Tutorial Business Analytics

Tutorial 10: Principal Component Analysis

Decision Sciences & Systems (DSS)

Department of Informatics

TU München

Tutorial Business Analytics

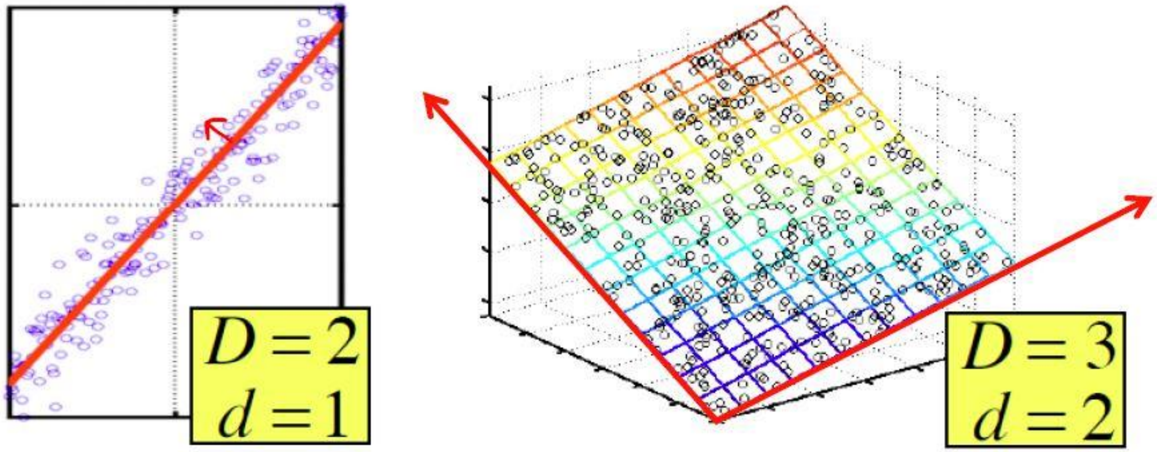
Agenda

- Dimensionality Reduction
- Principal Component Analysis
- PCA General Approach
- Reconstruction of Original Data

Tutorial Business Analytics

Dimensionality Reduction

- Reduce a complex dataset to a lower dimension
  - Simplify data understanding, visualization and manipulation (computation time!)
  - Reveal hidden underlying dynamics, e.g., latent variables, multicollinearity
  - Often the data lies on (or near) a **low dimensional** subspace



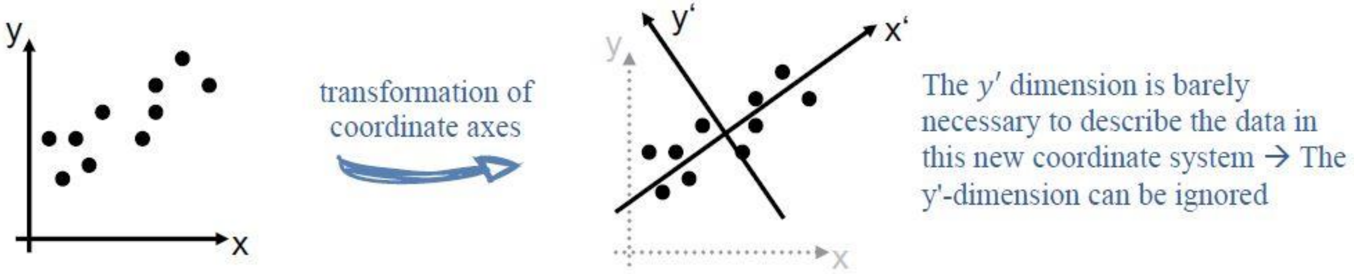
- We effectively need only d dimensions instead of D to describe the data!

Tutorial Business Analytics

Dimensionality Reduction

- Feature sub-selection
  - “Expert-driven” cut-off reduction (e.g., remove low-variance dimensions)
  - Features are often correlated → discarding whole features not always a good idea

- Linear transformations (PCA)
  - Linear transformation to represent data in a different coordinate system
  - Change of the basis (orthogonal basis transformations + potentially discarding dimensions)



The y' dimension is barely necessary to describe the data in this new coordinate system → The y'-dimension can be ignored

