Karlsruher Institut für Technologie

Geodatenanalyse I: Multivariate Statistik

Kathrin Menberg

Stundenplan

Vorläufiger Stundenplan		
Datum	**Thema**	**Dozent**
20.10.2021	Einführung in die Programmierung mit *Python*	Gabriel Rau
25.10.2021	Univariate Statistik und statistisches Testen	Kathrin Menberg
01.11.2021	*Feiertag*	
08.11.2021	Umgang und Berechnung von Datensätzen	Gabriel Rau
15.11.2021	Bivariate und schließende Statistik	Kathrin Menberg
22.11.2021	Datenvisualisierung mit *matplotlib*	Gabriel Rau
29.11.2021	Multivariate Statistik	Kathrin Menberg
06.12.2021	Datenformate, Datenspeicherung und Datenbanken	Gabriel Rau
13.12.2021	Monte-Carlo Methoden	Kathrin Menberg
20.12.2021	Analyse und Visualisierung von Geodaten	Gabriel Rau
27.12.2021	*Weihnachtsferien*	
03.01.2022	*Weihnachtsferien*	
10.01.2022	Sensitivitätsanalyse	Kathrin Menberg
17.01.2022	Datenethik, Lizensierung und Entwicklungstools	Gabriel Rau
24.01.2022	Räumliche Interpolation	Kathrin Menberg
31.01.2022	Fragen zur Programmierung	Gabriel Rau
07.02.2022	Regressionsanalyse	Kathrin Menberg

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:30	**Multivariate Statistik**
10:30 – 11:15	Übung
11:15 – 11:30	Pause
11:30 – 12:15	Fortsetzung Übung
12:15 – 12:30	Diskussion und Reflexion

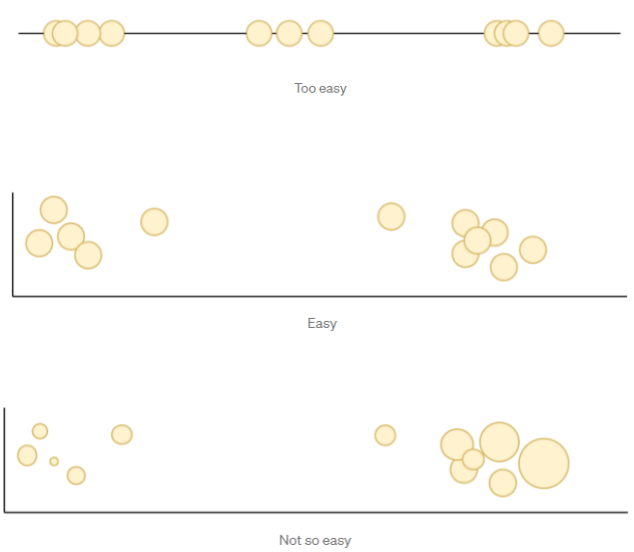
Lernziele

Am Ende der Stunde werden die Teilnehmer:

- … mit den statistischen Konzepten der Datentransformation, Eigenvektoren und Eigenwerten vertraut sein.
- … Methoden zur Reduzierung von Dimensionen auf Geodatensätze anwenden und die Ergebnisse graphisch darstellen können.

n-dimensionale Datensätze

- Beziehungen zwischen allen Parametern
  - Parameterraum (parameter space)
- Gemeinsame graphische Darstellung von vielen Parametern schwierig
  - Erkennen von Mustern usw.



https://towardsdatascience.com/pca-principal-component-analysis-explained-visually-in-5-minutes-20ce8a9ebf0f

Multivariate Statistik

- Dimensionen von Datensätzen reduzieren ohne viel Information zu verlieren

- 2D – Visualisierung von komplexen Beziehungen
  - **Hauptkomponentenanalyse** (principal component analysis)
  - Faktorenanalyse (factor analysis)
  - Unabhängigkeitsanalyse (Independent Component Analysis)

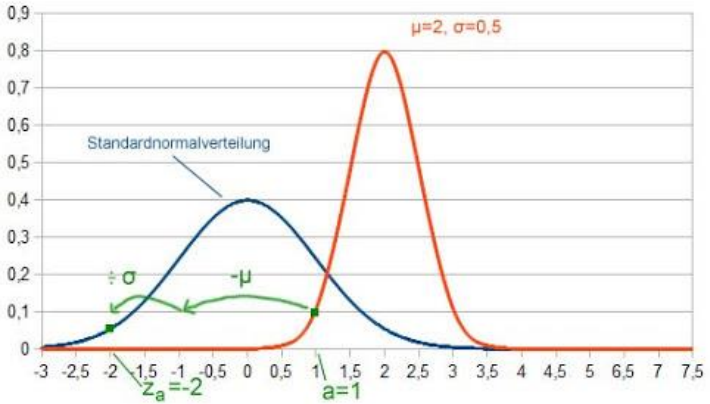
- Datenpunkte mit ähnlichen Eigenschaften identifizieren
  - Clusteranalyse (cluster analysis)
  - k-Means Algorithmus

