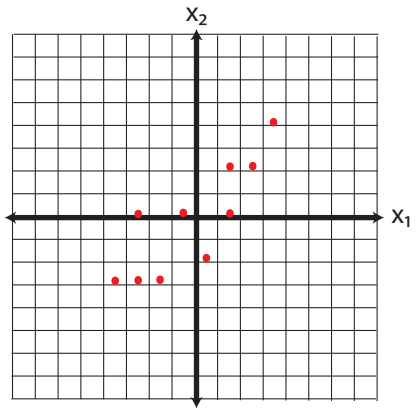
11.3.1 PCA for dimension reduction

It is quite common for an analyst to have too many variables. There are two different solutions to this problem:

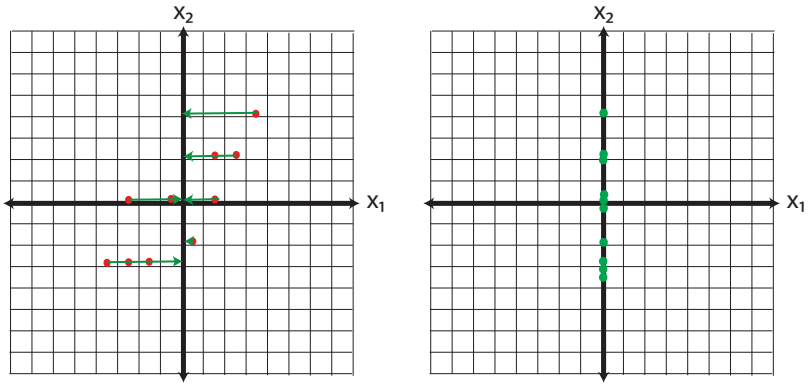
1. **Feature Selection**: Choose a subset of existing variables to be used in a model.
2. **Feature Extraction**: Create a new set of features which are combinations of original variables.

Feature Selection

Let's think for a minute about feature selection. What are we really doing when we consider a subset of our existing variables? Take the two dimensional data in Example 11.0.2 (while two-dimensions rarely necessitate dimension reduction, the geometrical interpretation extends to higher dimensions as usual!). The centered data appears as follows:



Now say we perform some kind of feature selection (there are a number of ways to do this, chi-square tests for instances) and we determine that the variable $\mathbf{x}_2$ is more important than $\mathbf{x}_1$. So we throw out $\mathbf{x}_2$ and we've reduced the dimensions from $p = 2$ to $k = 1$. Geometrically, what does our new data look like? By dropping $\mathbf{x}_1$ we set all of those horizontal coordinates to zero. In other words, we **project the data orthogonally** onto the $\mathbf{x}_2$ axis:



(a) Projecting Data Orthogonally (b) New One-Dimensional Data

Figure 11.2: Geometrical Interpretation of Feature Selection

Now, how much information (variance) did we lose with this projection? The total variance in the original data is

$$\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2.$$

The variance of our data reduction is

$$\|\mathbf{x}_2\|^2.$$

Thus, the proportion of the total information (variance) we've kept is

$$\frac{\|\mathbf{x}_2\|^2}{\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2} = \frac{6.01}{5.6 + 6.01} = 51.7\%.$$

Our reduced dimensional data contains only 51.7% of the variance of the original data. We've lost a lot of information!

The fact that feature selection omits variance in our predictor variables does not make it a bad thing! Obviously, getting rid of variables which have no relationship to a target variable (in the case of *supervised* modeling like prediction and classification) is a good thing. But, in the case of *unsupervised* learning techniques, where there is no target variable involved, we must be extra careful when it comes to feature selection. In summary,

- Feature Selection is important. Examples include:
  - Removing variables which have little to no impact on a target variable in supervised modeling (forward/backward/stepwise selection).
  - Removing variables which have obvious strong correlation with other predictors.
  - Removing variables that are not interesting in unsupervised learning (For example, you may not want to use the words "th" and "of" when clustering text).
- Feature Selection is an orthogonal projection of the original data onto the span of the variables you choose to keep.
- Feature selection should always be done with care and justification.
  - In regression, could create problems of endogeneity (errors correlated with predictors - omitted variable bias).
  - For unsupervised modelling, could lose important information.

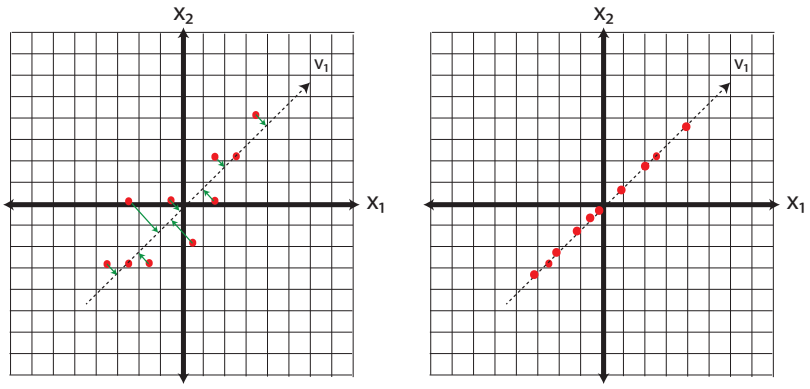
Feature Extraction

PCA is the most common form of feature extraction. The rotation of the space shown in Example 11.0.2 represents the creation of new features which are linear combinations of the original features. If we have p potential variables for a model and want to reduce that number to k, then the first k principal components combine the individual variables in such a way that is guaranteed to capture as much "information" (variance) as possible. Again, take our two-dimensional data as an example. When we reduce our data down to one-dimension using principal components, we essentially do the same orthogonal projection that we did in Feature Selection, only in this case we conduct that

projection in the new basis of principal components. Recall that for this data, our first principal component $\mathbf{v}_1$ was

$$\mathbf{v}_1 = \begin{pmatrix} 0.69 \\ 0.73 \end{pmatrix}.$$

Projecting the data onto the first principal component is illustrated in Figure 11.3 How much variance do we keep with k principal components? The



(a) Projecting Data Orthogonally (b) New One-Dimensional Data

Figure 11.3: Illustration of Feature Extraction via PCA

proportion of variance explained by each principal component is the ratio of the corresponding eigenvalue to the sum of the eigenvalues (which gives the total amount of variance in the data).

Theorem 11.3.1: Proportion of Variance Explained

The proportion of variance explained by the projection of the data onto principal component $\mathbf{v}_i$ is

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

Similarly, the proportion of variance explained by the projection of the data onto the first k principal components ($k < j$) is

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

In our simple 2 dimensional example we were able to keep

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{10.61}{10.61 + 1.00} = 91.38\%$$

of our variance in one dimension.