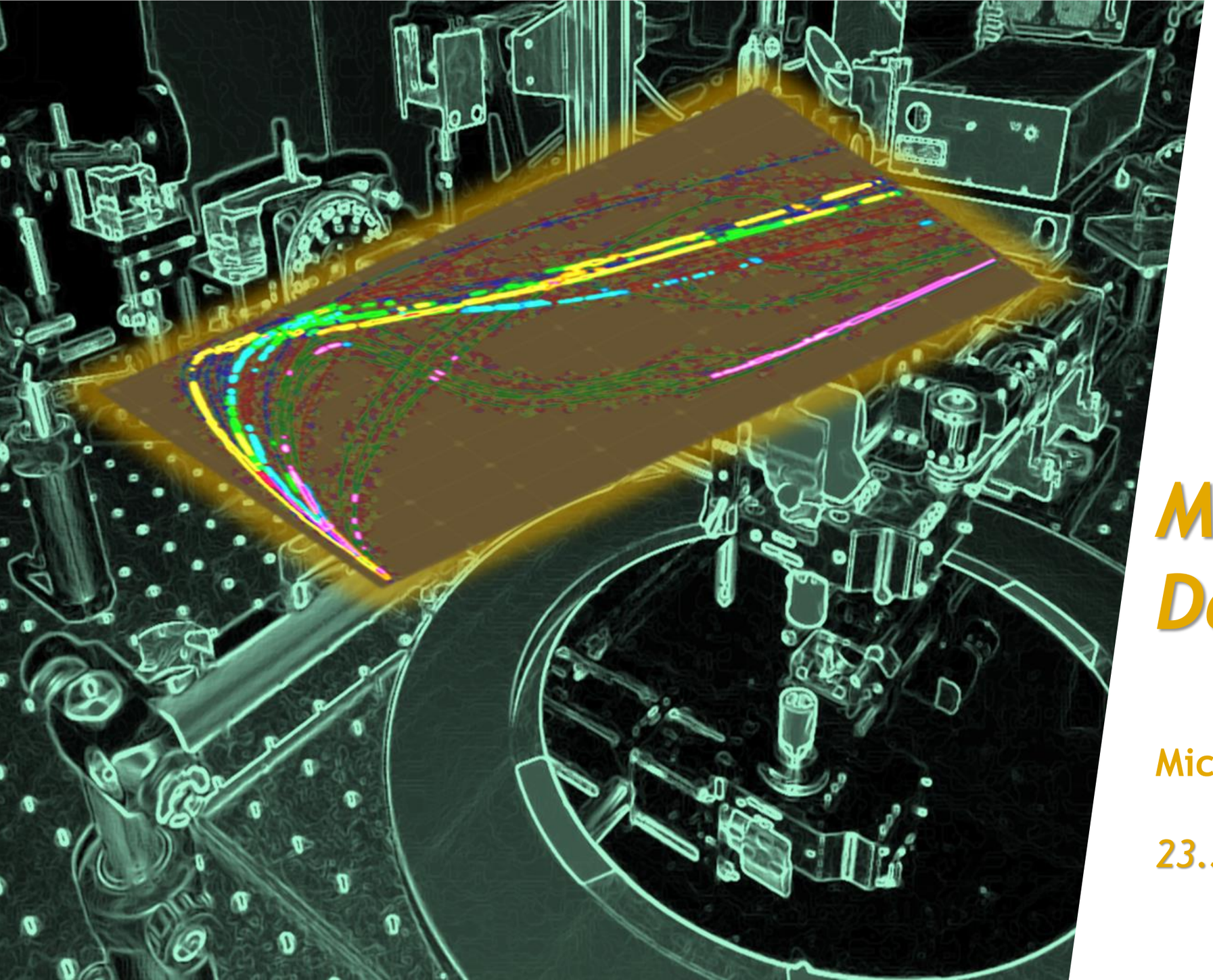


JOANNEUM RESEARCH POLICIES

Multivariate Datenanalyse

Michaela Dvorzak, Ulrike Kleb

23.3.2023

Ablauf und Inhalt

- Teil 1 - Vormittag
  - Kurzvorstellung JR-POL-DAT und unsere Arbeitsweise
  - Was ist multivariate Statistik?
  - Visualisierung
  - Exploration und Dimensionsreduktion
  - Clustering
  - Modellierung
- Teil 2 - Nachmittag
  - Use Cases – Fragestellungen und Diskussion

JOANNEUM RESEARCH POLICIES
Forschungsgruppe Data Analytics and Statistical Modelling

Systems Analysis & Data Acquisition	• Design of Experiments (DoE)
Processes	• Optimization of products and processes in design phase • Data analytics, modelling and optimization of industrial processes for Industry 4.0 ➔ “Digital Twin”
Metrology	• Calibration models for metrology and sensor systems • Development of Soft Sensors / Virtual Metrology
Reliability & Maintenance	• Reliability analysis and Lifetime modelling for components and systems • Predictive Maintenance
Predictive Analytics	• Customer-specific predictive models

Unser Vorgehensmodell … der „Data Science Process“

- Was ist die spezifische Fragestellung?
- Analysieren der Anforderungen
- Analysieren des Prozesses bzw. des Systems

- Daten aus bestehenden Quellen gewinnen
- Neue Datenquellen erschließen
- Daten fusionieren und aufbereiten

- Plausibilität prüfen
- Analysieren
- Visualisieren

- Detektion
- Klassifikation
- Regression
- Prognose
- Optimierung

- Validierung

- Neue Erkenntnisse vermitteln

- Präsentation
- Report
- Demo
- Software-Tool

Statistische Methoden

Datengewinnung

- Statistische Versuchsplanung / Design of Experiments (DoE)
  - Screening Designs
  - Full Factorial Designs
  - Response Surface Designs (CCD, …)
  - Taguchi Designs
- Stichprobenplan
- Fragebogen
- Filtern / Glätten
  - Signalzerlegung
  - FFT, Wavelets, …
  - MA, EWMA, LOESS, …

Analyse

- Kennzahlen, Tabellen
- SPC (Stat. Process Control)
- Plots
  - Histogramme, Boxplots
  - Scatterplot (2D/3D)
- Korrelation
- Zeitreihenanalyse
  - Autokorrelation
  - Periodizität
- Multivariate Verfahren
  - Clustering
  - Dimensionsreduktion (PCA, UMAP, …)
- Etc.!

Modellierung

- Regression / Klassifikation („supervised“)
  - Linear Models, RSM
  - GLM, GAM, GAMLSS
  - PLS
- Reliability / Survival
  - Weibull, Proportional Hazard (Cox
- Zeitreihenmodelle
  - ARIMA
- Functional Data Regression
  - e.g. Spektralkurven
- Tensor Regression
- Uncertainty / Bayesian Models
- Etc.

Statistische Methoden

Datengewinnung

- Statistische Versuchsplanung / Design of Experiments (DoE)
  - Screening Designs
  - Full Factorial Designs
  - Response Surface Designs (CCD, …)
  - Taguchi Designs
- Stichprobenplan
- Fragebogen
- Filtern / Glätten
  - Signalzerlegung
  - FFT, Wavelets, …
  - MA, EWMA, LOESS, …

Analyse

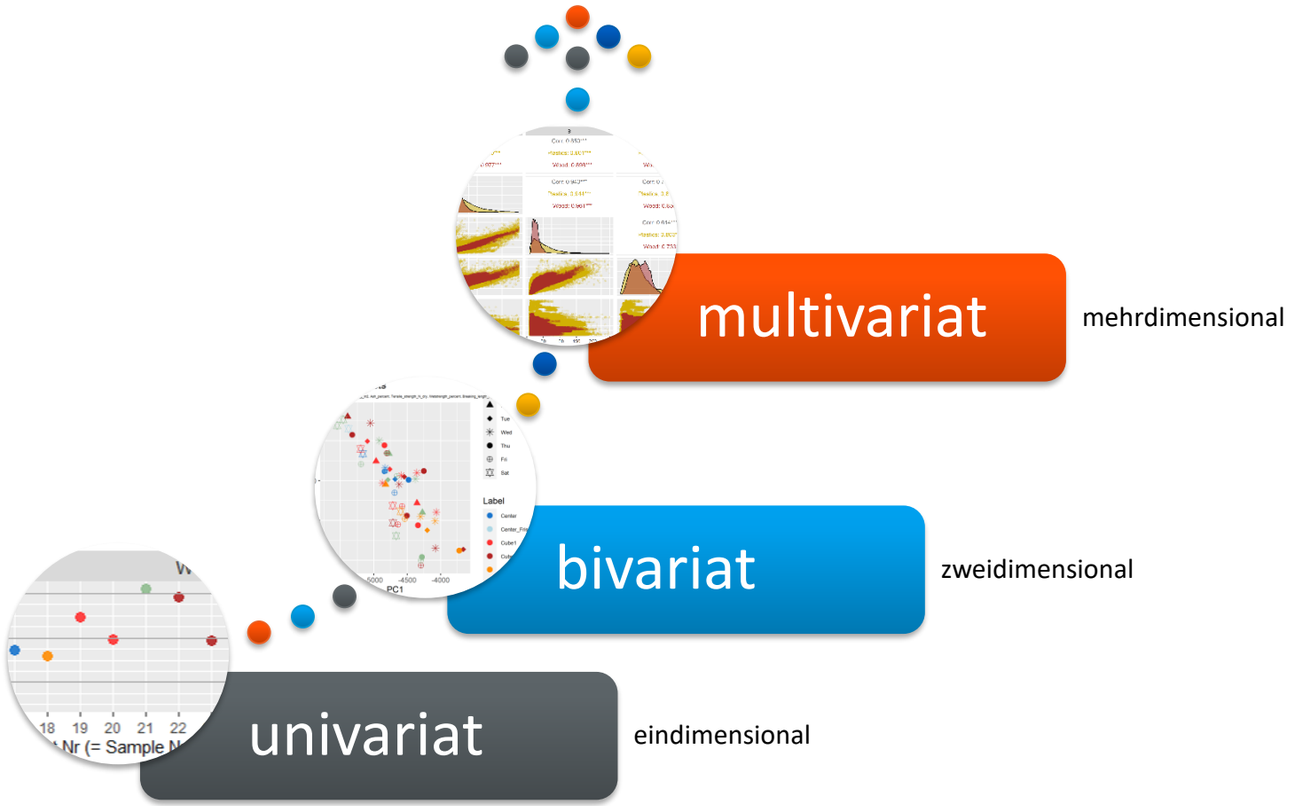
- Kennzahlen, Tabellen
- SPC (Stat. Process Control)
- Plots
  - Histogramme, Boxplots
  - Scatterplot (2D/3D)
- Korrelation
- Zeitreihenanalyse
  - Autokorrelation
  - Periodizität
- Multivariate Verfahren
  - Clustering
  - Dimensionsreduktion (PCA, UMAP, …)
- Etc.!

Modellierung

- Regres[illegible]n / Kl[illegible]ssifikation („supe[illegible]ed“
  [illegible]SM
  [illegible]LSS
  [illegible]rtional Hazard
  [illegible]delle
  [illegible]l Data Regression
  [illegible]r Regression
- Un[illegible]ertainty / Bayesian Models
- Etc.

XAI

Was sind multivariate Daten?



- Multivariate Daten
  - Für jede Beobachtung / Objekt / Person gibt es (Mess-)Werte von mehreren Merkmalen / Variablen
- … zum Beispiel
  - Fragebogen mit mehreren Fragen ➔ mehrere Antworten pro Teilnehmer*in
  - Bestimmung mehrerer Laborparameter pro Blutprobe
  - Messung verschiedener Qualitätsmerkmale pro Werkstück
  - Etc.

