Principal Component Analysis is

- a tool to analyze the data
- a tool to do dimensionality reduction
Auto data set

![Scatter plots showing relationships between variables: mpg, cylinders, horsepower, and weight.](image-url)
Basic concepts needed

• data analysis
  measures of center and spread, covariance and correlation

• linear algebra
  eigenvectors, eigenvalues, dot product, basis
How two features are related

Both covariance and correlation indicate how closely two features relationship follows a straight line.

**Covariance** \( \text{cov}(X, Y) \) is a measure of the joint variability of two random variables \( X \) and \( Y \)

\[
\text{cov}(X, Y) = E[(X - EX)(Y - EY)]
\]

The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables. Therefore normalize the covariance → **Pearson correlation** coefficient

\[
-1 \leq \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \leq +1
\]
Covariance matrix of features $A_1, \ldots, A_m$

$$\text{COV}(A_1, \ldots, A_m) = \begin{pmatrix}
\text{var}(A_1) & \text{cov}(A_1, A_2) & \ldots & \text{cov}(A_1, A_m) \\
\text{cov}(A_2, A_1) & \text{var}(A_2) & \ldots & \text{cov}(A_2, A_m) \\
\ldots & \ldots & \ldots & \ldots \\
\text{cov}(A_m, A_1) & \text{cov}(A_m, A_2) & \ldots & \text{var}(A_m)
\end{pmatrix}$$
```r
> cov(Auto[,c("mpg", "cylinders", "horsepower", "weight")])

<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cylinders</th>
<th>horsepower</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>60.91814</td>
<td>-10.35293</td>
<td>-233.85793</td>
<td>-5517.441</td>
</tr>
<tr>
<td>cylinders</td>
<td>-10.35293</td>
<td>2.909696</td>
<td>55.34824</td>
<td>1300.424</td>
</tr>
<tr>
<td>horsepower</td>
<td>-233.85793</td>
<td>55.348244</td>
<td>1481.56939</td>
<td>28265.620</td>
</tr>
<tr>
<td>weight</td>
<td>-5517.44070</td>
<td>1300.424363</td>
<td>28265.62023</td>
<td>721484.709</td>
</tr>
</tbody>
</table>

> cor(Auto[,c("mpg", "cylinders", "horsepower", "weight")])

<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cylinders</th>
<th>horsepower</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>1.0000000</td>
<td>-0.7776175</td>
<td>-0.7784268</td>
<td>-0.8322442</td>
</tr>
<tr>
<td>cylinders</td>
<td>-0.7776175</td>
<td>1.0000000</td>
<td>0.8429834</td>
<td>0.8975273</td>
</tr>
<tr>
<td>horsepower</td>
<td>-0.7784268</td>
<td>0.8429834</td>
<td>1.0000000</td>
<td>0.8645377</td>
</tr>
<tr>
<td>weight</td>
<td>-0.8322442</td>
<td>0.8975273</td>
<td>0.8645377</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>
```
**Eigenvector u, eigenvalue λ**: $A \cdot u = \lambda u$

- $u$ does not change its direction under the transformation
- $\lambda u$ scales a vector $u$ by $\lambda$; it changes its length, not its direction

1. The covariance matrix of $X$ is an $m \times m$ symmetric matrix given by $\frac{1}{n-1}XX^\top$

2. Any symmetric matrix $m \times m$ has a set of orthonormal eigenvectors $v_1, v_2, \ldots, v_m$ associated with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$
   - for any $i$, $A \cdot v_i = \lambda_i v_i$
   - $\|v_i\| = 1$
   - $v_i \cdot v_j = 0$ if $i \neq j$

3. A is a symmetric $m \times m$ matrix and $E$ is an $m \times m$ matrix whose $i$-th column is the $i$-th eigenvector of $A$. The eigenvectors are ordered in terms of decreasing values of their associated eigenvalues. Then there is a diagonal matrix $D$ such that $A = E \cdot D \cdot E^\top$

4. If the rows of $E$ are orthogonal, then $E^{-1} = E^\top$
**Basis** of $\mathcal{R}^m$ is a set of linearly independent vectors $u_1, \ldots, u_m$