Transformieren von Datensätzen

- Rohdaten oft nicht normalverteilt, Unterschiede in Varianzen zwischen einzelnen Parametern, usw.

- Standardisieren von Daten
- Standard-Normalverteilung
  - $X \sim N(0,1)$

$$standardized\ x_i = \frac{x_i - \bar{x}}{std\ (x)}$$



Hauptkomponentenanalyse

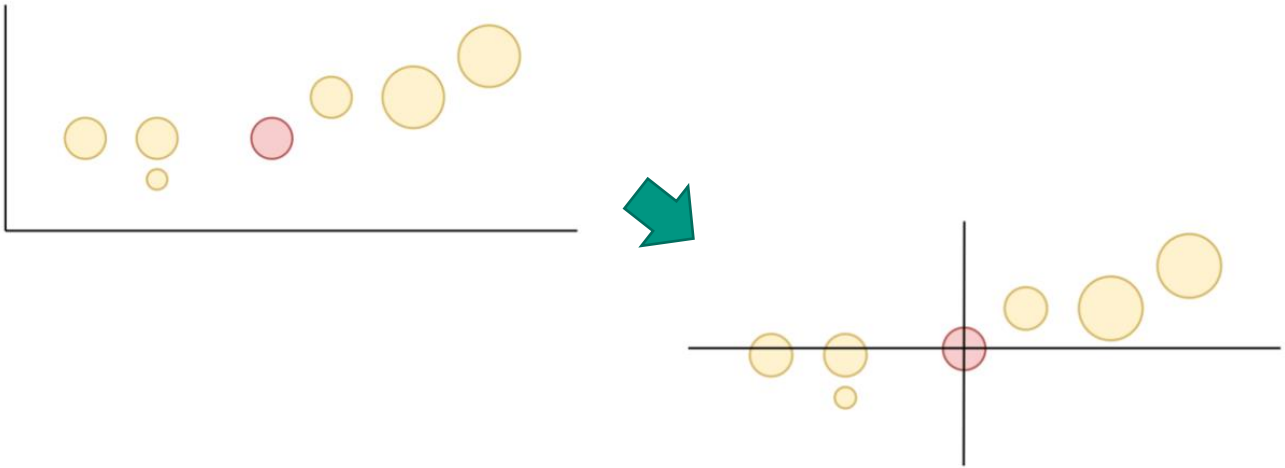
- Ziel: Reduzieren von Dimensionen (meist zur Visualisierung)
- Beispiel 3D → 2D:

x	y	z
1	2	2
2	2	2
2	1	1
4	3	2
5	3	3
6	4	3

https://towardsdatascience.com

Hauptkomponentenanalyse

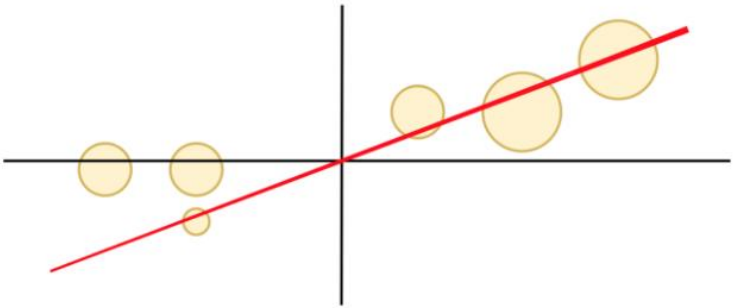
- 1. Alle Datenpunkte standardisieren
  - Mittelwerte berechnen und Datenpunkte „zentrieren“



https://towardsdatascience.com

Hauptkomponentenanalyse

- 2. Linie (bzw. Achse) mit der besten Übereinstimmung finden

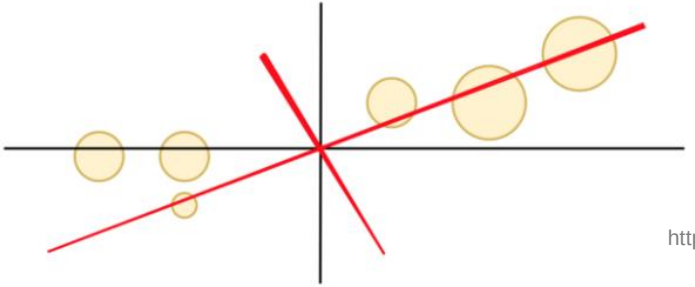


https://towardsdatascience.com

- erste Hauptkomponente (PC1)
- Linearkombination aus *x*, *y* und *z*
- Erklärt einen signifikanten Anteil der Varianz im Datensatz