Multivariate Datenanalyse ist …

Simultane Untersuchung mehrerer Zufallsvariablen / Merkmale

Gleichzeitige Betrachtung von zwei oder mehr abhängigen Variablen

Analyse von Abhängigkeiten zwischen mehr als 2 Variablen

Analysieren von komplexen Zusammenhängen

Aufdecken von Strukturen in (großen) Datenmengen

… vielseitig und flexibel, aber anspruchsvoll / komplex

"Multivariate analysis refers to statistical methods used to analyze data sets that involve multiple variables. It is an extension of univariate (one variable) and bivariate (two variables) analysis, where multiple variables are simultaneously examined to understand their relationships and patterns.

Multivariate analysis methods can be categorized into various types, such as descriptive and inferential multivariate analysis. In descriptive multivariate analysis, techniques like factor analysis, cluster analysis, and principal component analysis may be used to structure the data and identify patterns.

In contrast, inferential multivariate analysis involves statistical tests and models that examine the relationships between variables to form and test hypotheses. Examples of inferential multivariate analysis techniques include **multiple regression,** canonical correlation analysis, and discriminant analysis.

Multivariate analysis is used in many different fields, such as finance, marketing research, biology, and psychology, to name just a few."

Quelle: ChatGPT

Multivariate Datenanalyse – Überblick und Einteilung der Methoden

Visualisierung / Plots

- Histogramm und Verteilung
- Boxplots
- Scatterplot und Scatterplotmatrix, 2D und 3D
- Heatmaps
- Parallele Koordinaten
- Spider Charts
- Andrew‘s-Curves
- Glyphen ➔ Chernoff-Faces

Exploration und Dimensionsreduktion

- Korrelationsanalysen und Korrelationsplots
- Mahalanobisdistanz
- Transformation & Dimensionsreduktion (PCA, UMAP) und zugehörige Visualisierung

Clustering

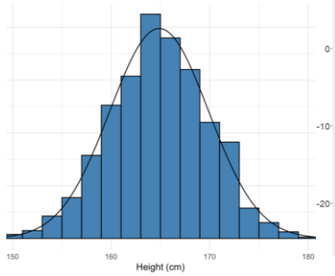
- Hierarchische Clusteranalyse
- Partitionierende Clusteranalyse
- Dichtebasierte Clusteranalyse

Modellierung

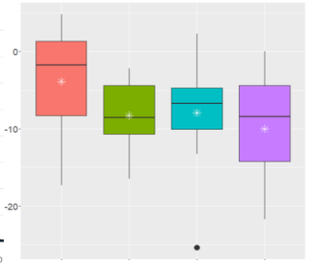
- Multiple Regression
- Diskriminanzanalyse
- PLS Regression und Classification

