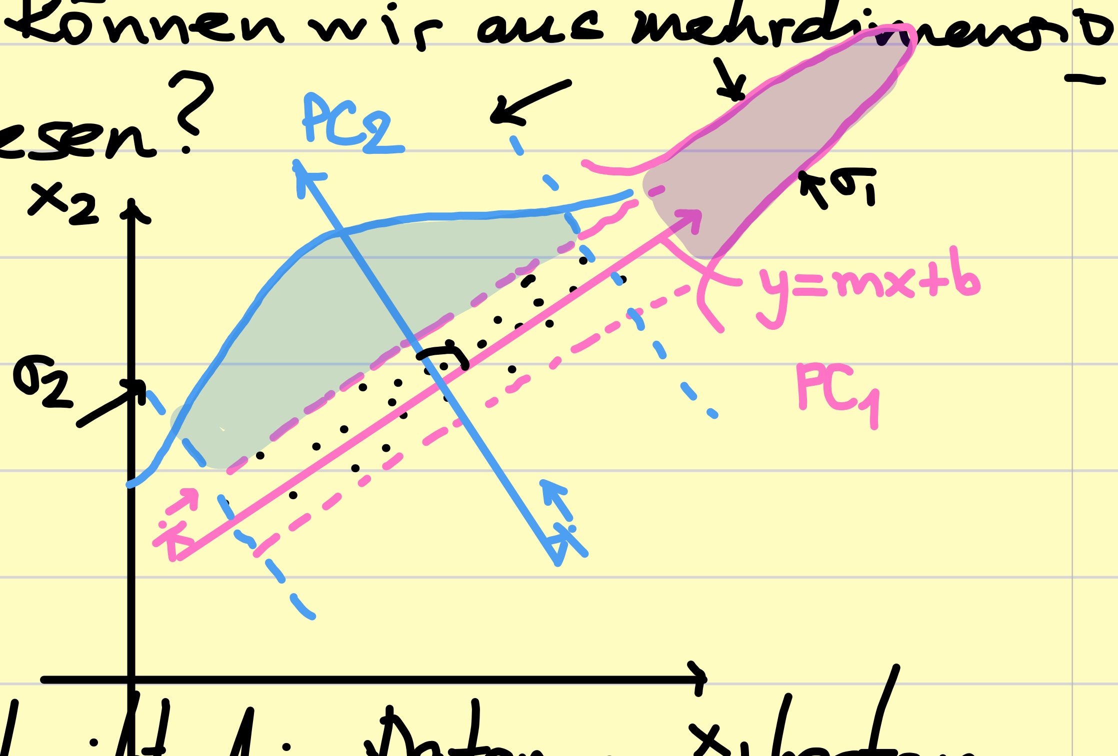
Hauptkomponentenanalyse

Principal Component Analysis (PCA)

Wie viel Variabilität können wir aus mehrdimensionalen Daten herauslesen?

· Graphische Intuition:



· Der Vektor → beschreibt die Daten am besten.

· Diese Richtung in dem die Variabilität minimal ist, erklärt ~~die~~ Daten am besten und heißt Hauptkomponente #1. (PC1)

· SENKRECHT zu PC1 ist PC2. PC2 erklärt wenigere Variabilität ($\sigma_2 \gg \sigma_1$) von den Daten

---

> Die Eigenvektoren der Kovarianzmatrix (normiert) sind die Hauptkomponenten eines Datensatzes.

Eigenvektoren einer Matrix A werden f.m. ermittelt:

$$\boxed{A\cdot\vec{v} = \lambda\cdot\vec{v}} \qquad \vec{v}: \text{Eigenvektoren} \quad \lambda: \text{Eigenwerte}$$

$$^{1)}\ \det(A-\lambda I)=0 \xrightarrow{2)} \lambda \xrightarrow{3)} \vec{v}$$

Beispiel: gegeben wird ein Kennzahlensystem mit 2 Kennzahlen (DLZ (Tage) und Qualität (ppm)). Ermitteln Sie die erste Hauptkomponente des Systems.

