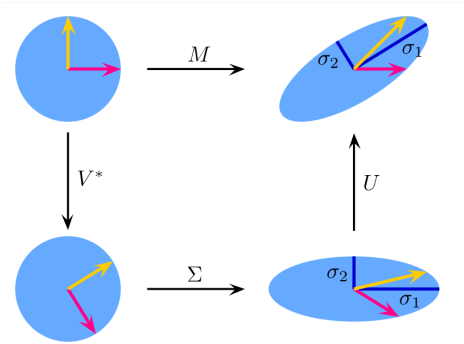
Kapitel 6

Hauptkomponentenanalyse und Singulärwertzerlegung



$$M = U \cdot \Sigma \cdot V^*$$

Inhalt

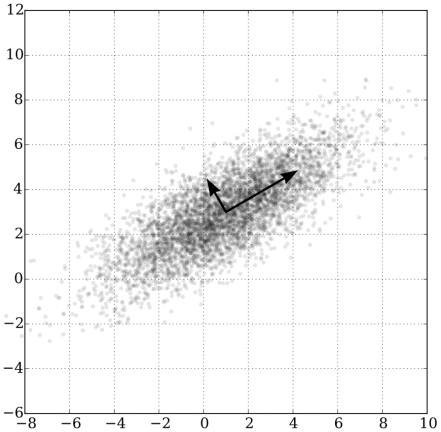
Hauptkomponentenanalyse: Idee (1)

- auch Principal Component Analysis oder kurz PCA
- Wir haben einen Datensatz mit p numerischen Merkmalen und n Beobachtungen.
- Wir repräsentieren den Datensatz in einer Matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$.
- Jede Zeile ist eine Beobachtung und stellt einen Punkt im $\mathbb{R}^p$ dar.
- Wir wollen nun die Beobachtungen in einen Raum niedriger Dimension überführen, also als Matrix $\mathbf{B} \in \mathbb{R}^{n \times q}$ mit $q < p$ darstellen.
- Dazu lassen wir aber nicht einfach Merkmale weg, sondern die neuen q Merkmale sollen jeweils eine geeignte Linearkombination der alten p Merkmale sein.
- Der Informationsverlust sollte gering sein.
- Wie sollen wir diese Linearkombinationen wählen?

Hauptkomponentenanalyse: Idee (2)

- Nutze Korrelationen zwischen den Merkmalen!
- Finde die Richtung mit maximaler Varianz im höherdimensionalen Raum und nutze diese Richtung für die Projektion in den Raum niedriger Dimension.
- Die Richtung der maximalen Varianz bildet die 1. Hauptkomponente.
- Die k-te Hauptkomponente ist orthogonal zu den Hauptkomponenten $1, \ldots, k-1$.
- Rechnerisch erhalten wir die Hauptkomponenten als Eigenvektoren der Kovarianzmatrix.
- Die zugehörigen Eigenwerte sagen etwas über den Anteil der Varianz in Richtung der Hauptkomponente in Bezug zur Gesamtvarianz aus.
- Wenn durch die ersten q Hauptkomponenten ein Großteil der Varianz abgedeckt ist, können wir die Daten mit geringem Informationsverlust in den $\mathbb{R}^q$ transformieren.

Beispiel: Hauptkomponenten einer zweidimensionalen Normalverteilung



- Mittelwert $(1, 3)$,
- Standardabweichung circa 3 in Richtung $(0.866, 0.5)^T$ und 1 in die dazu orthogonale Richtung.
- Die Vektoren sind die Eigenvektoren der Kovarianzmatrix und haben als Länge die Wurzel des zugehörigen Eigenwertes.

Kovarianz und Korrelationskoeffizient

Definition 6.1

Es seien X und Y zwei Zufallsvariablen, die auf dem gleichen Wahrscheinlichkeitsraum definiert sind. Weiterhin gelte $E(X) = \mu_X$ und $E(Y) = \mu_Y$.

Dann bezeichnet $\mathrm{Cov}(X, Y)$ die Kovarianz von X und Y, die definiert ist durch

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

Existieren die Standdardabweichungen σ_X und σ_Y für X bzw. Y und gilt $\sigma_X, \sigma_Y > 0$, dann ist

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

der Korrelationskoeffizient von X und Y.