Methoden der Visualisierung

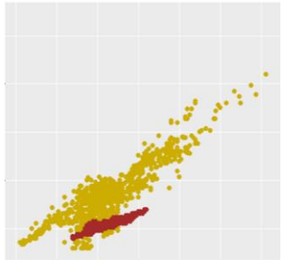
Histogramm & Verteilung



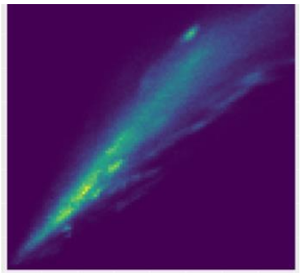
Boxplots



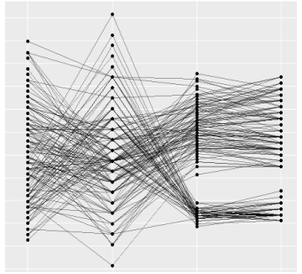
Scatterplot



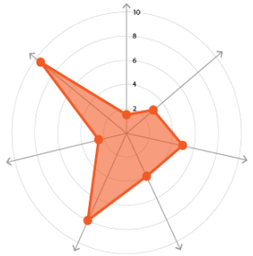
Heatmaps



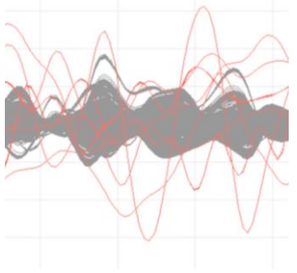
Parallele Koordinaten



Spider Charts



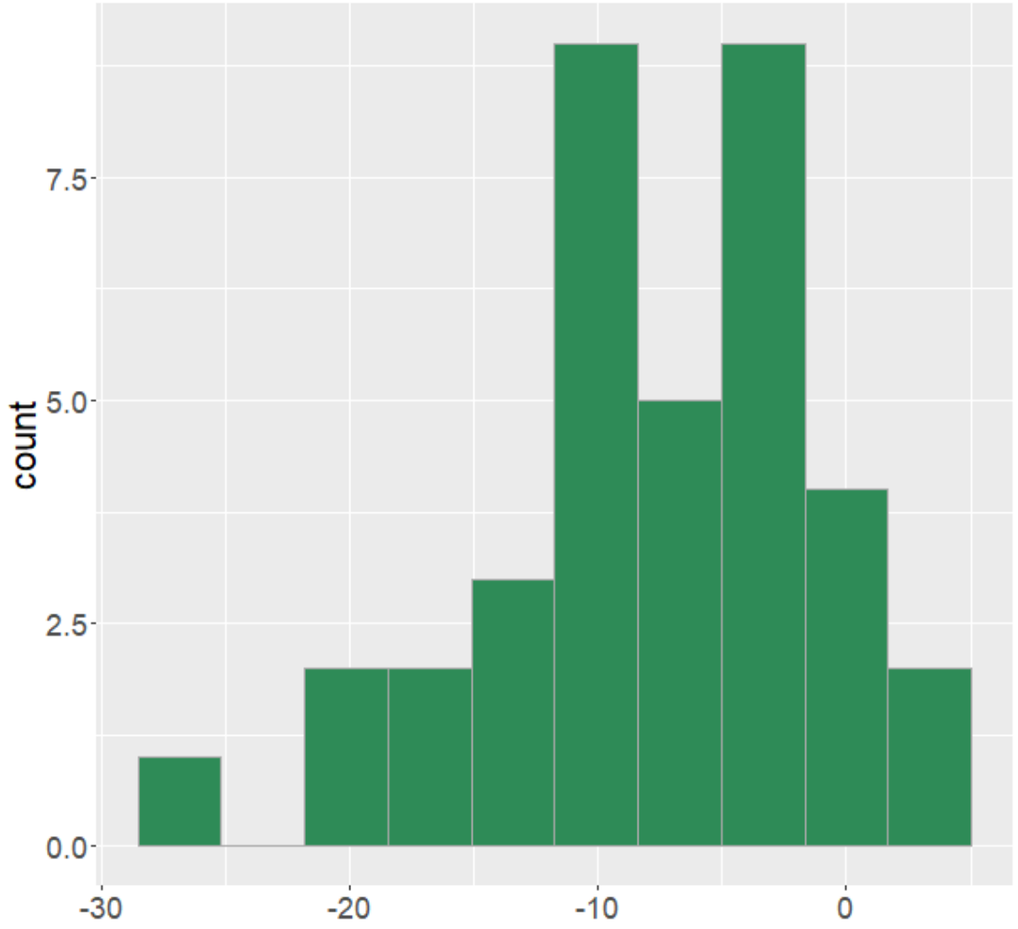
Andrew‘s Curves

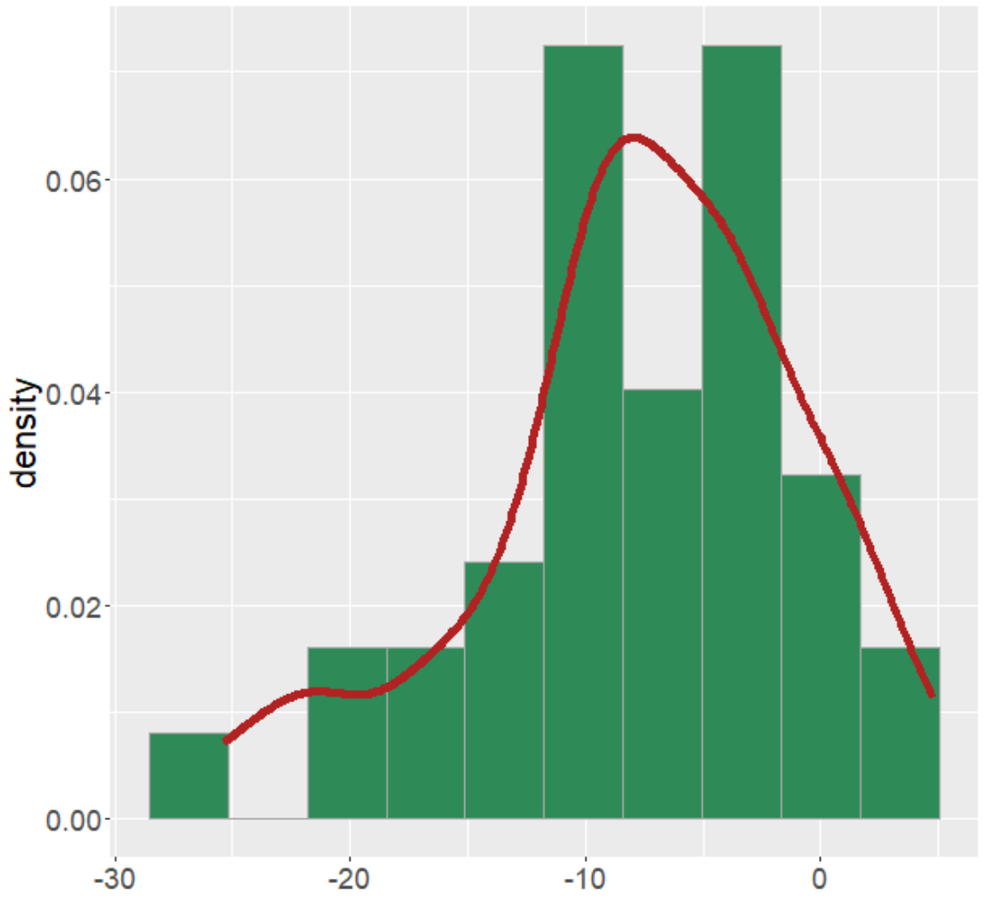


Chernoff Faces

JOANNEUM RESEARCH POLICIES

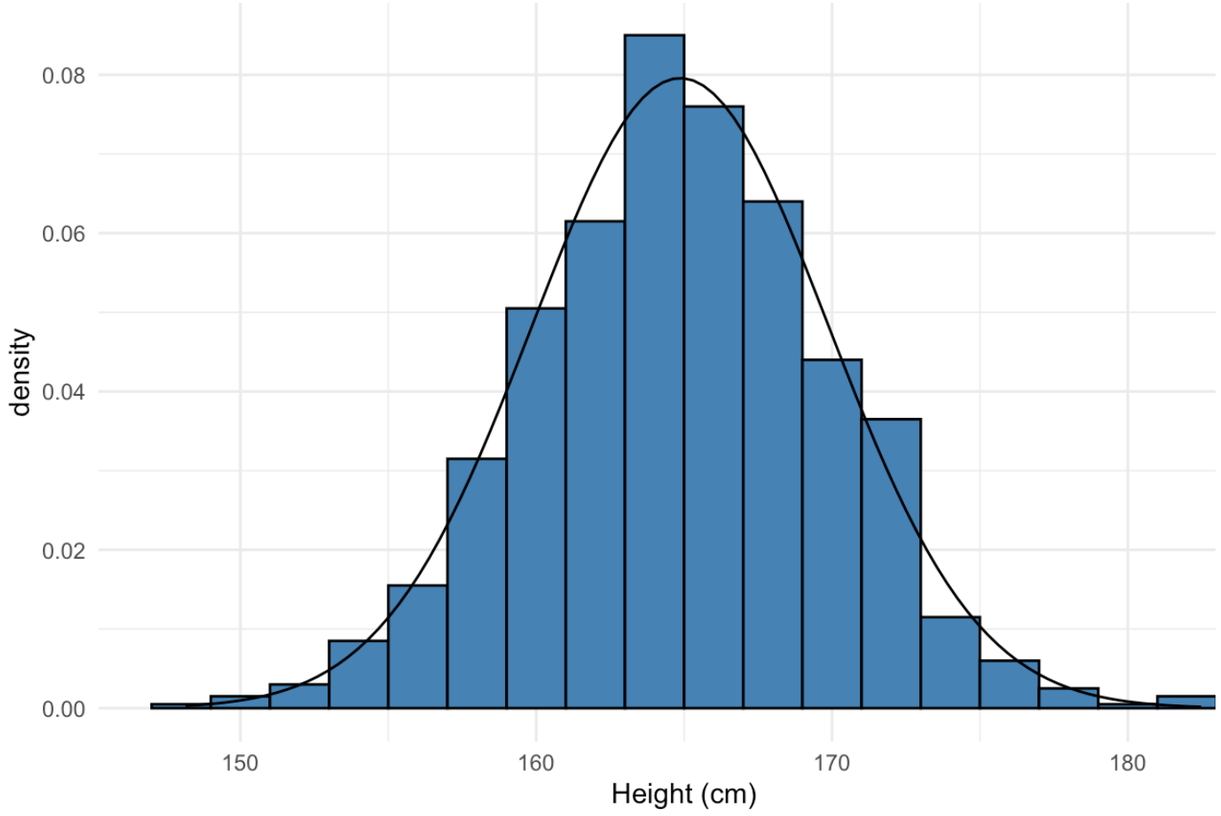
Visualisierung von Daten ➔ Histogramm und Dichteschätzung



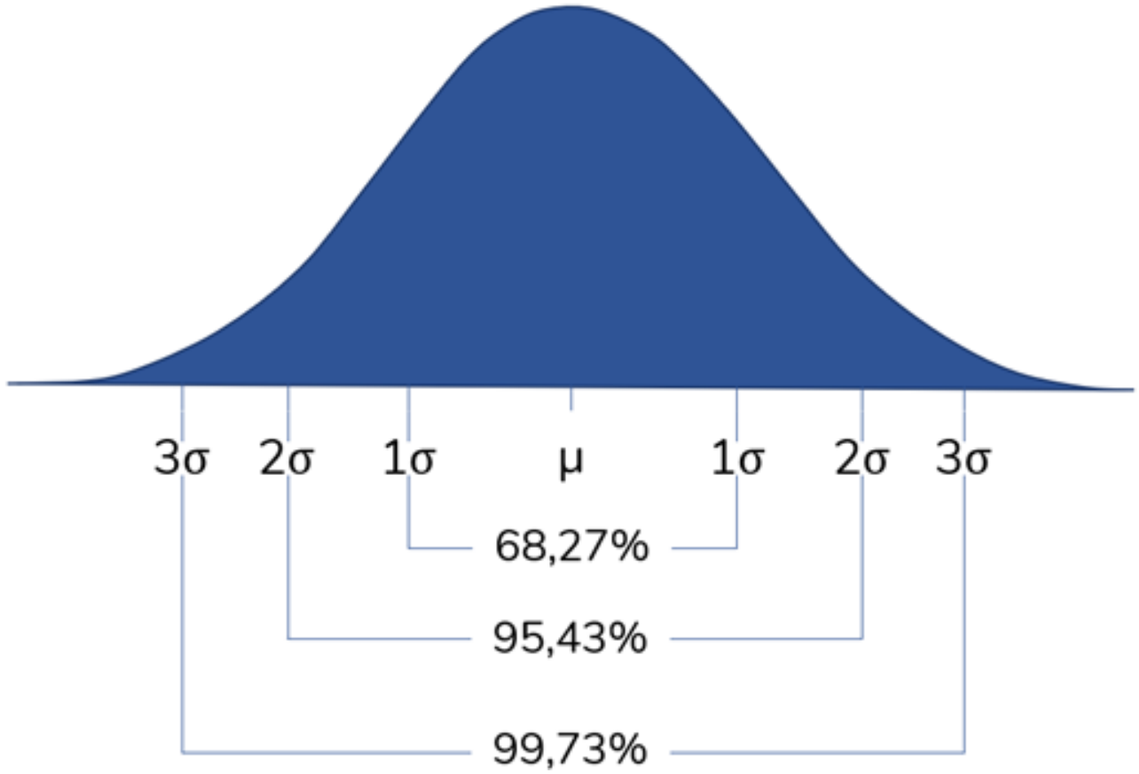


Grundlage: Klassierung / Diskretisierung stetiger Merkmale (Informationsverdichtung) -> Häufigkeiten je Klasse
Histogramm: Darstellung der Häufigkeiten

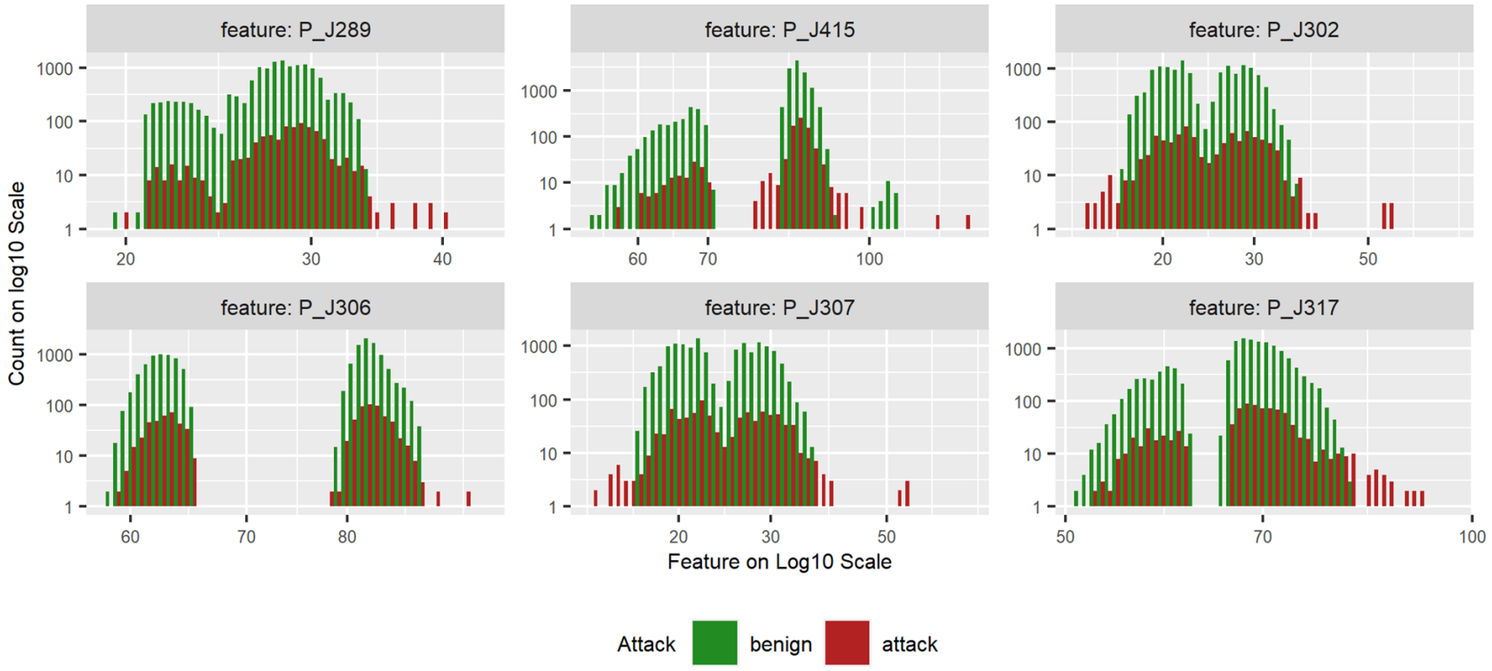
Die Normalverteilung / Gauss-Verteilung!