- none of them is a linear combination of other vectors
- $u_i \cdot u_j = 0$, $i, j = 1, \ldots, m$, $i \neq j$
- any $u = c_1 u_1 + \cdots + c_m u_m$
- for example, the standard basis of the 3-dimensional Euclidean space $\mathcal{R}^3$ consists of $x = \langle 1, 0, 0 \rangle$, $y = \langle 0, 1, 0 \rangle$, $z = \langle 0, 0, 1 \rangle$. It is an example of orthonormal basis, so called *naive* basis.
Representation of Data = \{x_i, x_i = \langle x_{1i}, \ldots, x_{mi}\rangle\}, |Data| = n for PCA

\[ X = \begin{pmatrix} x_{11} & \ldots & x_{1n} \\ x_{21} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots \\ x_{m1} & \ldots & x_{mn} \end{pmatrix} \]
Which features to keep?

- features that change a lot, i.e. high variance
- features that do not depend on others, i.e. low covariance

Which features to ignore?

- features with some noise, i.e. low variance
PCA principles

1. high correlation $\sim$ high redundancy

2. the most important feature has the largest variance
• **Question**

Is there any other representation of $X$ to extract the most important features?

• **Answer**

Use another basis

$$P^T \cdot X = Z$$

where $P$ transforms $X$ into $Z$
The principal components of $X$ are the vectors $p_i = \langle p_{1i}, \ldots, p_{mi} \rangle$

- **principal components** of $X$ are the vectors $p_i = \langle p_{1i}, \ldots, p_{mi} \rangle$
- **principal component loadings** of $p_i$ are the elements $p_{i1}, \ldots, p_{im}$

\[ P = \begin{pmatrix} p_{11} & \cdots & \cdots & p_{1m} \\ p_{21} & \cdots & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & \cdots & \cdots & p_{mm} \end{pmatrix} \]
PCA

i-principal component scores of $n$ instances are $p_i \cdot x_1, p_i \cdot x_2, \ldots, p_i \cdot x_n$
What is a good choice of $P$?

What features we would like $Z$ to exhibit?

**Goal:** $Z$ is a new representation of $X$

The new features are linear combinations of the original features whose weights are given by $P$.

The covariance matrix of $Z$ is diagonal and the entries on the diagonal are in descending order, i.e. the covariance of any pair of distinct features is zero, and the variance of each of our new features is listed along the diagonal.
• principal components are new basis vectors to represent \( x_j, j = 1, \ldots, n \)

• \( p_i \cdot x_j \) is a projection of \( x_j \) on \( p_i \)

• changing the basis does not change data, it changes their representation

Covariance matrix \( \text{cov}(A_1, A_2, \ldots, A_m) \)

• on the diagonal, large values correspond to interesting structure

• off the diagonal, large values correspond to high redundancy
Derivation of PCA

1. **Preprocessing Data**
   Mean normalization to get centered data $\rightarrow X$

2. **Covariance Matrix**
   \[
   \text{cov}(X) = A = \frac{1}{n-1} XX^\top
   \]

3. **Compute Eigenvectors**
   Compute eigenvectors $v_1, \ldots, v_m$ and eigenvalues $\lambda_1, \ldots, \lambda_m$ of $A$

4. **Order Eigenvalues**
   Take the eigenvectors, order them by eigenvalues, i.e. by significance, highest to lowest: $p_1, \ldots, p_m$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$

5. **Eigenvectors as Columns**
   The eigenvectors $p_1, \ldots, p_m$ become columns of $P$

   \[
   p_i = \begin{pmatrix}
   p_{1i} \\
   \vdots \\
   p_{mi}
   \end{pmatrix}
   \]
Properties of PCA

\[ P^\top \cdot X = Z \]

\[ Z = \begin{pmatrix} p_1 \cdot x_1 & \ldots & \ldots & p_1 \cdot x_n \\ p_2 \cdot x_1 & \ldots & \ldots & p_2 \cdot x_n \\ \ldots & \ldots & \ldots & \ldots \\ p_m \cdot x_1 & \ldots & \ldots & p_m \cdot x_n \end{pmatrix} \]

