PRINCIPAL COMPONENTS ANALYSIS
PCA - INTRODUCTION

\[ X = \begin{bmatrix}
1 & 2 & 4 \\
2 & 1 & 5 \\
3 & 4 & 10 \\
4 & 3 & 11 \\
\end{bmatrix} \]
### PCA - INTRODUCTION

<table>
<thead>
<tr>
<th>sepal_length</th>
<th>sepal_width</th>
<th>petal_length</th>
<th>petal_width</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4.4</td>
<td>2.9</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>52</td>
<td>6.9</td>
<td>3.1</td>
<td>4.9</td>
<td>1.5</td>
</tr>
<tr>
<td>35</td>
<td>5.0</td>
<td>3.2</td>
<td>1.2</td>
<td>0.2</td>
</tr>
<tr>
<td>127</td>
<td>6.1</td>
<td>3.0</td>
<td>4.9</td>
<td>1.8</td>
</tr>
<tr>
<td>96</td>
<td>5.7</td>
<td>2.9</td>
<td>4.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>
PCA - INTRODUCTION
PCA - INTRODUCTION

![PCA of IRIS dataset](image)

- setosa
- versicolor
- virginica
PRINCIPAL COMPONENT ANALYSIS

• A technique to find the directions along which the points (set of tuples) in high-dimensional data line up best.

• Treat a set of tuples as a matrix $M$ and find the eigenvectors for $MM^T$ or $M^TM$.

• The matrix of these eigenvectors can be thought of as a rigid rotation in a high-dimensional space.

• When this transformation is applied to the original data - the axis corresponding to the principal eigenvector is the one along which the points are most “spread out”.
PRINCIPAL COMPONENT ANALYSIS

• When this transformation is applied to the original data - the axis corresponding to the principal eigenvector is the one along which the points are most “spread out”.

• This axis is the one along which variance of the data is maximized.

• Points can best be viewed as lying along this axis with small deviations from this axis.

• Likewise, the axis corresponding the second eigenvector is the axis along which the variance of distances from the first axis is greatest, and so on.
PRINCIPAL COMPONENT ANALYSIS

• Principal Component Analysis (PCA) is a dimensionality reduction method.

• The goal is to embed data in high dimensional space, onto a small number of dimensions.

• It most frequent use is in exploratory data analysis and visualization.

• It can also be helpful in regression (linear or logistic) where we can transform input variables into a smaller number of predictors for modeling.
PRINCIPAL COMPONENT ANALYSIS

• Mathematically,

Given: Data set \( \{x_1, x_2, \ldots, x_n\} \)

where, \( x_i \) is the vector of p variable values for the i-th observation.

Return:

Matrix \([\phi_1, \phi_2, \ldots, \phi_p]\)

of linear transformations that retain maximal variance.

• You can think of the first vector \( \phi_1 \) as a linear transformation that embeds observations into 1 dimension

\[
Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p
\]
**PRINCIPAL COMPONENT ANALYSIS**

• You can think of the first vector $\phi_1$ as a linear transformation that embeds observations into 1 dimension

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

where $\phi_1$ is selected so that the resulting dataset \{\(z_i, \ldots, z_n\)\} has maximum variance.

• In order for this to make sense, mathematically, data has to be centered
  • Each $X_i$ has zero mean
  • Transformation vector $\phi_1$ has to be normalized, i.e., $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
PRINCIPAL COMPONENT ANALYSIS

• In order for this to make sense, mathematically, data has to be centered
  • Each $X_i$ has zero mean
  • Transformation vector $\phi_1$ has to be normalized, i.e., $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

• We can find $\phi_1$ by solving an optimization problem:

$$\max_{\phi_{11}, \phi_{21}, \ldots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \text{ s.t. } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Maximize variance but subject to normalization constraint.
PRINCIPAL COMPONENT ANALYSIS

• We can find $\phi_1$ by solving an optimization problem:

$$\max_{\phi_{11}, \phi_{21}, \ldots, \phi_{p1}} \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{1j} x_{ij} \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \phi_{1j}^2 = 1$$

Maximize variance but subject to normalization constraint.