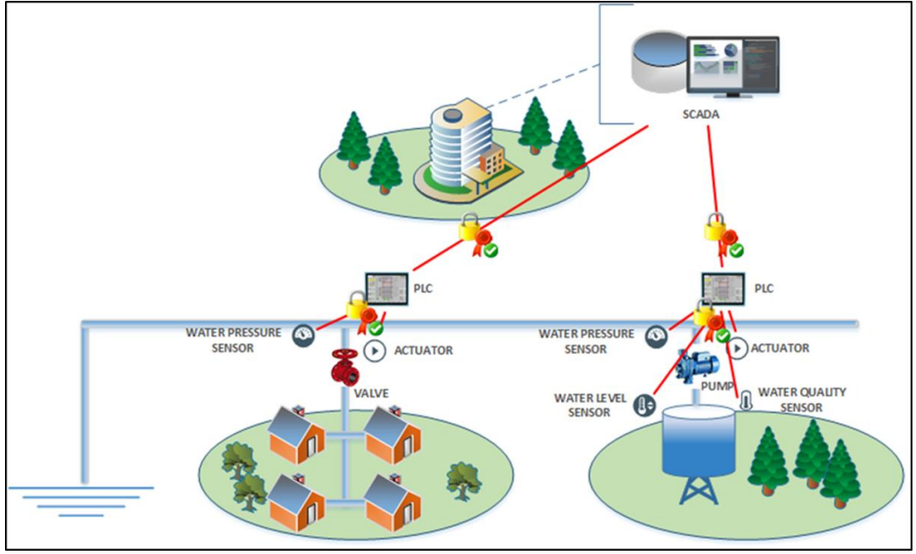


… „Sigma-Bereiche“



Histogramme mehrerer Variablen

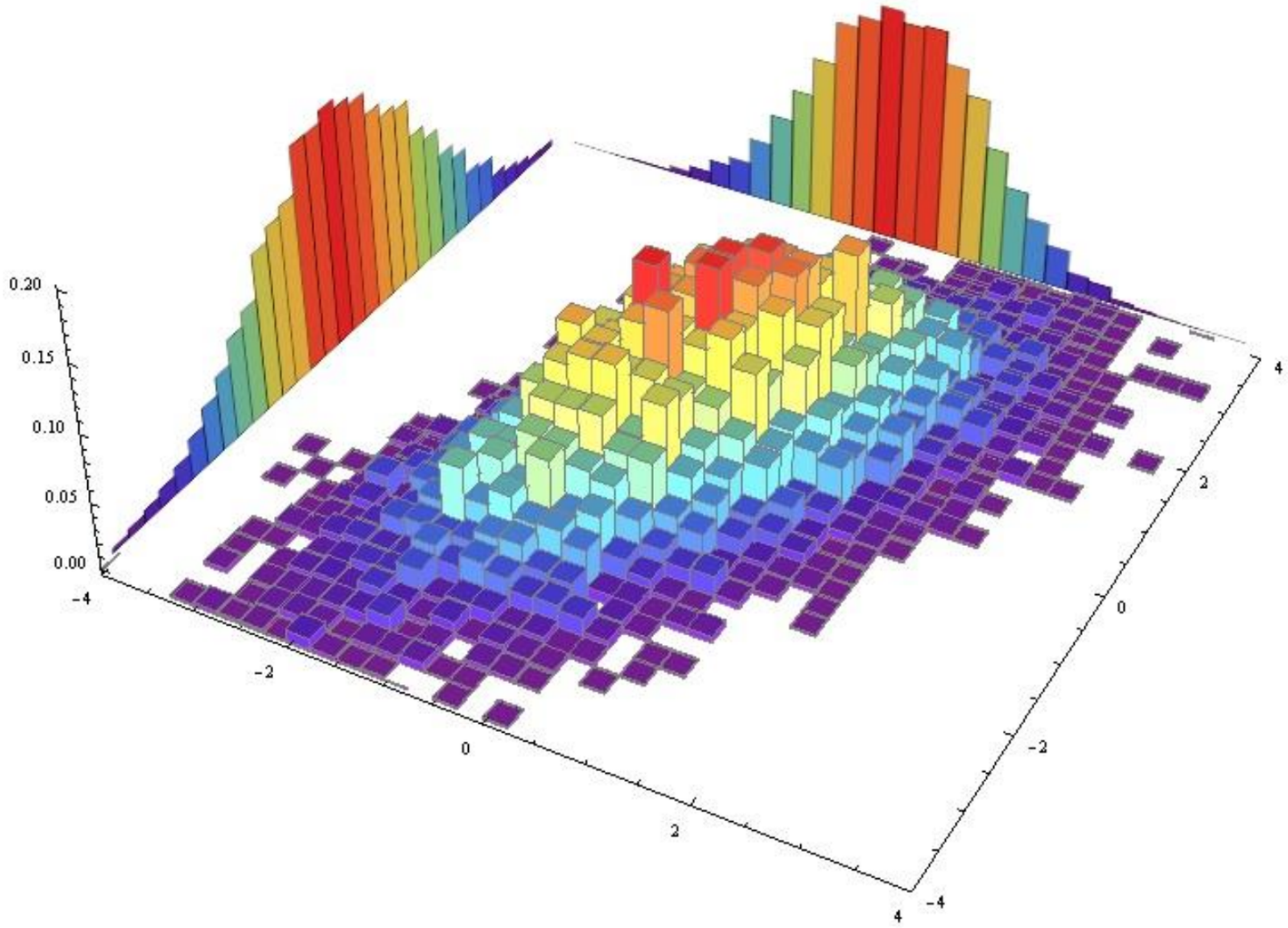




Beispiel

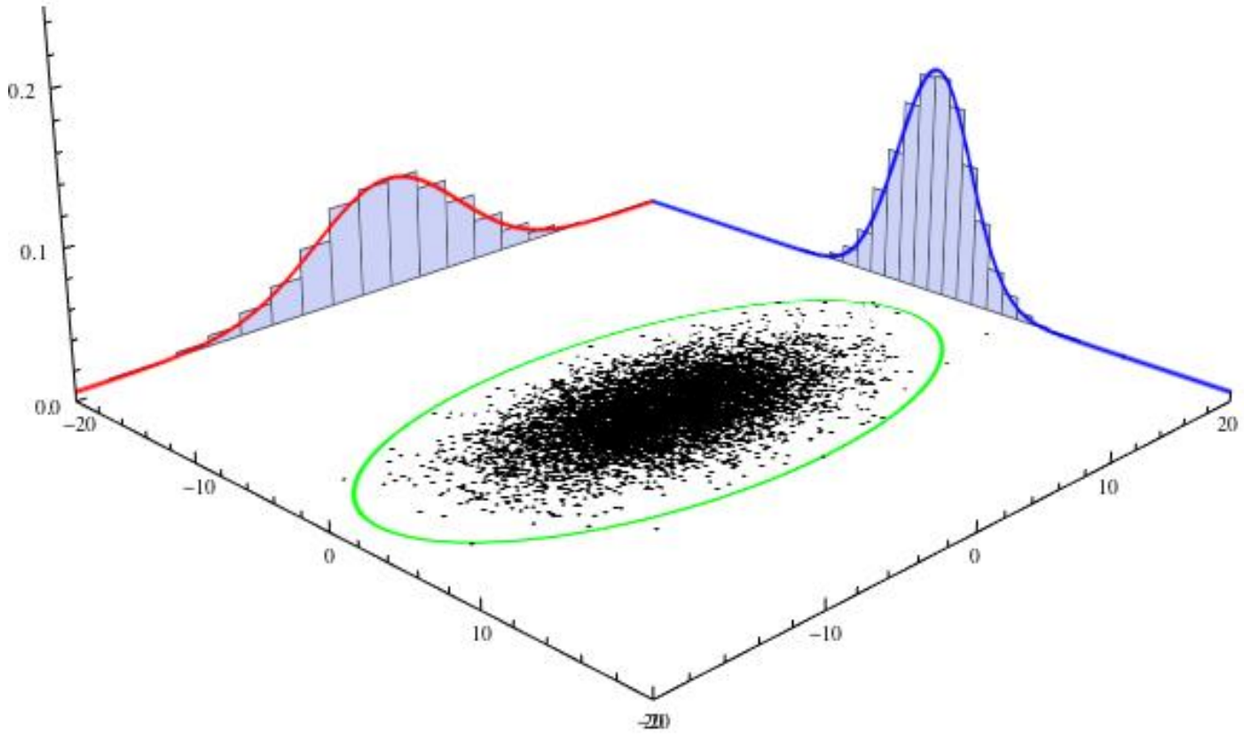
- Water distribution system, consisting of
  - pipes, junctions, storage tanks, pumps, valves, …
- SCADA readings
  - Water levels, flow, pressure, status (on/off), …
- Ddata
  - 1 observation/data row per hour
  - 43 Features
  - 14 (labelled) attacks