1. SCHRITT. Normierung der Daten

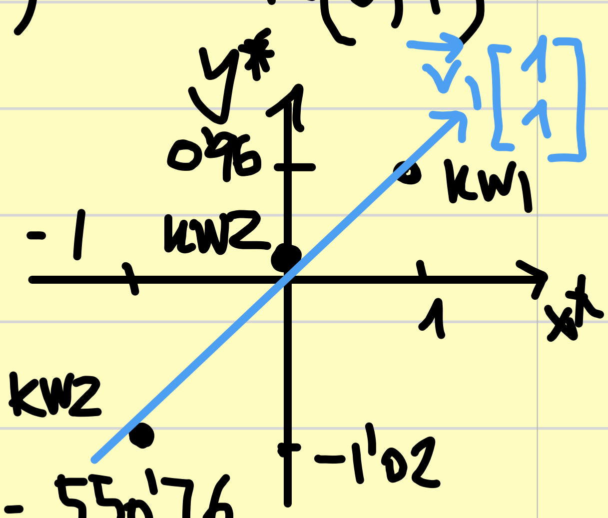
	DLZ (T)	Q (ppm)		DLZ^* (T)	Q^* (ppm)
KW1	17	3200		$\frac{17-15}{2}=1$	$\frac{3200-2666'67}{550'76}=0'96$
KW2	15	2700	$\rightarrow$	$\frac{15-15}{2}=0$	$\frac{2700-2666'67}{550'76}=0'06$
KW3	13	2100		$\frac{13-15}{2}=-1$	$\frac{2100-2666'67}{550'76}=-1'02$
				$N(0,1)$	$N(0,1)$

$$\bar{x}=\frac{17+15+13}{3}=15$$

$$\bar{y}=\frac{3200+2700+2100}{3}=2666'67$$

$$\sigma_x=\sqrt{\frac{(17-15)^2+(15-15)^2+(13-15)^2}{2}}=2$$

$$\sigma_y=\sqrt{\frac{(3200-2666'67)^2+(2700-2666'67)^2+(2100-2666'67)^2}{2}}=550'76$$



$$\overset{*}{A} = \begin{bmatrix} VAR(\overset{*}{X}) & KOV(\overset{*}{X},\overset{*}{Y}) \\ KOV(\overset{*}{X},\overset{*}{Y}) & VAR(\overset{*}{Y}) \end{bmatrix} = \begin{bmatrix} 1 & 0'99 \\ 0'99 & 1 \end{bmatrix}$$ Kov. Matrix normiert.

$$kov[\overset{*}{X},\overset{*}{Y}] = \frac{\sum(\overset{*}{x_i}-\overline{\overset{*}{x}})(\overset{*}{y_i}-\overline{\overset{*}{y}})}{n-1} = \frac{\sum \overset{*}{x_i}\overset{*}{y_i}}{n-1} = \frac{(1\cdot 0'96)+(-1\cdot -1'02)}{2} = 0'99$$

Schritt 2. $\det(\overset{*}{A}-\lambda I)=0$

$$\det\left(\begin{bmatrix} 1 & 0'99 \\ 0'99 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0 \rightarrow \det\left(\begin{bmatrix} 1-\lambda & 0'99 \\ 0'99 & 1-\lambda \end{bmatrix}\right) = 0 \rightarrow$$

$$\rightarrow (1-\lambda)^2 - 0'99^2 = 0 \rightarrow \lambda^2 - 2\lambda + 1 - 0'99^2 = 0 \rightarrow *$$

$$\boxed{\begin{aligned} &\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\cdot d - b\cdot c \\ &\det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + dhc - ceg - bdf - afh \end{aligned}}$$

$$* \rightarrow \lambda^2 - 2\lambda + 0'0199 = 0 \rightarrow \lambda = \frac{2 \pm \sqrt{2^2 - 4.1.0'0199}}{2} \rightarrow$$

$$\boxed{\begin{aligned} &ax^2+bx+c=0 \\ &x = \frac{-b \pm \sqrt{b^2-4ac}}{2a} \end{aligned}}$$

$$\rightarrow \lambda = \frac{2 \pm 1'98}{2} = \begin{cases} + \nearrow \lambda_1 = 1'99 \\ - \searrow \lambda_2 = 0'01 \end{cases}$$

Schritt 3. $\lambda_1 = 1'99$

$$A^* \cdot \vec{v_1} = \lambda_1 \cdot \vec{v_1} \rightarrow \begin{bmatrix} 1 & 0'99 \\ 0'99 & 1 \end{bmatrix} \cdot \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 1'99 \cdot \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \rightarrow$$

$$\rightarrow 1 \cdot v_{11} + 0'99\, v_{12} = 1'99\, v_{11} \rightarrow 0'99\, v_{12} = 0'99\, v_{11}$$

$$0'99\, v_{11} + 1 \cdot v_{12} = 1'99\, v_{12}$$

$$\rightarrow v_{11} = 1 \rightarrow v_{12} = 1 \rightarrow \boxed{\vec{v_1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}}$$

(wird gesetzt)

20p **Übung**: Gegeben wird ein Kennzahlensystem mit 2 ~~Kennzahlen~~: DLZ(X) und Output (Y).

9p a) Ermitteln Sie die normierte Kovarianzmatrix.

9p b) Ermitteln Sie die Eigenvektoren & -werte der Kovarianzmatrix.

2p c) Interpretieren Sie die Ergebnisse.

- DLZ(X): [17, 14, 12, 13, 9, 7] (Tage)
- OUTPUT(Y): [200, 250, 270, 240, 310, 330] (Stk/Zeit)