Diskussion: Kovarianz und Korrelationskoeffizient

- Sowohl die Kovarianz als auch der Korrelationskoeffizient messen die lineare Beziehung zwischen X und Y.
- Die Kovarianz $\mathrm{Cov}(X, Y)$ ist positiv, wenn $X - \mu_X$ und $Y - \mu_Y$ dazu tendieren, mit hoher Wahrscheinlichkeit das gleiche Vorzeichen zu haben.
- Die Kovarianz $\mathrm{Cov}(X, Y)$ ist negativ, wenn $X - \mu_X$ und $Y - \mu_Y$ dazu tendieren, mit hoher Wahrscheinlichkeit verschiedene Vorzeichen zu haben.
- Die Größe von $\mathrm{Cov}(X, Y)$ ist relativ bedeutungslos, da sie von der Varianz von X und Y abhängt.
- Der Korrelationskoeffizient eliminiert diese individuelle Variabilität.
- Es gilt

$$-1 \leq \rho_{X,Y} \leq 1.$$

Stichprobenkovarianz (1)

- Kovarianz und Korrelationskoeffizient sind theoretische Größen, deren Definition auf den Verteilungen von X und Y basiert.
- In der Datenanalyse kennen wir diese Verteilungen üblicherweise nicht. Wir sehen nur eine Stichprobe dieser Verteilungen.
- Wir behelfen uns, indem wir die Kovarianz oder den Korrelationskoeffizienten schätzen, z. B. durch die Stichprobenkovarianz.

Stichprobenkovarianz (2)

Definition 6.2

Ist $(x_1, y_1), \ldots, (x_n, y_n)$ eine Stichprobe zweier Zufallsvariablen X und Y, dann ist die Stichprobenkovarianz $s_{x,y}$ definiert durch

$$s_{x,y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

mit den arithmetischen Mittelwerten $\bar{x} = \sum_{i=1}^{n} x_i$ und $\bar{y} = \sum_{i=1}^{n} y_i$.

Diskussion: Stichprobenkovarianz

- Es gibt noch die korrigierte Stichprobenkovarianz $\hat{\sigma}_{x,y}$ mit

$$\hat{\sigma}_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- Die Stichprobenkovarianz entspricht dem Maximum-Likelihood-Schätzer

$$S_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

  für $\mathrm{Cov}(X, Y)$, aber dieser Schätzer ist nicht erwartungstreu.

- Konkret gilt

$$E(S_{X,Y}) = \frac{n-1}{n} \mathrm{Cov}(X, Y).$$

- Die korrigierte Stichprobenkovarianz ist jedoch erwartungstreu.
- In der induktiven Statistik wird daher die korrigierte Stichprobenkovarianz genutzt.
- Die Formel für die Stichprobenkovarianz kann verwendet werden, wenn ein ganzer Datensatz betrachtet wird und die Erwartungswerte bekannt oder ausreichend gut geschätzt werden können.
- Insbesondere bei wenigen Messwerten sollte man dagegen die Formel für die korrigierte Stichprobenkovarianz nutzen.

Kovarianzmatrix

- Für n Zufallsvariablen $X_1, \ldots, X_n$ heißt die Matrix $\mathbf{\Sigma} = (\sigma_{i,j}) \in \mathbb{R}^{n \times n}$ mit

$$\sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$$

  Kovarianzmatrix.
- Für eine Datenanalyse eines Datensatzes nutzen wir die Stichprobenkovarianzmatrix $\mathbf{S} = (s_{i,j}) \in \mathbb{R}^{n \times n}$ mit

$$s_{i,j} = s_{x_i, x_j}.$$

- Die Stichprobenkovarianzmatrix besteht also aus den Stichprobenkovarianzen der einzelnen Merkmale.

Beispiel 6.3

Datensatz:

x_1	x_2
1.0	1.41
2.0	1.56
2.0	2.19
4.0	2.79
5.0	3.04
6.0	2.23
9.0	3.74
9.0	3.84
9.0	2.80
13.0	4.18

Geschätze Erwartungswerte:

$$\bar{x}_1 = 6 \quad \text{und} \quad \bar{x}_2 = 2.78$$

Stichprobenkovarianzmatrix:

$$\mathbf{S} = \begin{pmatrix} 13.80 & 2.97 \\ 2.97 & 0.81 \end{pmatrix}$$

Vorgehen zur Hauptkomponentenanalyse

1. Zentriere die Daten
   Man subtrahiere von jedem der p Merkmal den Erwartungswert (geschätzt durch das arithmetische Mittel).
2. Optional: Normiere die Daten bzgl. der Varianz
   Berechne für jedes Merkmal die Standardabweichung und dividiere die Merkmalswerte durch diese.
3. Berechne die Stichprobenkovarianzmatrix $\mathbf{S}$
4. Berechne Eigenwerte und zugehörige Eigenvektoren von $\mathbf{S}$
   Es seien $\lambda_1, \ldots, \lambda_p$ die Eigenwerte. Dann gibt

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

den Anteil der zugehörigen Hauptkomponente an der Gesamtvarianz an.