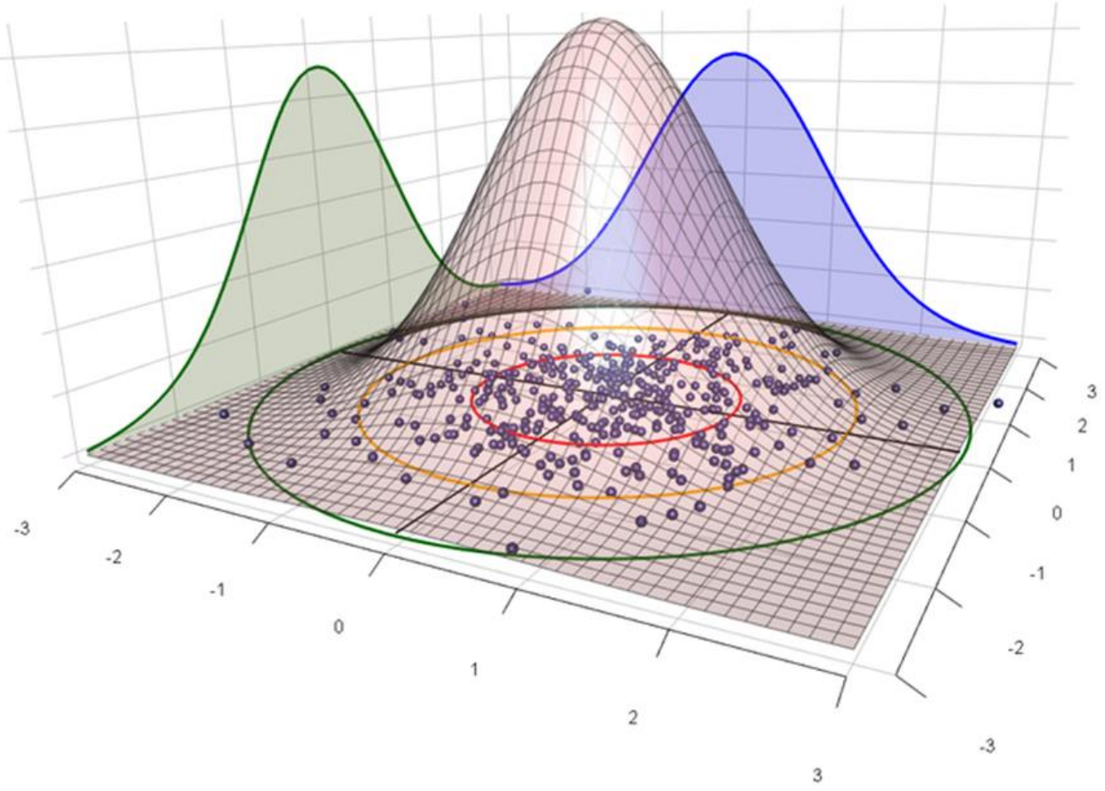
Bivariates Histogramm



JOANNEUM RESEARCH POLICIES

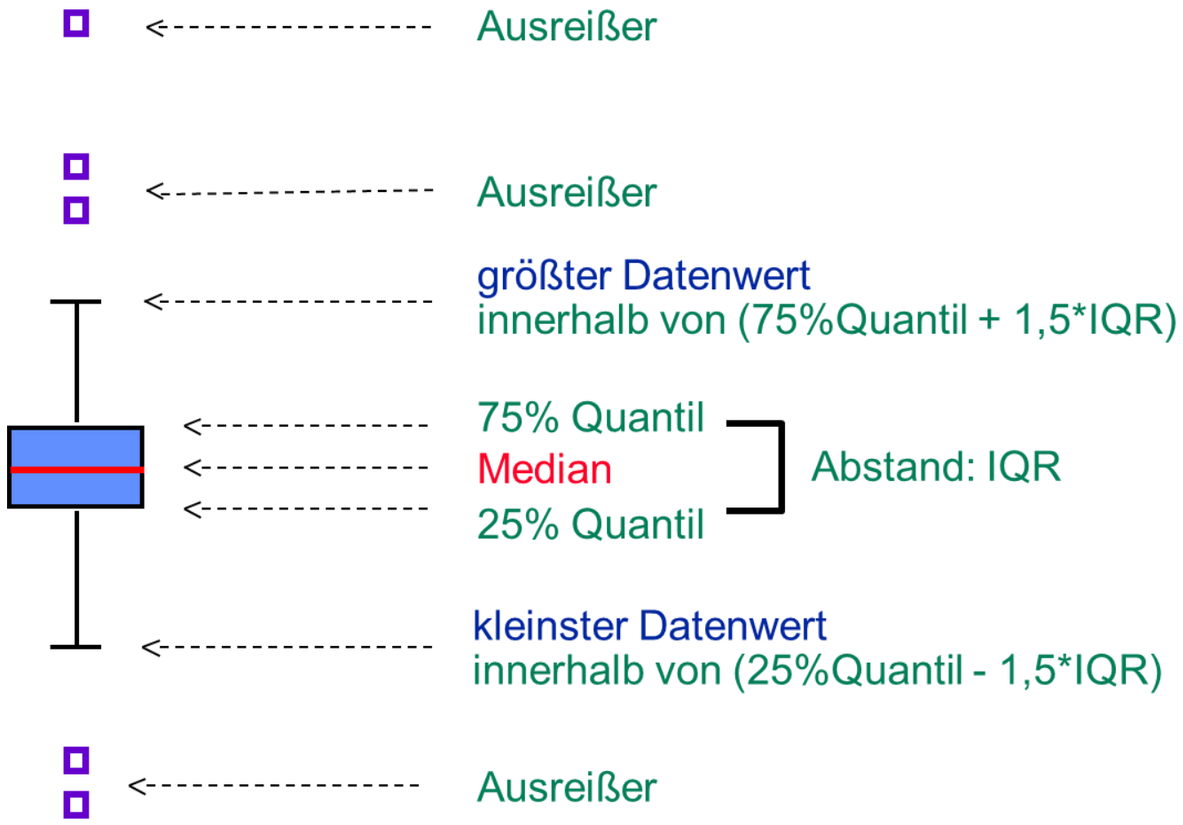
Bivariate Normalverteilung

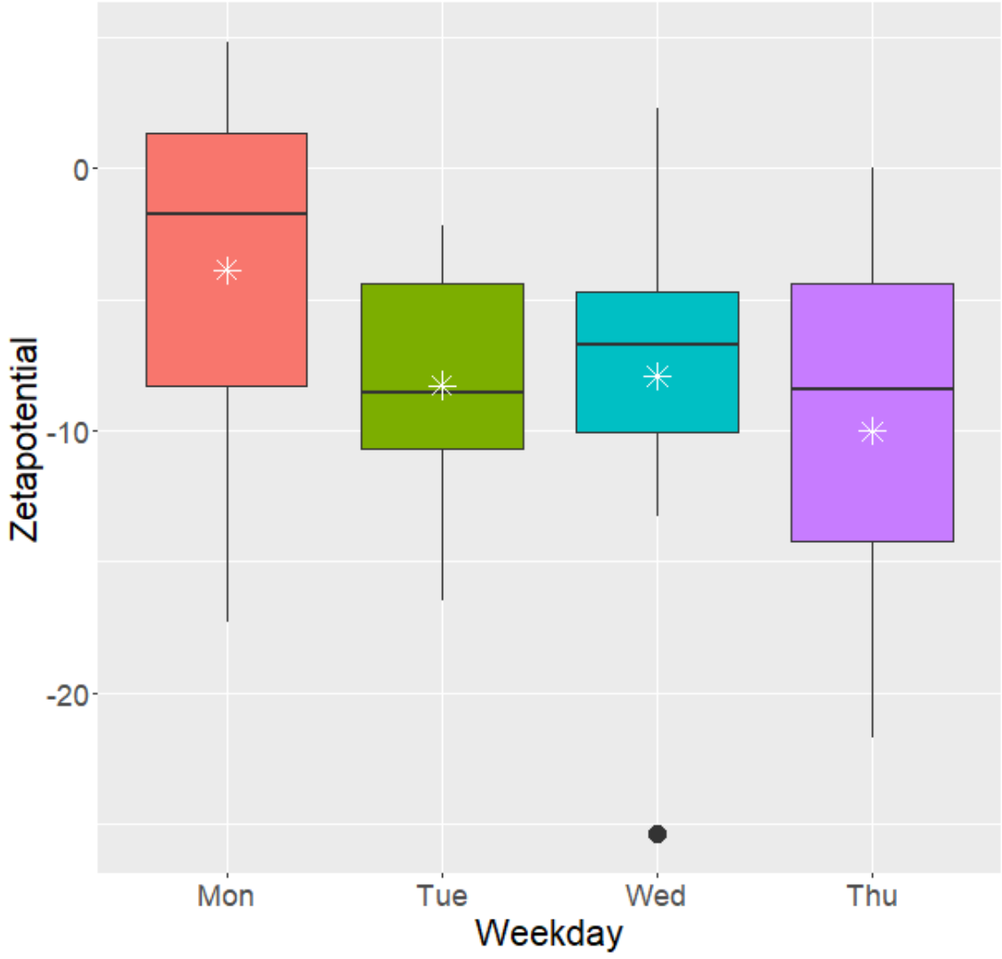




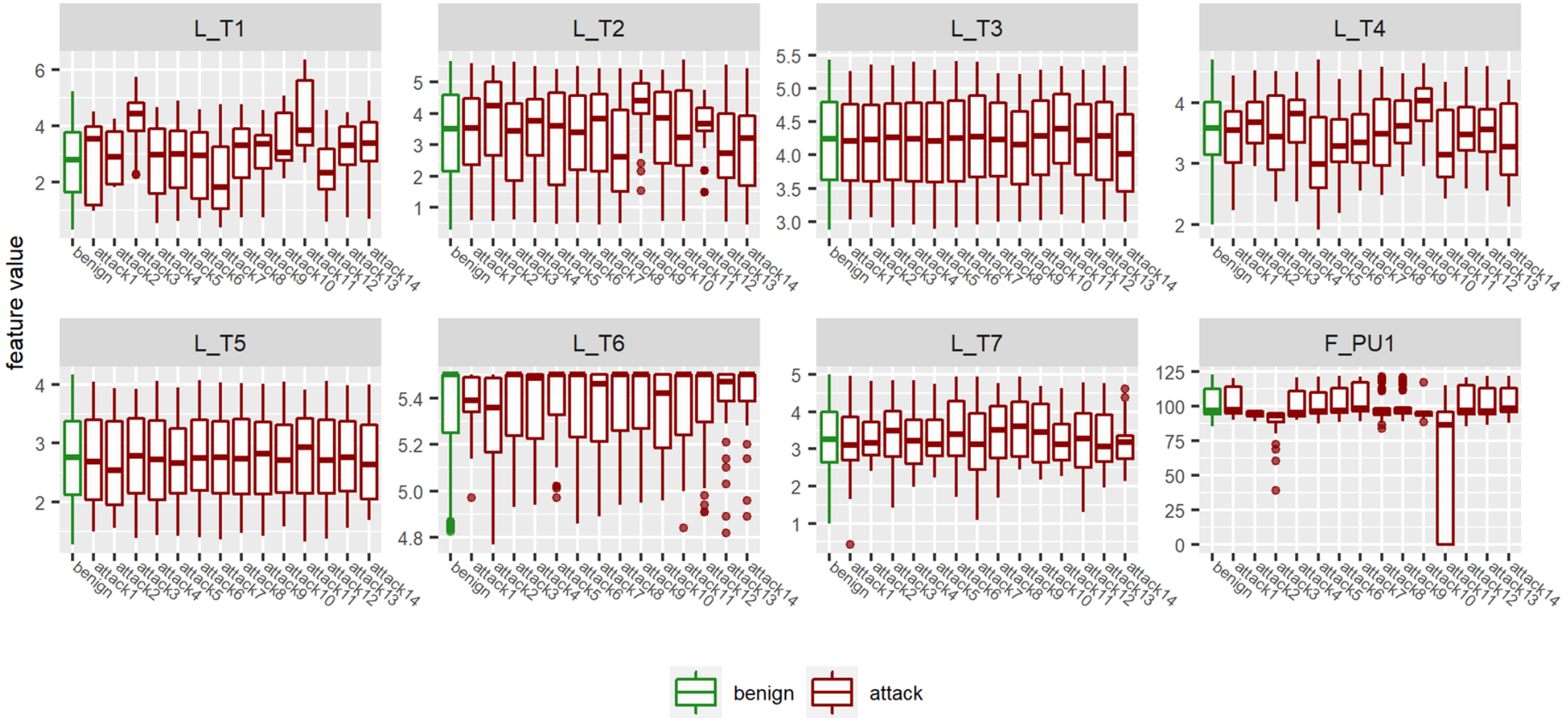
JOANNEUM RESEARCH POLICIES

Boxplots zur Visualisierung von Verteilungen und zum Gruppenvergleich





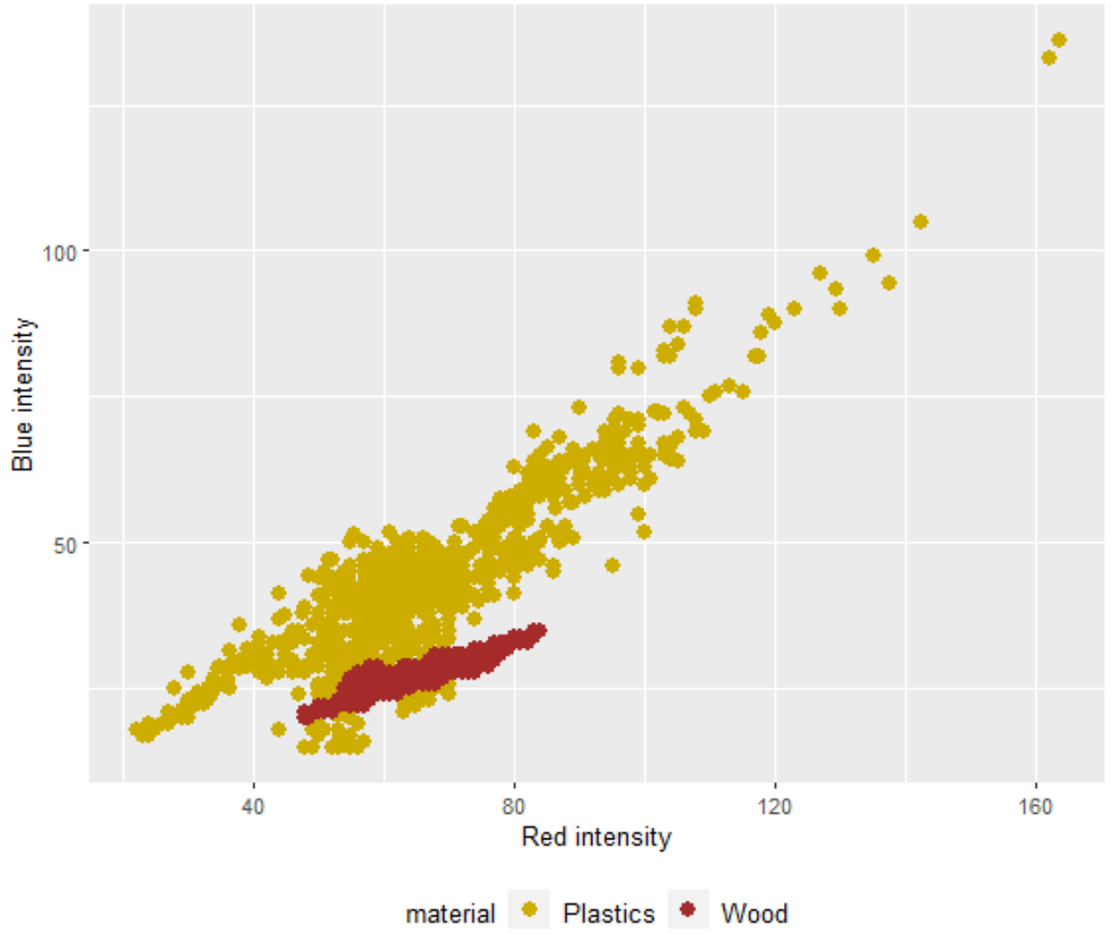
Gruppierte Boxplots für mehrere Merkmale