- The \( i \)-th diagonal value of \( \text{cov}(Z) \) is the variance of \( X \) along \( p_i \).
- We calculate a rotation of the original coordinate system such that all non-diagonal elements of the new covariance matrix become zero.
- The principal components define the basis of the new coordinate axes and the eigenvalues correspond to the diagonal elements of the new covariance matrix.
- So the eigenvalues, by definition, define the variance along the corresponding principal components.
Properties of PCA

\[ cov(P^T \cdot X) \overset{\text{see p.49.1}}{=} \frac{1}{n-1} (P^T \cdot X) \cdot (P^T \cdot X)^T = \]

\[ \frac{1}{n-1} P^T \cdot X \cdot X^T \cdot P \overset{\text{let } A = X \cdot X^T}{=} \frac{1}{n-1} P^T \cdot A \cdot P = \]

\[ \overset{\text{see p.49.3}}{=} \frac{1}{n-1} P^T \cdot (P \cdot D \cdot P^T) \cdot P \overset{\text{see p.49.4}}{=} \frac{1}{n-1} P^T \cdot (P^T)^{-1} D \cdot P^T \cdot (P^T)^{-1} = \frac{1}{n-1} D \]
A geometric interpretation for the first principal component \( p_1 \)

It defines a direction in feature space along which the data vary the most. If we project the \( n \) instances \( x_1, \ldots, x_n \) onto this direction, the projected values are the principal component scores \( z_{11}, \ldots, z_{n1} \) themselves.
The fraction of variance explained by a $k$-th principal component $\text{PVE}(p_k)$ is the ratio between the variance of that principal component and the total variance.

- **total variance in $X$:** $\sum_{j=1}^{m} \text{var}(A_j) = \sum_{i=1}^{m} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$
  (assuming feature normalization)

- **variance expressed by $p_k$:** $\frac{1}{n} \sum_{i=1}^{n} z_{ki}^2$

- $\text{PVE}(p_k) = \frac{\sum_{i=1}^{n} z_{ki}^2}{\sum_{i=1}^{m} \sum_{i=1}^{n} x_{ij}^2}$

- $\text{PVE}(p_1, \ldots, p_M) = \sum_{i=1}^{M} \text{PVE}(p_i), \ M \leq m$
PCA
Auto data set

```r
> a <- Auto[c("mpg", "cylinders", "horsepower", "weight")]
> pca.a <- prcomp(a, scale = TRUE)
> summary(pca.a)

# Importance of components:
#                 Comp.1  Comp.2  Comp.3  Comp.4
Standard deviation 1.8704  0.49540  0.40390  0.30518
Proportion of Variance 0.8746  0.06135  0.04078  0.02328
Cumulative Proportion 0.8746  0.93593  0.97672  1.00000
```
Scree plot
PCA
Auto data set

> pca.a$rotation

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>0.4833271</td>
<td>0.8550485</td>
<td>-0.02994982</td>
<td>0.1854453</td>
</tr>
<tr>
<td>cylinders</td>
<td>-0.5033993</td>
<td>0.3818233</td>
<td>-0.55748381</td>
<td>-0.5385276</td>
</tr>
<tr>
<td>horsepower</td>
<td>-0.4984381</td>
<td>0.3346173</td>
<td>0.79129092</td>
<td>-0.1159714</td>
</tr>
<tr>
<td>weight</td>
<td>-0.5143380</td>
<td>0.1055192</td>
<td>-0.24934614</td>
<td>0.8137252</td>
</tr>
</tbody>
</table>

- PC1 places approximately equal weight on cylinders, horsepower, weight with much higher weight on mpg.
- PC2 places most of its weight on mpg and less weight on the other three features.
A biplot displays both the PC scores and the PC loadings.
The biplot for the Auto data set is showing

- the scores of each example (i.e., cars) on the first two principal components with axes on the top and right – see the id cars in black
- the loading of each feature (i.e., mpg, weight, cylinders, horsepower) on the first two principal components with axes on the bottom and left – see the red arrows
In general, a $m \times n$ matrix $X$ has $\min(n - 1, m)$ distinct principal components.

- **Question**
  How many principal components are needed?

- **Answer**
  There is no single answer to this question. Study scree plots.
• Principal Component Analysis
  data analysis, derivation, scree plot, biplot