Hauptkomponentenanalyse

- 3. zweite Hauptkomponente finden

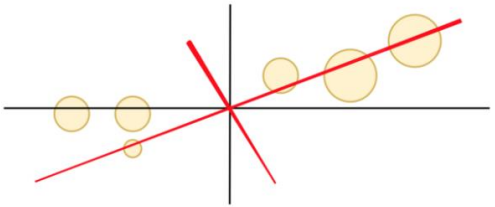


https://towardsdatascience.com

- Beste Übereinstimmung im rechten Winkel zu PC1
- Ebenso Linearkombination aus *x*, *y* und *z*
- Erklärt einen kleineren Anteil der Varianz im Datensatz als PC1

Hauptkomponentenanalyse

- Mathematische Lösung für n-dimensionale Datensätze über lineare Algebra

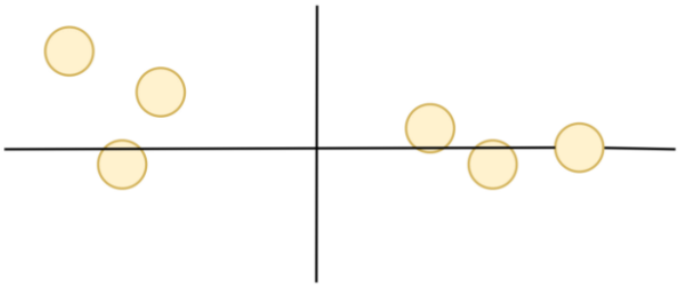


https://towardsdatascience.com

- Hauptkomponenten als Achsen (bzw. Vektoren) im Parameterraum
  - Betrag des Vektors: Eigenwert (PC1 > PC2, …) (Skalar)
  - Richtung des Vektors: Eigenvektor (*n* * 1, Vektor)

Hauptkomponentenanalyse

- 4. Rotation des Parameterraums auf die identifizierten Achsen



https://towardsdatascience.com

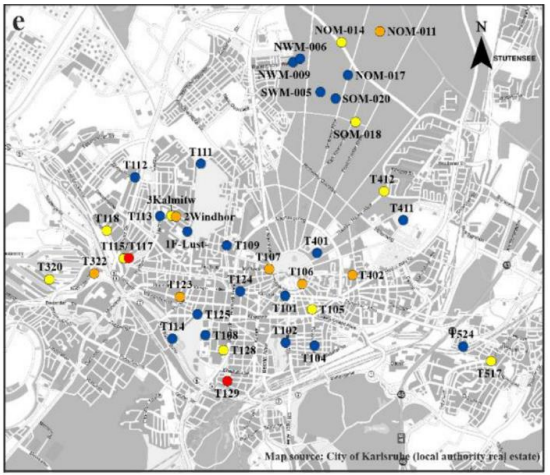
- Matrixmultiplikation:

  standardisierte Datenpunkte × Eigenvektoren (PC1, PC2)

Übung 5: Multivariate Statistik

- Grundwasserdatensatz Karlsruhe
  - Hauptkomponentenanalyse
  - Matrizenrechnung
  - Visualisierung

O_2
Temperatur
Eisen
Mangan
pH
Nitrat
Phosphat
Tiefe
usw.



Koch et al. (2020)

- Aufgaben in Jupyter Notebook: 05_Multivariate_Statistik_uebung

Literatur

- Trauth (2015) MATLAB Recipes for Earth Sciences (4th Ed.), Springer
- Koch et al. (2020) Groundwater fauna in an urban area: natural or affected?, Hydrology and Earth System Sciences Discussions
- Lever et al. (2017) Principal component analysis, Nature Methods 14(7), 641-642

Nützliche Weblinks:

- https://towardsdatascience.com/a-complete-guide-to-principal-component-analysis-pca-in-machine-learning-664f34fc3e5a

Karlsruher Institut für Technologie