Principal Components Analysis (PCA)

Exploratory data analysis of high-dimensional data sets.
Example: Consider a data set of heights and weights of people.
Example: Consider a data set of heights and weights of people
Example: Consider a data set of heights and weights of people.
PCA on this data set reframes data in terms of overall size and heavynessness.
The math behind PCA

Variance of one variable:

$$\text{Var}(X) = \frac{1}{n} \sum_{j} (\bar{x} - x_j)^2 = \sigma^2_X$$

Covariance of two variables:

$$\text{Cov}(X,Y) = \frac{1}{n} \sum_{j} (\bar{x} - x_j)(\bar{y} - y_j) = \sigma^2_{XY}$$
The math behind PCA

Covariance matrix of $n$ variables $X_1 \ldots X_n$:

$$
C = \begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}
\end{pmatrix}
$$
PCA diagonalizes the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \mathbf{U} \mathbf{D} \mathbf{U}^T$$

$$= \mathbf{U} \begin{pmatrix}
\lambda_1^2 & 0 & \cdots & 0 \\
0 & \lambda_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n^2
\end{pmatrix} \mathbf{U}^T$$
PCA diagonalizes the covariance matrix $C$:

\[ C = U D U^T \]

where $U$ is the rotation matrix and $D$ is the diagonal matrix with the eigenvalues $\lambda_i^2$ on its diagonal.
The math behind PCA

PCA diagonalizes the covariance matrix \( C \):

\[
C = U D U^T
\]

\[
= U \begin{pmatrix}
\lambda_1^2 & 0 & \cdots & 0 \\
0 & \lambda_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n^2
\end{pmatrix} U^T
\]

diagonal matrix
The math behind PCA

PCA diagonalizes the covariance matrix $C$:

$$C = UDU^T$$

$U$ is an orthogonal matrix of eigenvectors, and $D$ is a diagonal matrix of eigenvalues:

$$D = \begin{pmatrix}
\lambda_1^2 & 0 & \cdots & 0 \\
0 & \lambda_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n^2
\end{pmatrix}$$

The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the variances explained by each component.
The math behind PCA

PCA diagonalizes the covariance matrix $\mathbf{C}$:

$$
\mathbf{C} = \mathbf{UDU}^T
$$

$$
= \mathbf{U} \begin{pmatrix}
\lambda_1^2 & 0 & \cdots & 0 \\
0 & \lambda_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n^2
\end{pmatrix} \mathbf{U}^T
$$

covariance between components is zero (they are uncorrelated)
In our earlier example, overall size and heaviness are uncorrelated.
Doing a PCA in R

\begin{verbatim}
iris %>%
  select(-Species) %>% # remove Species column
  scale() %>% # scale to zero mean
  # and unit variance
  prcomp() -> pca # do PCA
  pca # store result
  # in variable "pca"
\end{verbatim}
Doing a PCA in R

```r
> pca
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.5210659</td>
<td>-0.37741762</td>
<td>0.7195664</td>
<td>0.2612863</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.2693474</td>
<td>-0.92329566</td>
<td>-0.2443818</td>
<td>-0.1235096</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.5804131</td>
<td>-0.02449161</td>
<td>-0.1421264</td>
<td>-0.8014492</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.5648565</td>
<td>-0.06694199</td>
<td>-0.6342727</td>
<td>0.5235971</td>
</tr>
</tbody>
</table>
```
Doing a PCA in R

```r
> pca

Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.5210659</td>
<td>-0.37741762</td>
<td>0.7195664</td>
<td>0.2612863</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.2693474</td>
<td>-0.92329566</td>
<td>-0.2443818</td>
<td>-0.1235096</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.5804131</td>
<td>-0.02449161</td>
<td>-0.1421264</td>
<td>-0.8014492</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.5648565</td>
<td>-0.06694199</td>
<td>-0.6342727</td>
<td>0.5235971</td>
</tr>
</tbody>
</table>
Squares of the std. devs represent the % variance explained by each PC.
Doing a PCA in R

> pca

Standard deviations:
[1]  1.7083611  0.9560494  0.3830886  0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.5210659</td>
<td>-0.37741762</td>
<td>0.7195664</td>
<td>0.2612863</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.2693474</td>
<td>-0.92329566</td>
<td>-0.2443818</td>
<td>-0.1235096</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.5804131</td>
<td>-0.02449161</td>
<td>-0.1421264</td>
<td>-0.8014492</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.5648565</td>
<td>-0.06694199</td>
<td>-0.6342727</td>
<td>0.5235971</td>
</tr>
</tbody>
</table>
The rotation matrix tells us which variables contribute to which PCs.
We can also recover each original observation expressed in PC coordinates

> pca$x
We can also recover each original observation expressed in PC coordinates

```r
> pca$x

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.25714118</td>
<td>-0.478423832</td>
<td>0.127279624</td>
<td>0.024087508</td>
</tr>
<tr>
<td>2</td>
<td>-2.07401302</td>
<td>0.671882687</td>
<td>0.233825517</td>
<td>0.102662845</td>
</tr>
<tr>
<td>3</td>
<td>-2.35633511</td>
<td>0.340766425</td>
<td>-0.044053900</td>
<td>0.028282305</td>
</tr>
<tr>
<td>4</td>
<td>-2.29170679</td>
<td>0.595399863</td>
<td>-0.090985297</td>
<td>-0.065735340</td>
</tr>
<tr>
<td>5</td>
<td>-2.38186270</td>
<td>-0.644675659</td>
<td>-0.015685647</td>
<td>-0.035802870</td>
</tr>
<tr>
<td>6</td>
<td>-2.06870061</td>
<td>-1.484205297</td>
<td>-0.026878250</td>
<td>0.006586116</td>
</tr>
<tr>
<td>7</td>
<td>-2.43586845</td>
<td>-0.047485118</td>
<td>-0.334350297</td>
<td>-0.036652767</td>
</tr>
<tr>
<td>8</td>
<td>-2.22539189</td>
<td>-0.222403002</td>
<td>0.088399352</td>
<td>-0.024529919</td>
</tr>
<tr>
<td>9</td>
<td>-2.32684533</td>
<td>1.111603700</td>
<td>-0.144592465</td>
<td>-0.026769540</td>
</tr>
<tr>
<td>10</td>
<td>-2.17703491</td>
<td>0.467447569</td>
<td>0.252918268</td>
<td>-0.039766068</td>
</tr>
<tr>
<td>11</td>
<td>-2.15907699</td>
<td>-1.040205867</td>
<td>0.267784001</td>
<td>0.016675503</td>
</tr>
<tr>
<td>12</td>
<td>-2.31836413</td>
<td>-0.132633999</td>
<td>-0.093446191</td>
<td>-0.133037725</td>
</tr>
<tr>
<td>13</td>
<td>-2.21104370</td>
<td>0.726243183</td>
<td>0.230140246</td>
<td>0.002416941</td>
</tr>
</tbody>
</table>
Plot of iris plants in PC coordinates reveals differences among species.
These differences are much harder to see in the original variables.