Tutorial Business Analytics

Principal Component Analysis

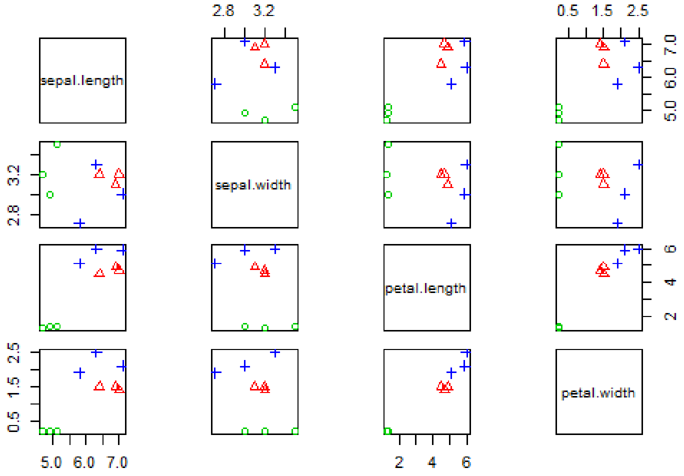
- **Goal**: Transform the data, such that the covariance between the new dimensions is 0 and we maximize the variance along the axes
  - Find a coordinate system in which the variables are linearly uncorrelated
  - The dimensions with no or low variance can then be ignored

- PCs → **Principal Components** allow us to summarize a large set of correlated variables, with a smaller number of representative variables that collectively explain most of the variability in the original set.
  - PC directions are directions in feature space along which the original data are **highly variable**
  - The first PC has the largest possible variance, and each succeeding component has the highest variance possible under the constraint that it is orthogonal to the preceding

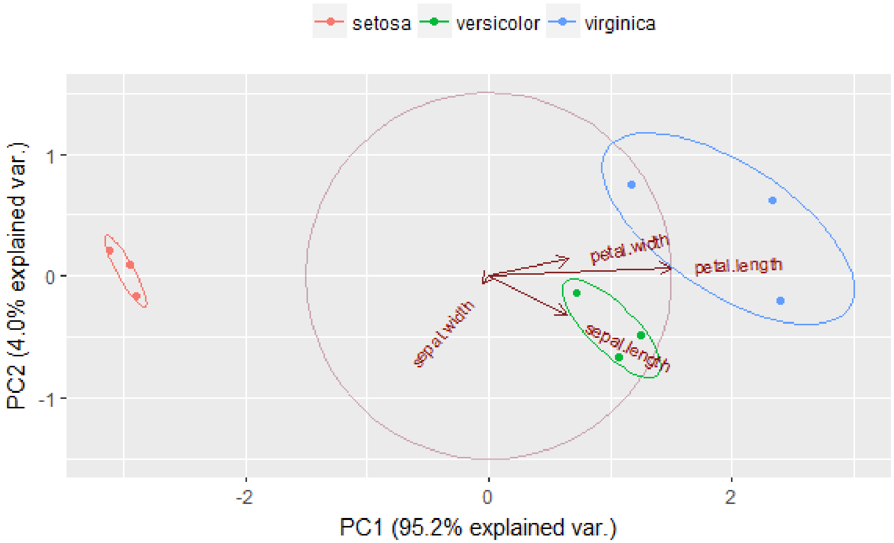
Tutorial Business Analytics

Principal Component Analysis Illustrated

- Visualization of the iris dataset – 4 numerical features (only 9 observations in the example)
  - Very hard to make sense of the data
  - Some features could be correlated



- After applying PCA we discover that the first two components explain more than 99% of the data variance. The 2D transformation is easy to visualize and automatically clusters the different species.

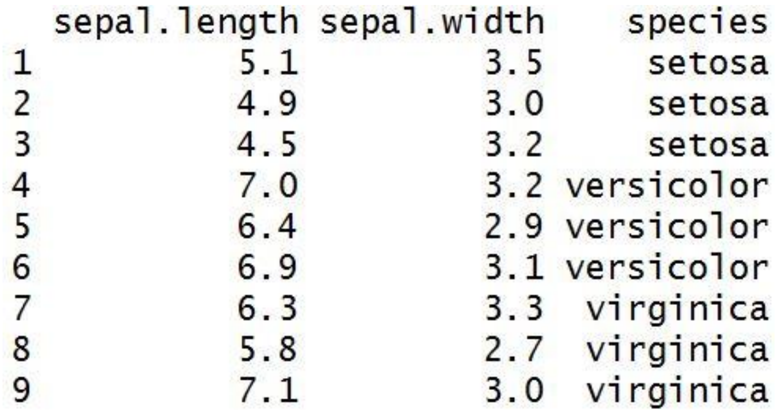


Tutorial Business Analytics

PCA General Approach

- Step 1: Center the data → subtract the mean from each data dimension
- Step 2: Compute the covariance matrix ∑ (or the correlation matrix)
- Step 3: Use the eigenvector decomposition to transform the coordinate system → find the eigenvectors of the covariance matrix and order them by the corresponding largest eigenvalues
- Step 4: Reduce dimensionality and form feature vector (principal components)
- Step 5: Derive the new data (projection on the subspace → PC scores)