5. Sortiere die Eigenwerte absteigend nach ihren Betragswerten
   Zusammen mit den zugehörigen Eigenvektoren ergeben sich dadurch die Hauptkomponenten.
   Es lohnt sich, hier den Anteil der Hauptkomponente an der Gesamtvarianz auszugeben.
6. Optional: Verwerfe die betragsmäßig kleinsten Eigenwerte und -vektoren und transformiere die Daten
   - Baue aus den verbleibenden Eigenvektoren eine Transformationsmatrix $\mathbf{W}$.
   - Generiere durch
     $$\mathbf{x}' = \mathbf{W}x$$
     einen neuen Datensatz mit niedriger Dimension. Dabei ist $\mathbf{x}$ ein Datenpunkt im alten zentrierten oder normierten Datensatz.

Hauptkomponentenanalyse für den IRIS Datensatz

Klassifikation von Schwertlilien

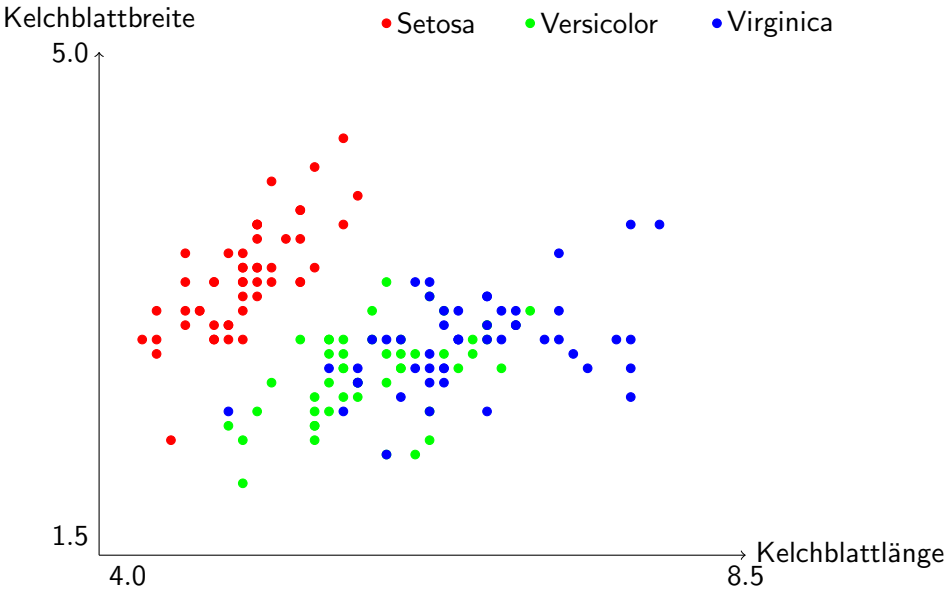
Attribute:

- Kelchblattlänge
- Kelchblattbreite
- Kronblattlänge
- Kronblattbreite
- Art

Alle Längen und Breiten in cm.
$n = 150$
$p = 4$

Auszug aus den Daten:

5.1	3.5	1.4	.2	Setosa
4.9	3	1.4	.2	Setosa
⋮	⋮	⋮	⋮	⋮
7	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	Versicolor
⋮	⋮	⋮	⋮	⋮
6.3	3.3	6	2.5	Virginica
5.8	2.7	5.1	1.9	Virginica


Kelchblattbreite
Setosa
Versicolor
Virginica
5.0
1.5
4.0
8.5
Kelchblattlänge

Der Datensatz sei als Matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ gegeben.