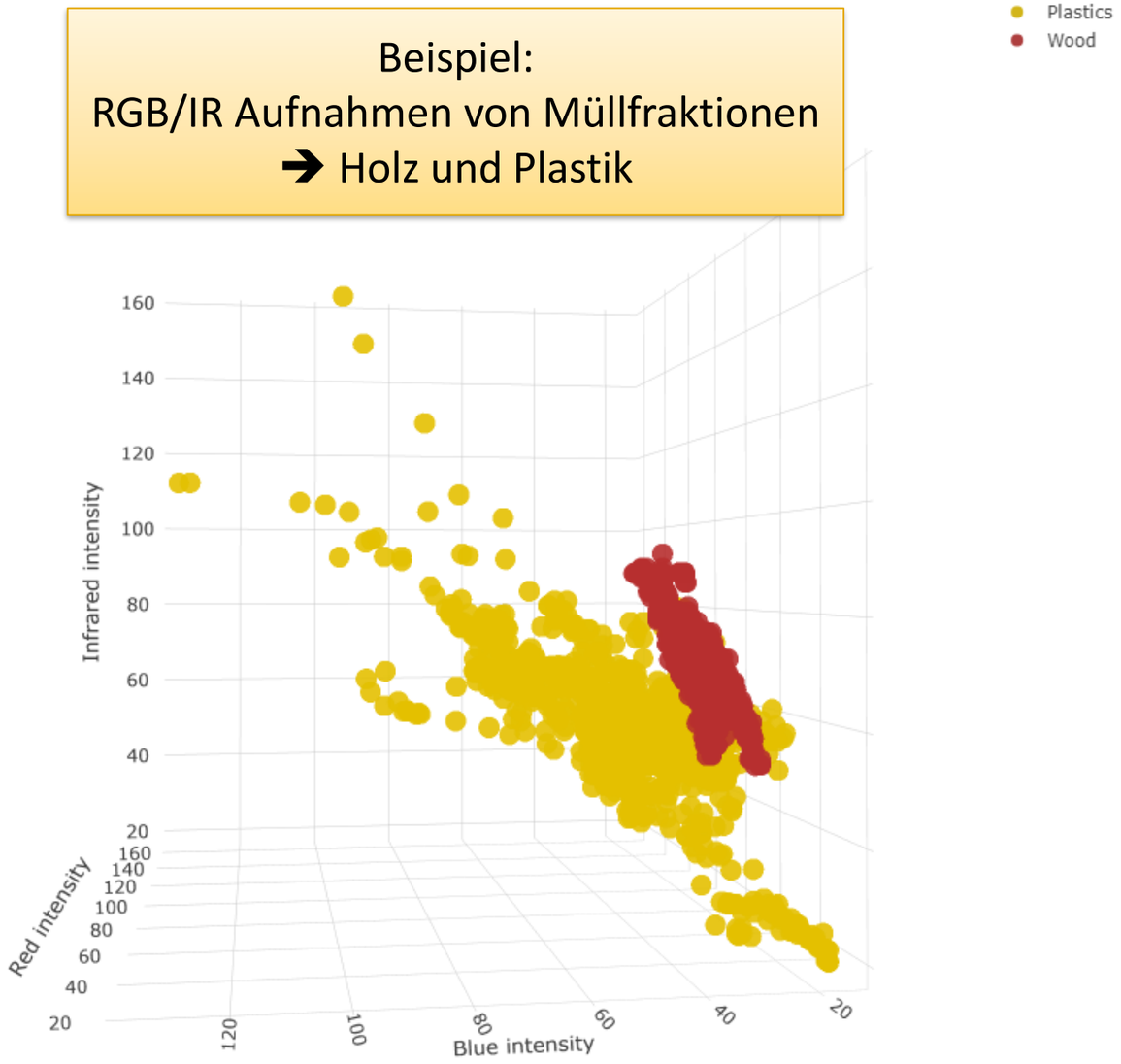


JOANNEUM RESEARCH POLICIES

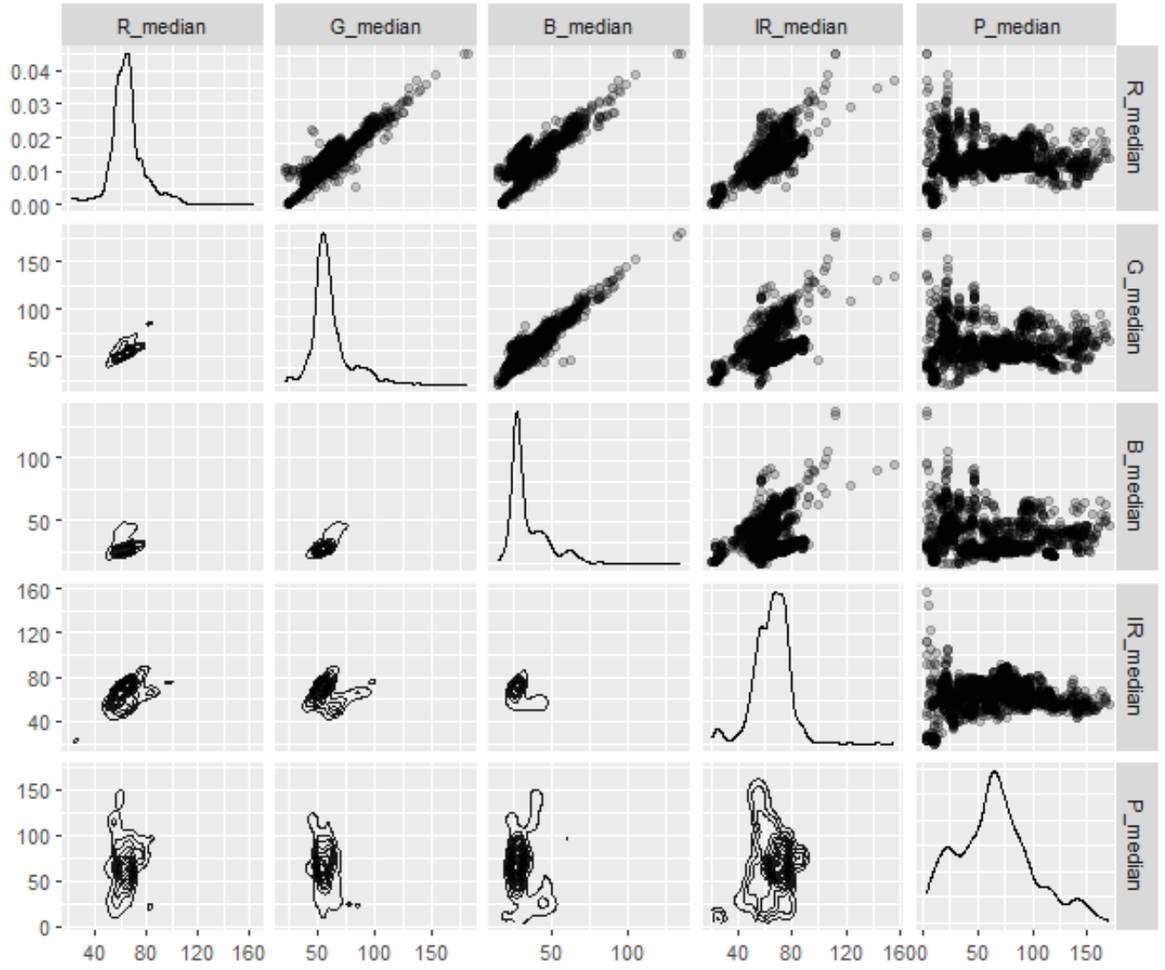
Scatterplot 2D und 3D

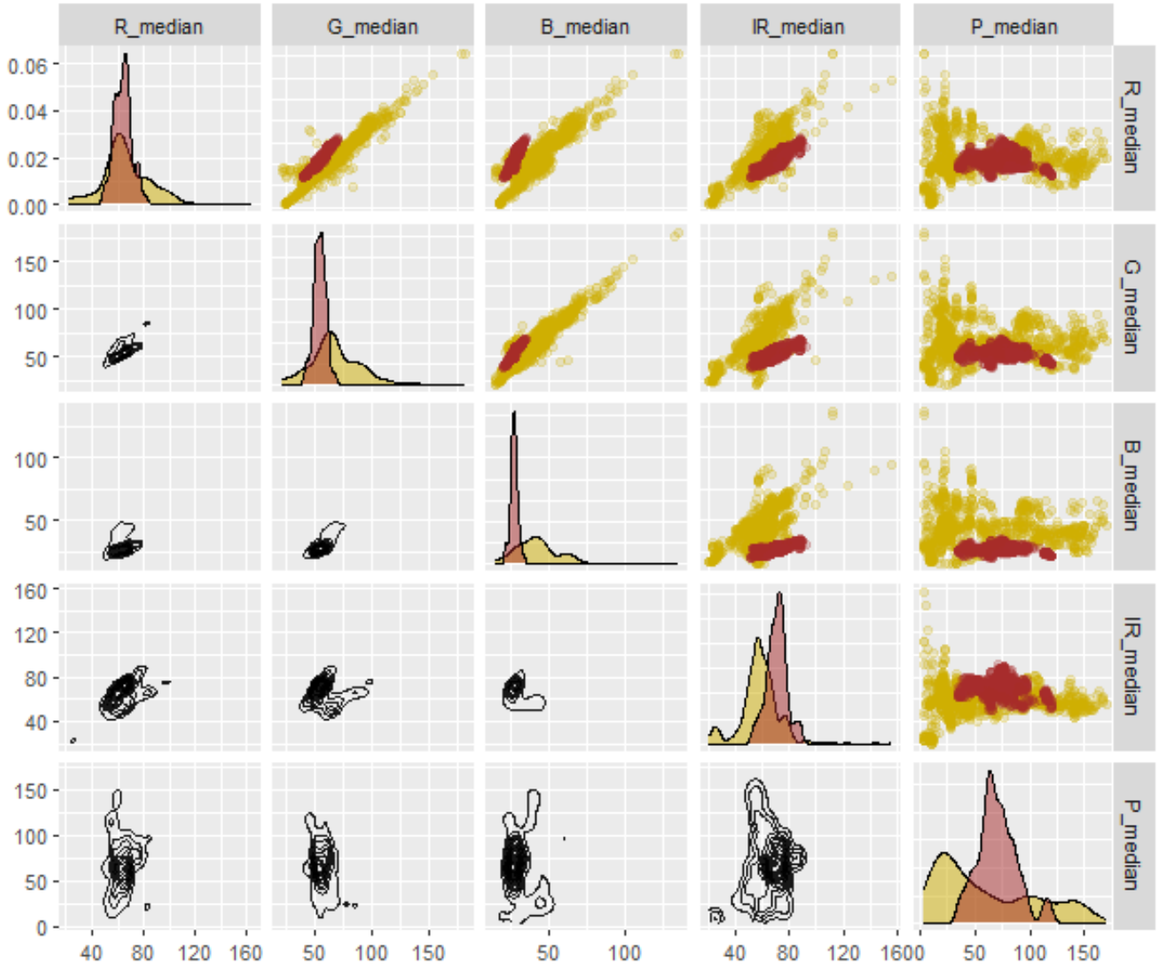
Beispiel:
RGB/IR Aufnahmen von Müllfraktionen
➔ Holz und Plastik