- ***Example:*** Iris dataset
  - We will take only two features for ease of calculation
  - 9 observations – 3 for each species



```
  sepal.length sepal.width    species
1          5.1         3.5     setosa
2          4.9         3.0     setosa
3          4.5         3.2     setosa
4          7.0         3.2 versicolor
5          6.4         2.9 versicolor
6          6.9         3.1 versicolor
7          6.3         3.3  virginica
8          5.8         2.7  virginica
9          7.1         3.0  virginica
```

Tutorial Business Analytics

PCA Step 1: Center the data

- Calculate the mean of each data dimension: `sepal.length` (d_1) and `sepal.width` (d_2)

$$\bar{d}_j = \frac{1}{N}\sum_{i=1}^{N} d_{ij}$$

$$\bar{d}_1 = \frac{1}{9} \cdot (5.1 + 4.9 + 4.5 + 7.0 + 6.4 + 6.9 + 6.3 + 5.8 + 7.1) = \frac{1}{9} \cdot 54 = 6$$

$$\bar{d}_2 = 3.1$$

- We transform our dataset to a zero means dataset by subtracting the means:

$$x_j = d_j - \bar{d}_j \quad \Rightarrow \quad X = \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.5 & 3.2 \\ 7.0 & 3.2 \\ 6.4 & 2.9 \\ 6.9 & 3.1 \\ 6.3 & 3.3 \\ 5.8 & 2.7 \\ 7.1 & 3.0 \end{bmatrix} - \begin{bmatrix} 6 & 3.1 \\ 6 & 3.1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} -0.9 & 0.4 \\ -1.1 & -0.1 \\ -1.5 & 0.1 \\ 1 & 0.1 \\ 0.4 & -0.2 \\ 0.9 & 0 \\ 0.3 & 0.2 \\ -0.2 & -0.4 \\ 1.1 & -0.1 \end{bmatrix}$$

Tutorial Business Analytics

PCA Step 2: Compute the covariance matrix

- The covariance matrix of the centered dataset is computed by determining the variances $var(x_j)$ for each dimension and the covariance $cov(x_{j_1}, x_{j_2})$ between dimensions.

$$\Sigma_x = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & cov(x_1, x_3) \\ cov(x_2, x_1) & var(x_2) & cov(x_2, x_3) \\ cov(x_3, x_1) & cov(x_3, x_2) & var(x_3) \end{bmatrix}$$

- Given that the means of the feature vectors are now 0, we can use the following formulas:

$$var(x_j) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij}{}^2$$

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij_1} - \bar{x}_{j_1}) \cdot (x_{ij_2} - \bar{x}_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij_1} x_{ij_2}$$

Tutorial Business Analytics

PCA Step 2: Compute the covariance matrix

- Applying the formulas for our dataset:

$$\boldsymbol{var}(\boldsymbol{x_1}) = \frac{1}{9-1} \cdot \left((-0.9)^2 + (-1.1)^2 + (-1.5)^2 + 1^2 + 0.4^2 + 0.9^2 + 0.3^2 + (-0.2)^2 + 1.1^2\right)$$

$$= \tfrac{1}{8} \cdot (7.58) = 0.9475$$

$$\boldsymbol{var}(\boldsymbol{x_2}) = 0.055$$

$$\boldsymbol{cov}(\boldsymbol{x_1}, \boldsymbol{x_2}) = \tfrac{1}{9-1} \cdot ((-0.9) \cdot 0.4 + (-1.1) \cdot (-0.1) + (-1.5) \cdot 0.1 + 1 \cdot 0.1 + 0.4 \cdot (-0.2) + 0.9 \cdot 0 + 0.3 \cdot 0.2 + (-0.2) \cdot (-0.4) + 1.1 \cdot (-0.1))$$

$$= \tfrac{1}{8} \cdot (-\,0.35) = -0.04375$$

$$\boldsymbol{cov}(\boldsymbol{x_2}, \boldsymbol{x_1}) = cov(x_1, x_2)$$

- The covariance matrix: $\sum_x = \begin{bmatrix} 0.9475 & -0.04375 \\ -0.04375 & 0.055 \end{bmatrix}$

- Next: Transform the coordinate system so that the covariance in between the new axes is 0. According to the spectral theorem, the eigenvectors of a symmetric matrix form an orthogonal basis. The largest eigenvector of the covariance matrix always points into the direction of the largest variance of the data.