• The second transformation, $\phi_2$ is obtained similarly with the added constraint that $\phi_2$ is orthogonal to $\phi_1$

• Taken together $[\phi_1, \phi_2]$ define a pair of linear transformations of the data into 2 dimensional space

$$Z_{n \times 2} = X_{n \times p}[\phi_1, \phi_2]_{p \times 2}$$
PRINCIPAL COMPONENT ANALYSIS

• Taken together $[\phi_1, \phi_2]$ define a pair of linear transformations of the data into 2 dimensional space

$$Z_{n\times2} = X_{n\times p} [\phi_1, \phi_2]_{p\times2}$$

• Each of the columns of the $Z$ matrix are called Principal components.

• The units of the PCs are meaningless.

• In practice we may also scale $X_j$ to have unit variance.

• In general if variables $X_j$ are measured in different units (e.g., miles vs. liters vs. dollars), variables should be scaled to have unit variance.
SPECTRAL THEOREM

Using Spectral theorem

\[(X^TX)\phi = \lambda \phi\]

\[XX^T(X\phi) = \lambda(X\phi)\]

Conclusion:

The matrices \(XX^T\) and \(X^TX\) share the same nonzero eigenvalues.

To get an eigenvector of \(XX^T\) from \(X^TX\) multiply \(\phi\) on the left by \(X\).

Very powerful, particularly if number of observations, \(m\), and the number of predictors, \(n\), are drastically different in size.

For PCA:

\[\text{Cov}(X, X) = XX^T\]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

Eigen Values and Eigen Vectors?
EXAMPLE - PCA

\[
X = \begin{bmatrix}
1 & 2 \\
2 & 1 \\
3 & 4 \\
4 & 3
\end{bmatrix}
\]

From spectral theorem:

\[
(X^T X) \phi = \lambda \phi
\]

\[
X^T X = \begin{bmatrix}
1 & 2 & 3 & 4 \\
2 & 1 & 4 & 3
\end{bmatrix} \begin{bmatrix}
1 & 2 \\
2 & 1 \\
3 & 4 \\
4 & 3
\end{bmatrix} = \begin{bmatrix}
30 & 28 \\
28 & 30
\end{bmatrix}
\]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

From spectral theorem:

\[(X^TX)\phi = \lambda \phi \implies (X^TX)\phi - \lambda I\phi = 0\]

\[ ((X^TX) - \lambda I)\phi = 0 \]

\[
\begin{bmatrix}
30 - \lambda & 28 \\
28 & 30 - \lambda
\end{bmatrix} = 0 \implies \lambda = 58 \text{ and } \lambda = 2
\]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

From spectral theorem:

\[ (X^T X) \phi = \lambda \phi \]

\[ \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} \phi_{11} \\ \phi_{12} \end{bmatrix} = 58 \begin{bmatrix} \phi_{11} \\ \phi_{12} \end{bmatrix} \implies \phi_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

From spectral theorem:

\[ (X^T X)\phi = \lambda \phi \]

\[ \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} \phi_{11} \\ \phi_{12} \end{bmatrix} = 58 \begin{bmatrix} \phi_{11} \\ \phi_{12} \end{bmatrix} \implies \phi_1 = \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \]

\[ \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} \phi_{21} \\ \phi_{22} \end{bmatrix} = 2 \begin{bmatrix} \phi_{21} \\ \phi_{22} \end{bmatrix} \implies \phi_2 = \begin{bmatrix} -1 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

From spectral theorem:

\[(X^T X) \phi = \lambda \phi\]

\[ \phi_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \lambda_1 = 58 \]

\[ \phi_2 = \begin{bmatrix} -1 \\ \sqrt{2} \end{bmatrix} \quad \lambda_2 = 2 \]

\[ \phi = \begin{bmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & \sqrt{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \quad \phi_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \lambda_1 = 58 \quad \phi_2 = \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \lambda_2 = 2 \]

\[ Z = X\phi = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{3}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \]
EXAMPLE - PCA

\[ X = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \]

\[ Z = \begin{bmatrix} \frac{3}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \]
HOW MANY PRINCIPAL COMPONENTS?

• How many PCs should we consider in post-hoc analysis?

• One result of PCA is a measure of the variance to each PC relative to the total variance of the dataset.

• We can calculate the percentage of variance explained for the m-th PC:

\[
PVE_m = \frac{\sum_{i=1}^{n} z_{im}^2}{p \sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}
\]
HOW MANY PRINCIPAL COMPONENTS?

- We can calculate the percentage of variance explained for the m-th PC:

\[
PVE_m = \frac{\sum_{i=1}^{n} z_{im}^2}{p \sum_{j=1}^{n} \sum_{i=1}^{n} x_{ij}^2}
\]