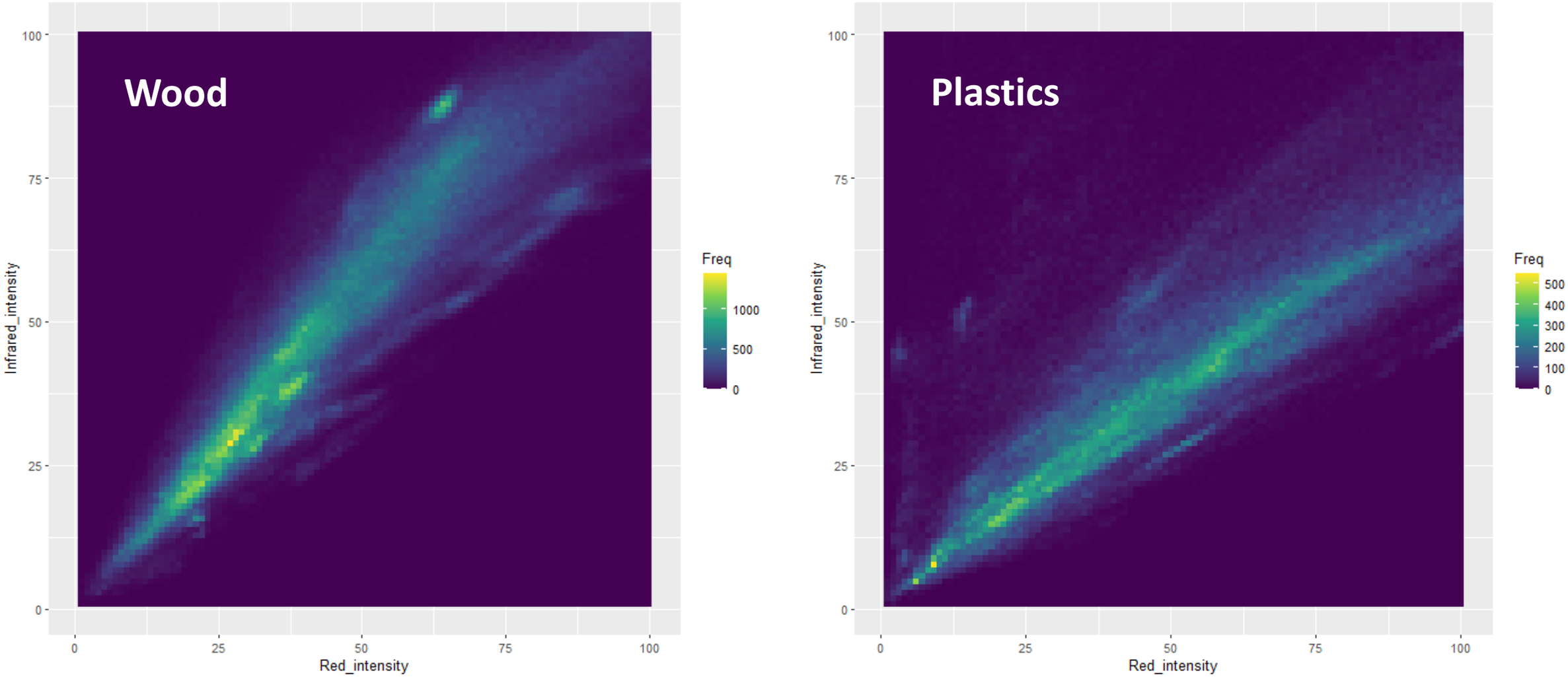
Visualisierung von Zusammenhängen ➔ Scatterplot-Matrix





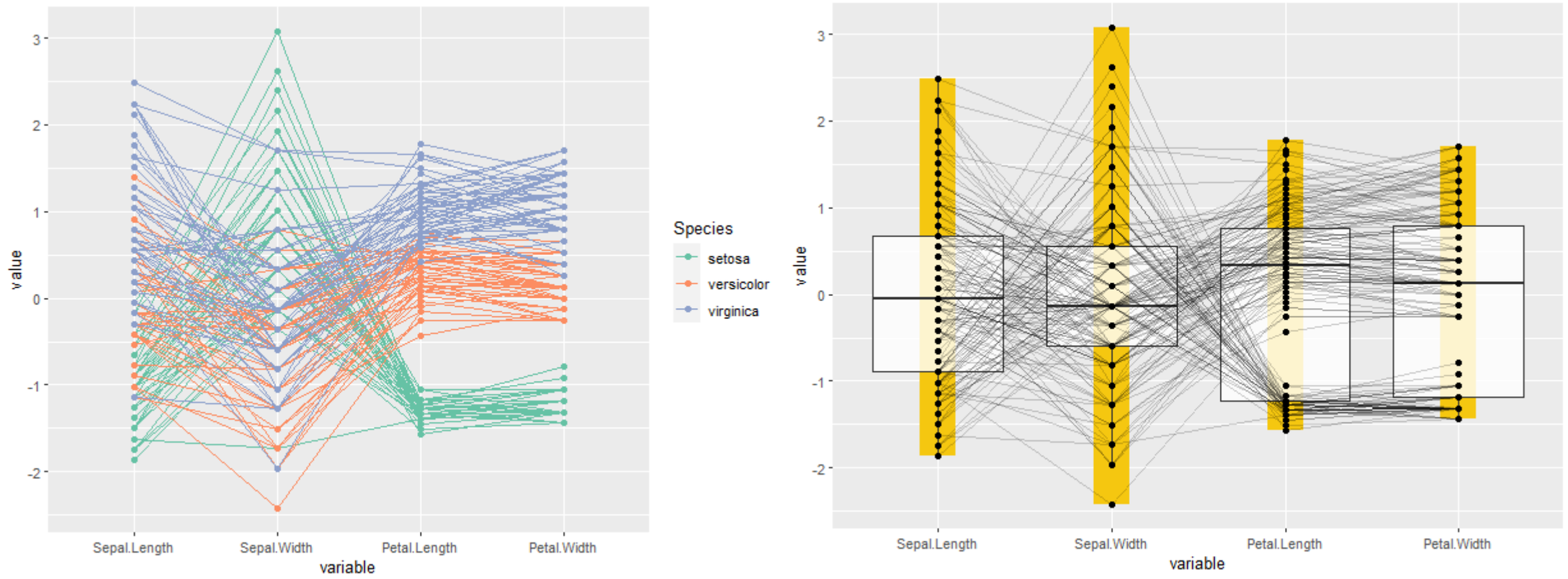
JOANNEUM RESEARCH POLICIES

Visualisierung von Zusammenhängen ➔ Heatmaps



JOANNEUM RESEARCH POLICIES

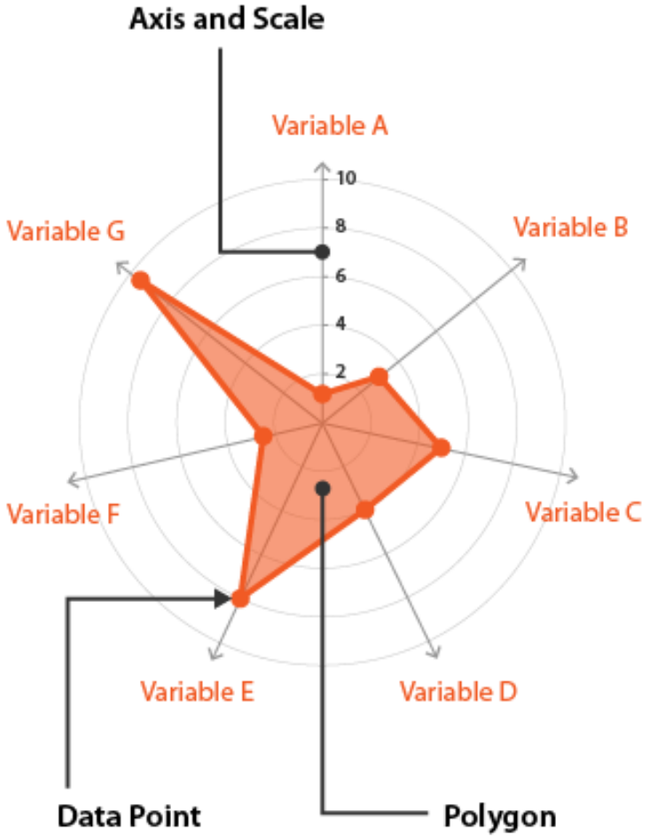
Parallele Koordinaten



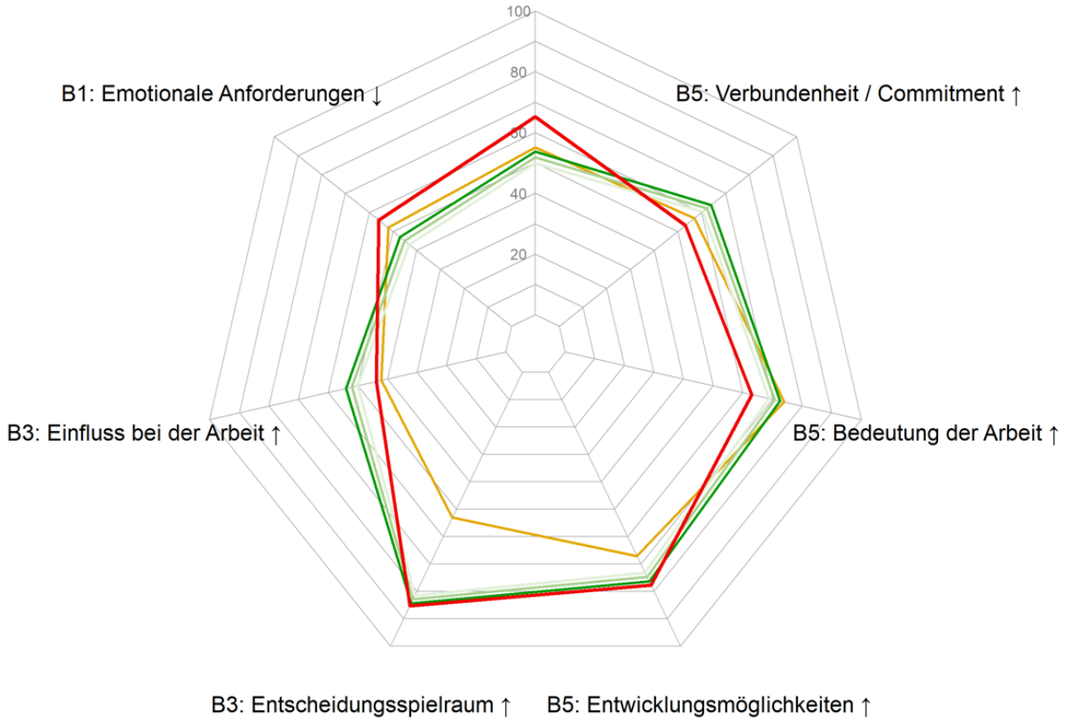
- Koordinatensystem: Parallele Achsen mit äquidistantem Abstand
- jede Achse wird mit einer Variable assoziiert (Punkte auf den Achsen mit Linien verbunden)
- Anordnung der Achsen (Variablen) (!), Skalierung der Achsen

Spider Chart (Spinnennetz-Diagramm, Radar Chart)

- Sternförmige Anordnung der Merkmalsachsen (Sternförmige Koordinaten)

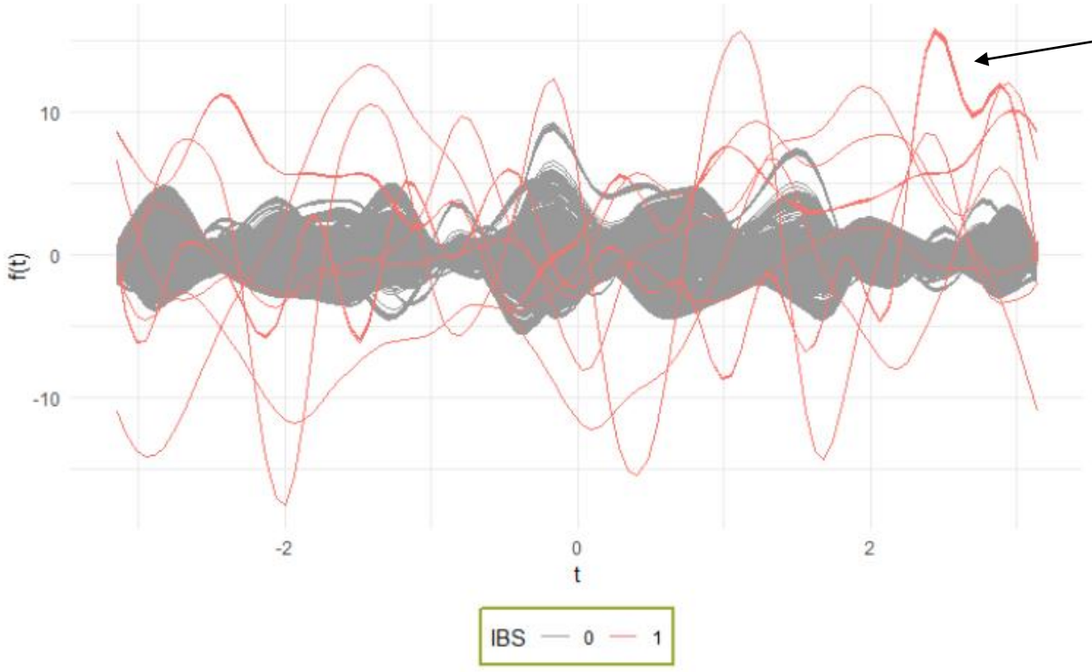


Beispiel: „Arbeitszufriedenheit“ in Unternehmen



Andrew's Curves

- Visualisierung von Strukturen in multivariaten Datensätzen (Andrews, 1972)
- Abbildung der Merkmale auf eine zweidimensionale Kurve $f_i(t)$ im Intervall $t \in [-\pi, \pi]$ (Datentransformation)