Tutorial Business Analytics

PCA Step 3: Calculate the eigenvalues and eigenvectors

- To compute the eigenvalues of the covariance matrix $\sum_x$ of size p, we need to solve the characteristic equation $\left|\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_p}\right| = 0$. First, we derive the characteristic polynomial of $\sum_{\boldsymbol{x}}$:

$$\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_2} = \begin{bmatrix} 0.9475 & -0.04375 \\ -0.04375 & 0.055 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.9475 - \lambda & -0.04375 \\ -0.04375 & 0.055 - \lambda \end{bmatrix}$$

Hence,

$$\begin{aligned} |\textstyle\sum_x - \lambda I_2| &= (0.9475 - \lambda)(0.055 - \lambda) - (-0.04375)(-0.04375) \\ &= 0.0521125 \; - \; \lambda(0.9475 + 0.055) + \lambda^2 - 0.0019140625 \\ &= \lambda^2 - 1.0025 \cdot \lambda + 0.0501984375 \end{aligned}$$

- Then, we solve the characteristic equation for λ:

$$\lambda^2 - 1.0025 \cdot \lambda + 0.0501984375 = 0$$

The roots of this equation will give us the two eigenvalues:

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{1.0025 \pm \sqrt{(-1.0025)^2 - 4 \cdot 1 \cdot 0.0501984375}}{2 \cdot 1} = \frac{1.0025 \pm \sqrt{0.8042125}}{2}$$

$$= \frac{1.0025 \pm 0.896779}{2} \qquad \Rightarrow \qquad \begin{cases} \lambda_1 = 0.94963948 \\ \lambda_2 = 0.05286052 \end{cases}$$

Tutorial Business Analytics

PCA Step 3: Calculate the eigenvalues and eigenvectors

- *Reminder* - How to compute the determinant of a 2×2 matrix:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

- *Reminder* - How to compute the determinant of a 3×3 matrix:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = \mathrm{a} \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

- *Reminder* - Laplace formula to compute the determinant of an $n \times n$ matrix A:

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} M_{i,j}$$

The minor $M_{i,j}$ is defined by the determinant of the $(n-1) \times (n-1)$ matrix that results from removing the i^{th} row and j^{th} column from A.

Tutorial Business Analytics

PCA Step 3: Calculate the eigenvalues and eigenvectors

- The corresponding eigenvectors are found by using these values of λ in the equation $(\Sigma_{\boldsymbol{x}} - \lambda \boldsymbol{I_p})v = 0$.

  For $\lambda_1 = 0.94963948$:

$$(\Sigma_x - 0.94963948\, I_2)v = 0 \quad \Rightarrow \begin{bmatrix} -0.00213948 & -0.04375 \\ -0.04375 & -0.8946395 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\Rightarrow \begin{cases} -0.00213948\, v_1 - 0.04375\, v_2 = 0 \\ -0.04375\, v_1 - 0.8946395 v_2 \quad = 0 \end{cases}$$

$$\Rightarrow \quad v_1 = -20.4489 v_2$$

  Thus, the eigenvectors of Σ_x corresponding to $\lambda_1 = 0.94963948$ are of the form $r \begin{bmatrix} -20.4489 \\ 1 \end{bmatrix}$, where r is a scalar.

- We constrain the eigenvector loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

$$\sum_{i=1}^{p} v_i^2 = 1 \quad \Rightarrow \quad \text{eigenvector}_1 = \begin{bmatrix} -0.99880642 \\ 0.04884401 \end{bmatrix}$$

Tutorial Business Analytics

PCA Step 3: Calculate the eigenvalues and eigenvectors

For $\lambda_2 = 0.05286052$:

$$(\Sigma_x - 0.05286052\, I_2)v = 0 \quad \Rightarrow \begin{bmatrix} 0.8946395 & -0.04375 \\ -0.04375 & 0.00213948 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$\Rightarrow \begin{cases} 0.8946395\, v_1 - 0.04375\, v_2 & = 0 \\ -0.04375\, v_1 + 0.00213948\, v_2 & = 0 \end{cases}$$

$$\Rightarrow v_2 = 20.4489 v_1$$

Thus, the eigenvectors of Σ_x corresponding to $\lambda_2 = 0.05286052$ are of the form $r\begin{bmatrix} 1 \\ 20.4489 \end{bmatrix}$, where r is a scalar.

$$\sum_{i=1}^{p} v_i^2 = 1 \quad \Rightarrow \quad \text{eigenvector}_2 = \begin{bmatrix} -0.04884401 \\ -0.99880642 \end{bmatrix}$$

- Resulting in $\text{eigenvectors} = \begin{bmatrix} -0.99880642 & -0.04884401 \\ 0.04884401 & -0.99880642 \end{bmatrix}$

- As expected, the two eigenvectors are orthogonal to each other: $\Phi_1 \Phi_2^T = 0$