1. Mittelwerte der einzelnen Merkmale:

Kelchblattlänge $\bar{x}_1$	Kelchblattbreite $\bar{x}_2$	Kronblattlänge $\bar{x}_3$	Kronblattbreite $\bar{x}_4$
5.843333	3.057333	3.758000	1.199333

   Damit zentrieren wir den Datensatz, d. h. wir bilden die Matrix $\mathbf{B} = (b_{i,j}) \in \mathbb{R}^{n \times p}$ mit

$$b_{i,j} = a_{i,j} - \bar{x}_j.$$

2. Keine Normierung der Daten bzgl. der Varianz.

3. Da die Daten nun zentriert sind, nutzen wir konkret die Formel

$$s_{i,j} = \frac{1}{n} \sum_{k=1}^{n} b_{k,i} b_{k,j}$$

zur Berechnung der Stichprobenkovarianz von Merkmal i und Merkmal j.

Damit ergibt sich die Stichprobenkovarianzmatrix

$$\mathbf{S} = \begin{pmatrix} 0.6811 & -0.0422 & 1.2658 & 0.5128 \\ -0.0422 & 0.1887 & -0.3275 & -0.1208 \\ 1.2658 & -0.3275 & 3.0955 & 1.2870 \\ 0.5128 & -0.1208 & 1.2870 & 0.5771 \end{pmatrix}.$$

4. Wir berechnen mit dem QR-Verfahren die Eigenwerte und Eigenvektoren von $\mathbf{S}$.
   Für eine einzelne QR-Zerlegung nutzen wir das Gram-Schmidt-Verfahren.
   Die Eigenwerte sind

$$\lambda_1 = 4.2001, \quad \lambda_2 = 0.2411, \quad \lambda_3 = 0.0777, \quad \lambda_4 = 0.0237.$$

   Der Anteil der Hauptkomponenten an der Varianz beträgt

$$92.46\%, \quad 5.31\%, \quad 1.71\%, \quad 0.52\%.$$

   Die Eigenvektoren und damit die Hauptkomponenten sind

1. HK	2. HK	3. HK	4. HK
0.3614	0.6566	−0.5820	0.3155
−0.0845	0.7302	0.5979	−0.3197
0.8567	−0.1734	0.0762	−0.4798
0.3583	−0.0755	0.5458	0.7537

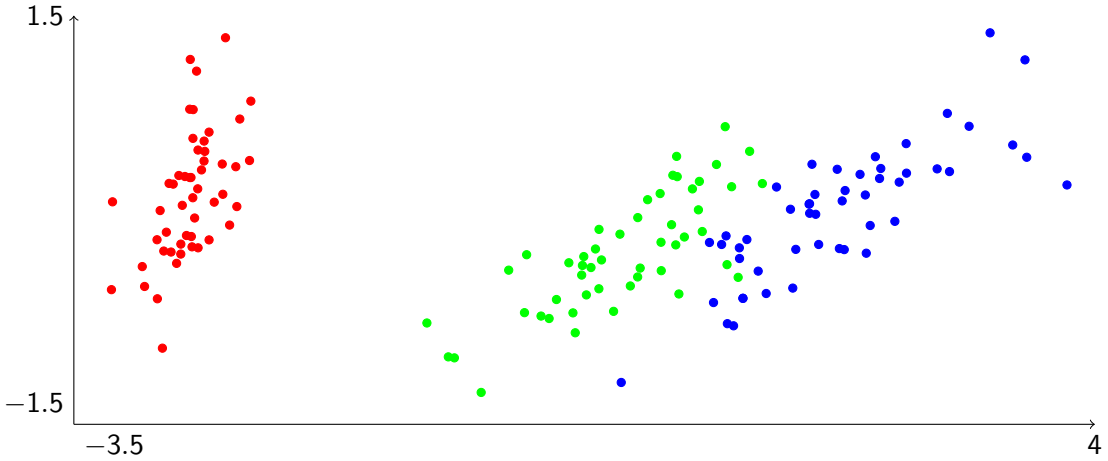
5. Die Eigenwerte und die zugehörigen Eigenvektoren sind zufälliger Weise schon richtig sortiert.
6. Wir nehmen die ersten beiden Hauptkomponenten ($q = 2$), da sie zusammen fast 98% der Varianz erklären.
   Mit diesen beiden Hauptkomponenten definieren wir die Transformatrix

$$\mathbf{W} = \begin{pmatrix} 0.3614 & 0.6566 \\ -0.0845 & 0.7302 \\ 0.8567 & -0.1734 \\ 0.3583 & -0.0755 \end{pmatrix} \in \mathbb{R}^{p \times q}$$

   und bauen mit

$$\mathbf{C} = \mathbf{BW}$$

   den transformierten Datensatz $\mathbf{C} \in \mathbb{R}^{n \times q}$.


1.5
−1.5
−3.5
4