Tutorial Business Analytics

PCA Step 4: Order the eigenvectors and select the principal components

- The eigenvector with the highest corresponding eigenvalue is the principal component of the dataset. We order the eigenvector by their eigenvalues, highest to lowest. This gives us the components in order of significance:

$$\lambda_1 = 0.94963948 > \lambda_2 = 0.05286052$$

- Our eigenvectors are already ordered; therefore, our principal component loading vectors are:

$$\Phi = \begin{bmatrix} -0.99880642 & -0.04884401 \\ 0.04884401 & -0.99880642 \end{bmatrix}$$

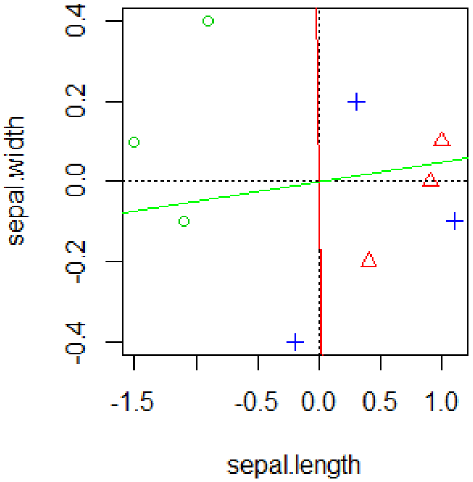
- We can now decide to leave out the component of lesser significance. For this, we calculate the variance explained by each component, using the eigenvalues:

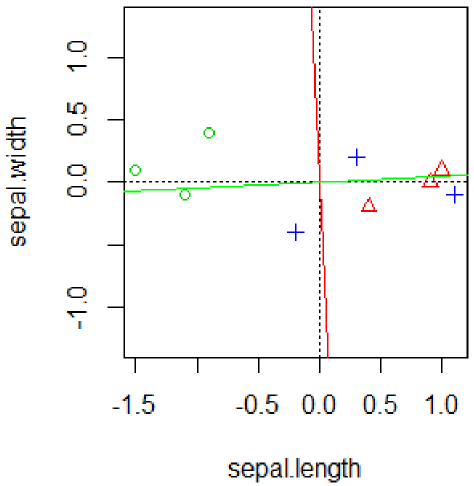
$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \quad \Rightarrow \quad \frac{0.94963948}{0.94963948 + 0.05286052} = \frac{0.94963948}{1.0025} > 94.7\% \quad \text{of the variance explained}$$

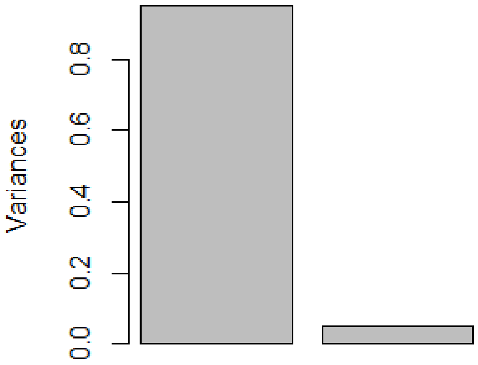
- Therefore, we can keep just the first component and reduce the dimensionality of our data down to 1.

Tutorial Business Analytics

PCA Step 4: Order the eigenvectors and select the principal components







<u>Above</u>: Plots of the eigenvectors on top of the centered original dataset X. In the second plot, we normalize the features.
<u>Below</u>: Plot of the variances explained by each component.

Tutorial Business Analytics

PCA Step 5: Project the transformed data

- The general formula for projecting the transformed data is:

$$Z = X\Phi$$

- For the 1D projection, we multiply the centered dataset with the first principal component:

$$Z = X\Phi_1 = \begin{bmatrix} -0.9 & 0.4 \\ -1.1 & -0.1 \\ -1.5 & 0.1 \\ 1 & 0.1 \\ 0.4 & -0.2 \\ 0.9 & 0 \\ 0.3 & 0.2 \\ -0.2 & -0.4 \\ 1.1 & -0.1 \end{bmatrix} \begin{bmatrix} -0.99880642 \\ 0.04884401 \end{bmatrix} = \begin{bmatrix} 0.9184634 \\ 1.0938027 \\ 1.5030940 \\ -0.9939220 \\ -0.4092914 \\ -0.8989258 \\ -0.2898731 \\ 0.1802237 \\ -1.1035715 \end{bmatrix}$$

Tutorial Business Analytics

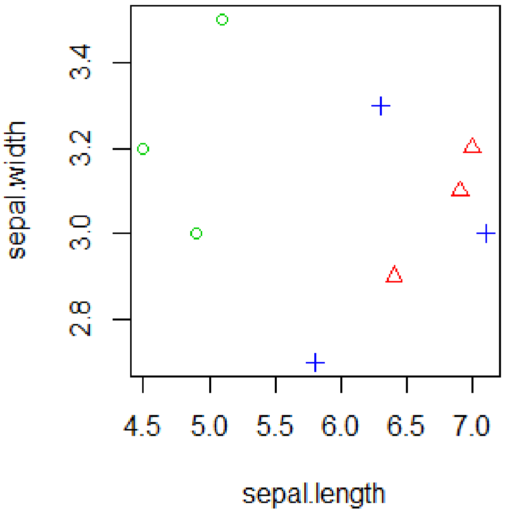
PCA Step 5: Project the transformed data

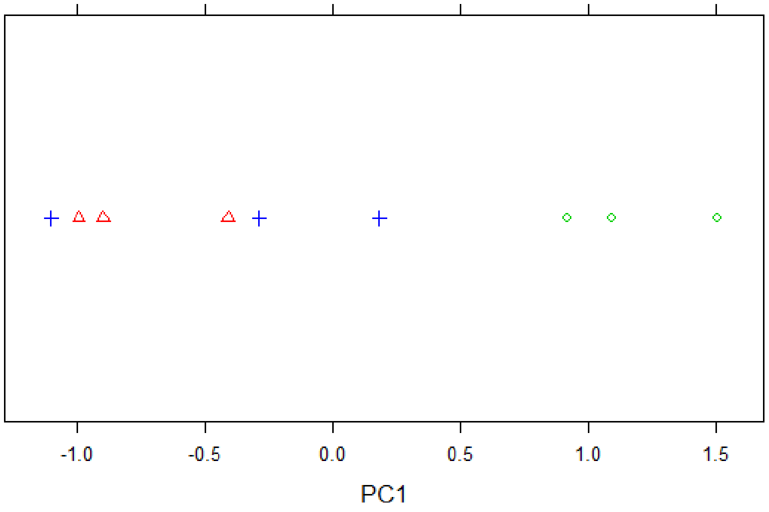
- For the 2D projection, we multiply the centered dataset with the both principal components:

$$Z = X\Phi = \begin{bmatrix} -0.9 & 0.4 \\ -1.1 & -0.1 \\ -1.5 & 0.1 \\ 1 & 0.1 \\ 0.4 & -0.2 \\ 0.9 & 0 \\ 0.3 & 0.2 \\ -0.2 & -0.4 \\ 1.1 & -0.1 \end{bmatrix} \begin{bmatrix} -0.99880642 & -0.04884401 \\ 0.04884401 & -0.99880642 \end{bmatrix} = \begin{bmatrix} 0.9184634 & -0.35556296 \\ 1.0938027 & 0.15360905 \\ 1.5030940 & -0.02661462 \\ -0.9939220 & -0.14872465 \\ -0.4092914 & 0.18022368 \\ -0.8989258 & -0.04395961 \\ -0.2898731 & -0.21441449 \\ 0.1802237 & 0.40929137 \\ -1.1035715 & 0.04615223 \end{bmatrix}$$

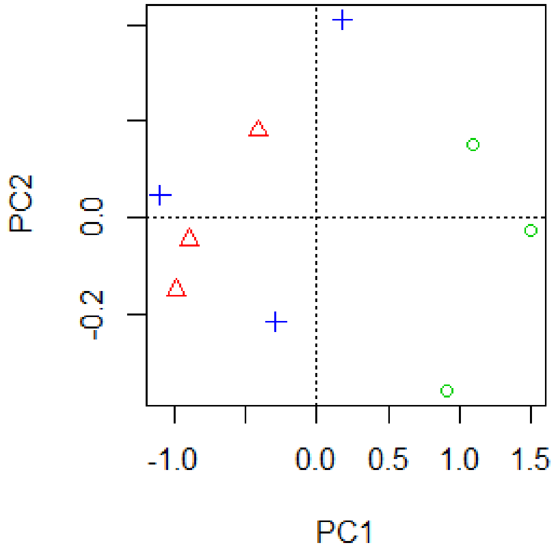
Tutorial Business Analytics

PCA Step 5: Project the transformed data





Plots of the original dataset and its 1D and 2D PCA projections



Tutorial Business Analytics

Reconstruction of Original Data

- To restore the original dataset, we multiply the projected data with the transposed eigenvectors and add the original dimension means:

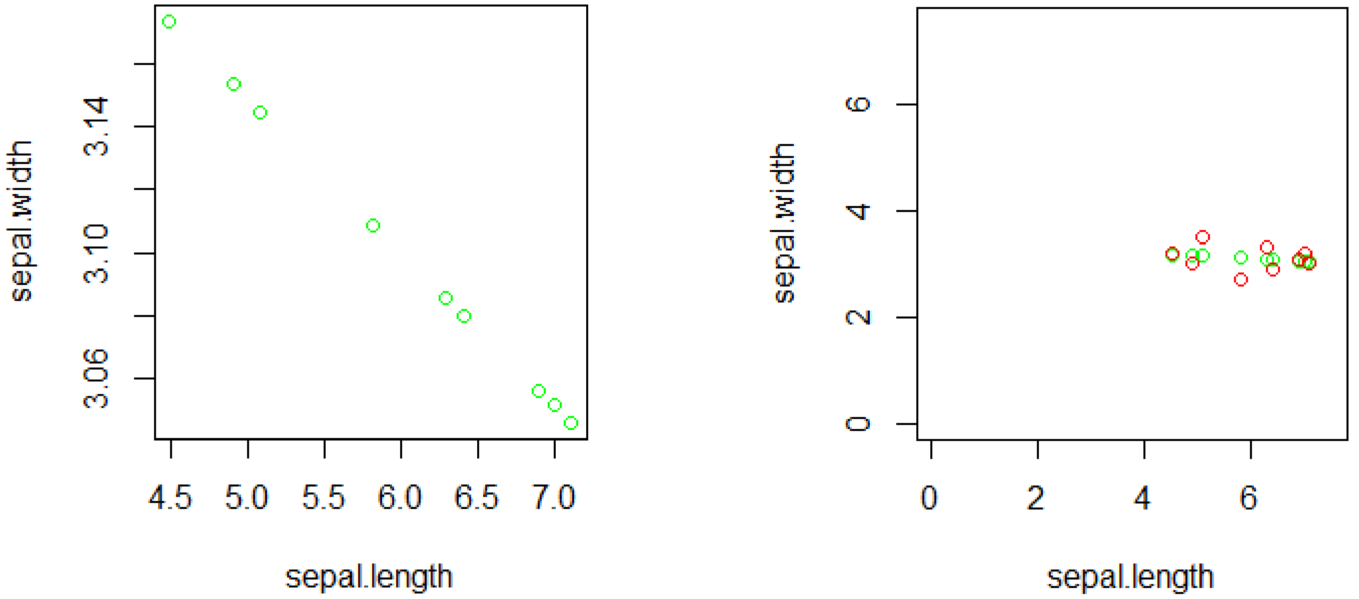
$$D \approx Z\Phi^T + \text{means}$$

- From the 1D projection:

$$D \approx \begin{bmatrix} 0.9184634 \\ 1.0938027 \\ 1.5030940 \\ -0.9939220 \\ -0.4092914 \\ -0.8989258 \\ -0.2898731 \\ 0.1802237 \\ -1.1035715 \end{bmatrix} \begin{bmatrix} -0.99880642 & 0.04884401 \end{bmatrix} + \begin{bmatrix} 6 & 3.1 \\ 6 & 3.1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} 5.082633 & 3.144861 \\ 4.907503 & 3.153426 \\ 4.4987 & 3.173417 \\ 6.992736 & 3.051453 \\ 6.408803 & 3.080009 \\ 6.897853 & 3.056093 \\ 6.289527 & 3.085841 \\ 5.819991 & 3.108803 \\ 7.102254 & 3.046097 \end{bmatrix}$$

Tutorial Business Analytics

Reconstruction of Original Data



- If we reduce the dimensionality, then, when reconstructing the data, we lose those dimensions we chose to discard. Nevertheless, the information loss is relatively small. In the above plots, we can see the reconstructed observations in green against the original ones in red.

Tutorial Business Analytics

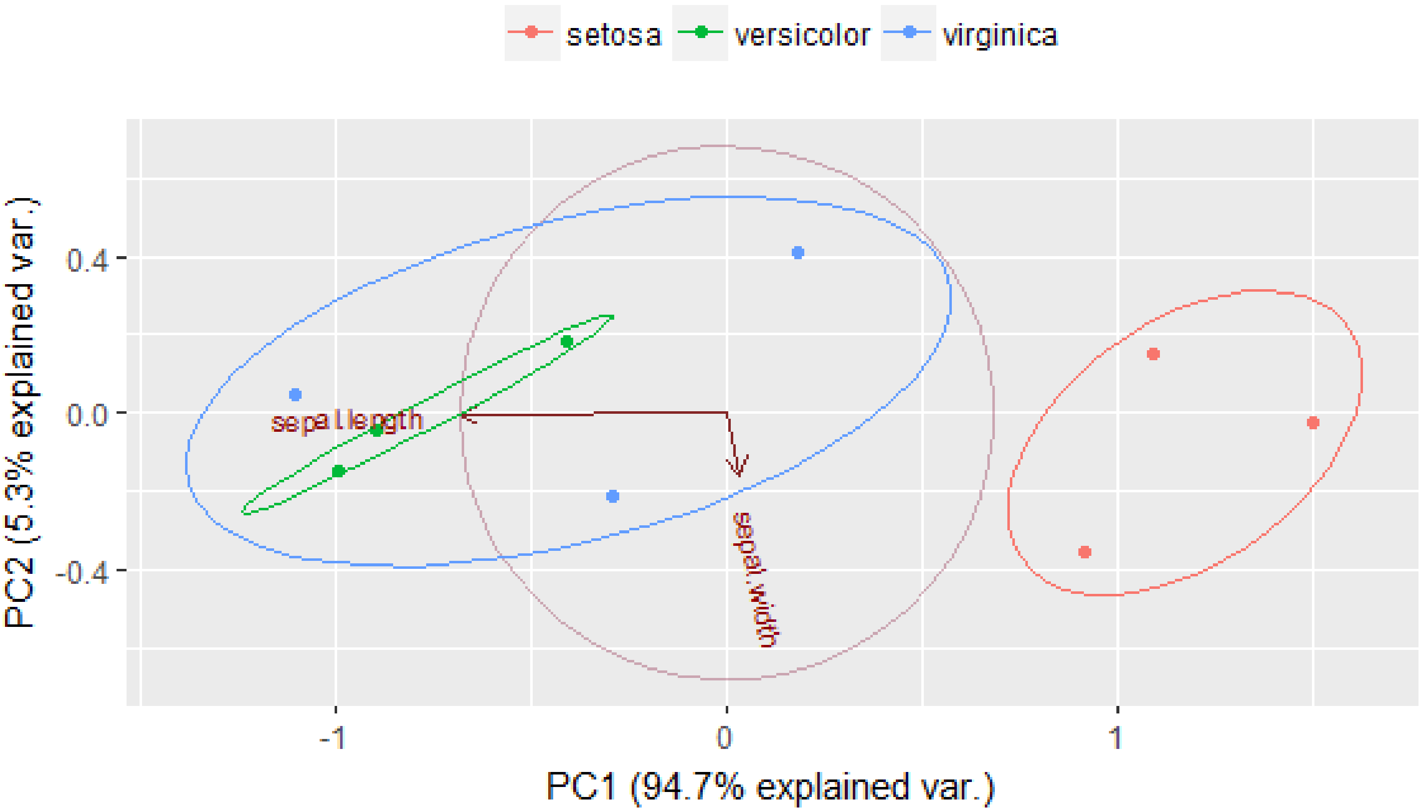
Reconstruction of Original Data

- From the 2D projection:

$$D \approx \begin{bmatrix} 0.9184634 & -0.35556296 \\ 1.0938027 & 0.15360905 \\ 1.5030940 & -0.02661462 \\ -0.9939220 & -0.14872465 \\ -0.4092914 & 0.18022368 \\ -0.8989258 & -0.04395961 \\ -0.2898731 & -0.21441449 \\ 0.1802237 & 0.40929137 \\ -1.1035715 & 0.04615223 \end{bmatrix} \begin{bmatrix} -0.99880642 & 0.04884401 \\ -0.04884401 & -0.99880642 \end{bmatrix} + \begin{bmatrix} 6 & 3.1 \\ 6 & 3.1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.5 & 3.2 \\ 7.0 & 3.2 \\ 6.4 & 2.9 \\ 6.9 & 3.1 \\ 6.3 & 3.3 \\ 5.8 & 2.7 \\ 7.1 & 3.0 \end{bmatrix} = D$$

Tutorial Business Analytics

Reconstruction of Original Data



Tutorial Business Analytics

Formulas Cheat Sheet

- Calculate the dimension means: $\bar{d}_j = \frac{1}{N}\sum_{i=1}^{N} d_{ij}$
- Subtract means: $x_j = d_j - \bar{d}_j$
- The covariance matrix: $\sum_x = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & cov(x_1, x_3) \\ cov(x_2, x_1) & var(x_2) & cov(x_2, x_3) \\ cov(x_3, x_1) & cov(x_3, x_2) & var(x_3) \end{bmatrix}$
- Calculate the covariance matrix:

$$cov\left(x_{j_1}, x_{j_2}\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij_1} - \bar{x}_{j_1}\right) \cdot \left(x_{ij_2} - \bar{x}_{j_2}\right) = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij_1}\, x_{ij_2}$$

- Find the eigenvalues by solving the characteristic equation: $\left|\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_p}\right| = 0$
- Calculate the eigenvectors: $\left(\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_p}\right) v = 0$
- Calculate the variance explained by each component: $\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$
- Projecting the transformed data: $Z = X\Phi$
- Restoring the original dataset: $D \approx Z\Phi^T + \text{means}$

Tutorial Business Analytics

Agenda

- Dimensionality Reduction
- Principal Component Analysis
- PCA General Approach
- Reconstruction of Original Data
- PCR Regression