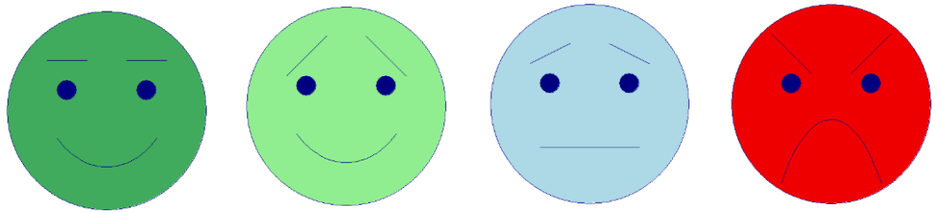
Höhere Variabilität

→ Erkennen von Strukturen / Gruppen

Beispiel: Analyse von Mikrobiomdaten

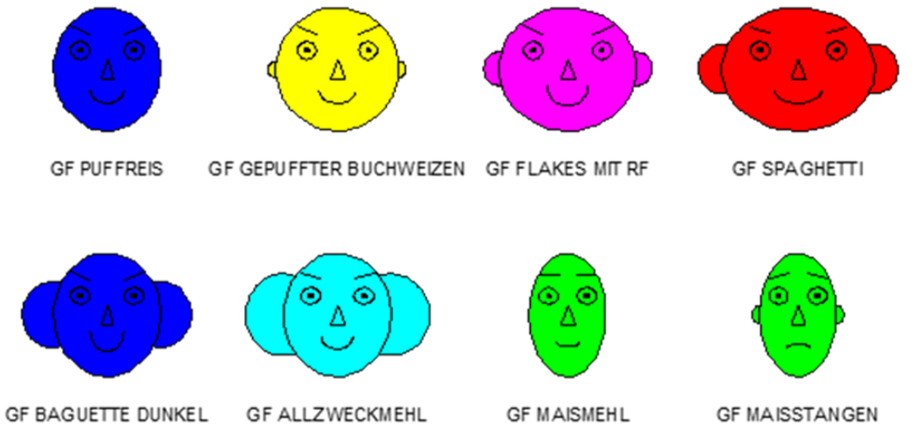
→ Reizdarmsyndrom („Irritable bowel syndrom", IBS)

Glyphen-basierte Ansätze -> Chernoff Faces



- Glyphe = graphisches Objekt, das ein multivariates Datenobjekt repräsentiert
- Zuordnung (Abbildung) der Merkmale (Dimension) zu graphischen Attributen des Objekts (Form, Farbe, Größe, Orientierung, etc.)
- Beispiel Chernoff Faces → Visualisierung von multivariaten Daten durch menschliche Gesichter
- Idee:
  - jeder Variable eines Datensatzes wird ein Gesichtsmerkmal zugeordnet
  - für jede Beobachtung entsteht individuelles Gesicht
  - bis zu 18 Merkmale darstellbar (Augen, Augenbrauen, Mund, Nase, Größe des Gesichts, Farben etc.)
- → Menschen können Komplexität multivariater Daten leichter erfassen
- Herausforderung
  - Welche Variable für welches Gesichtsmerkmal? → gut / schlecht soll unterscheidbar sein

Beispiel:
Bewertung verschiedener glutenfreier Lebensmittel



Exploration & Dimensionsreduktion

Exploration

- Ziel: Erkennen von Zusammenhängen, Strukturen und Besonderheiten in den Daten

Korrelationsanalysen

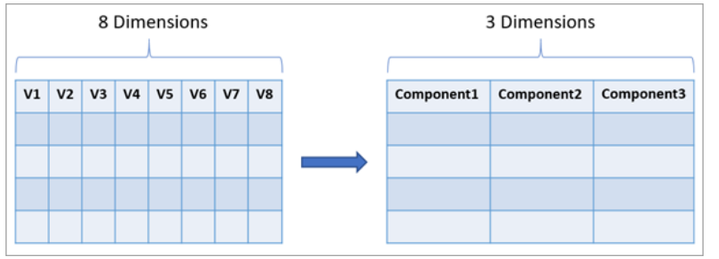
Korrelationsplots

Mahalanobisdistanz

Transformation & Dimensionsreduktion

- Ziel: Reduktion der Dimension des Datensatzes unter Bewahrung relevanter Information in den Daten

Transformation aus hochdimensionalen Raum -> niedrigdimensionalen Raum



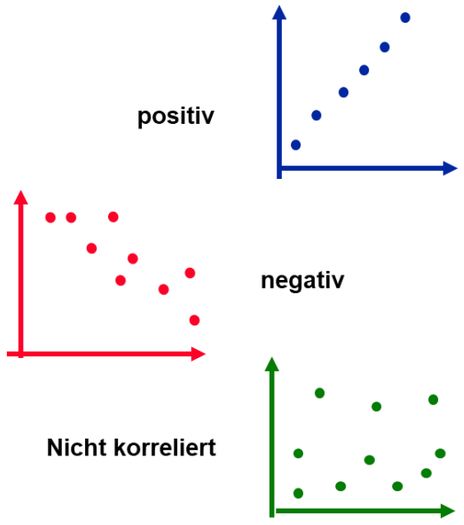
Principal Component Analysis (PCA)

Uniform Manifold Approximation and Projection (UMAP)

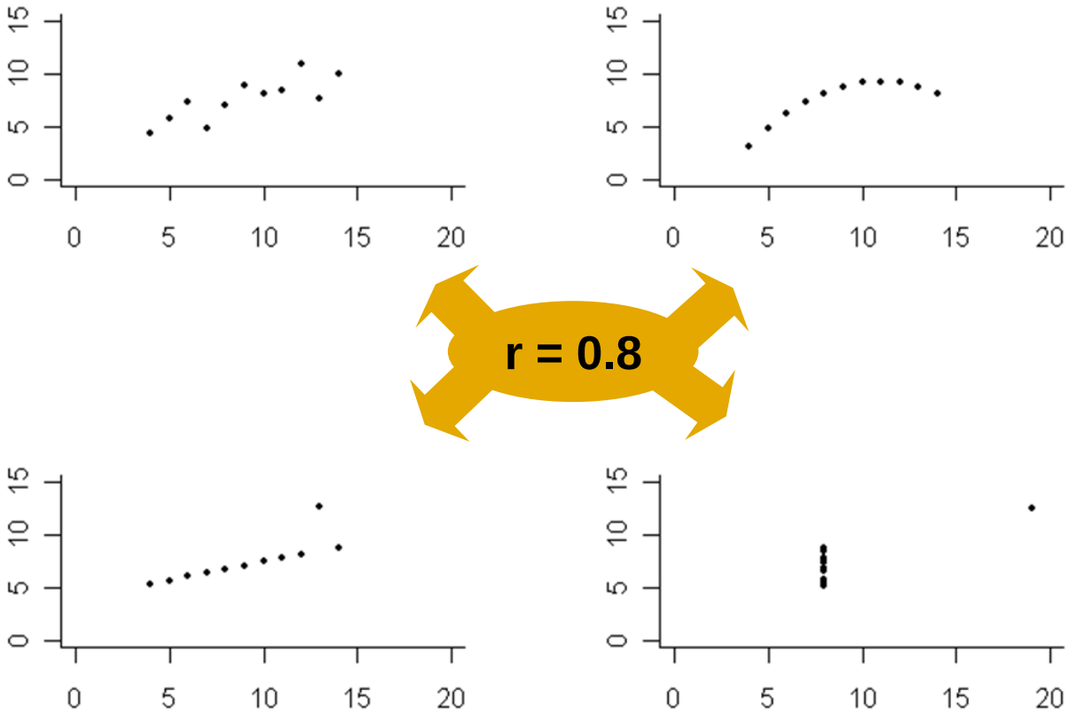
Visualisierungsmethoden (Andrew‘s Curves)

Korrelationsanalyse

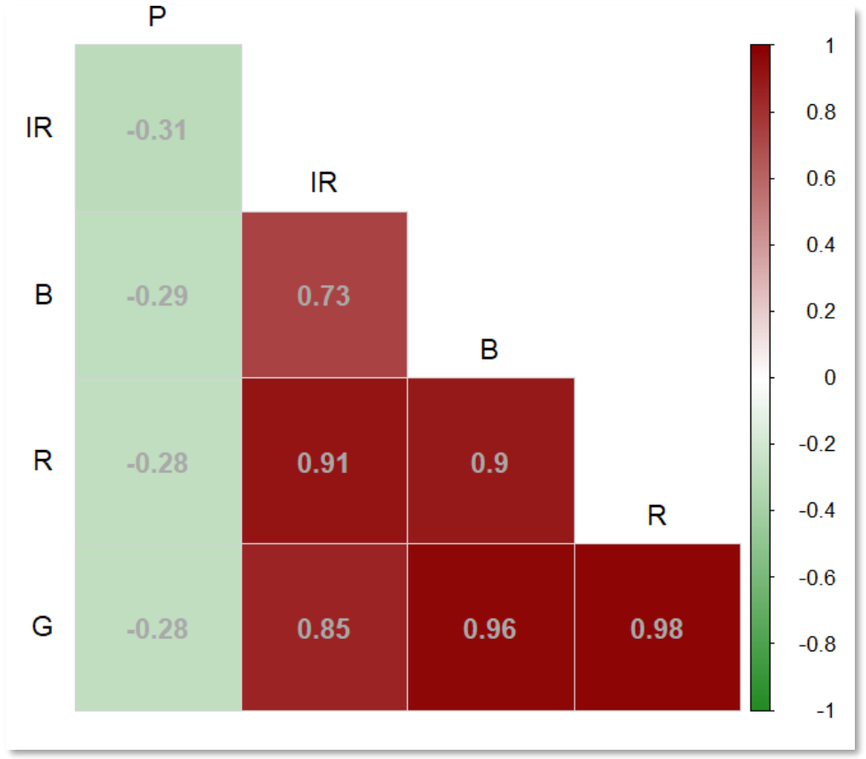
- Grad des (linearen) Zusammenhangs zwischen zwei Merkmalen / Datenvektoren
- Messwerte ➔ Pearson-Korrelationskoeffizient
- Ordinale Daten ➔ Spearman-Korrelationskoeffizient
- Wertebereich: $-1 \leq r \leq 1$
  - Positiv korreliert
    - $r > 0$
    - Direkt proportional
  - Negativ korreliert
    - $r < 0$
    - Indirekt proportional
  - unkorreliert
    - $r \approx 0$
    - kein linearer Zusammenhang

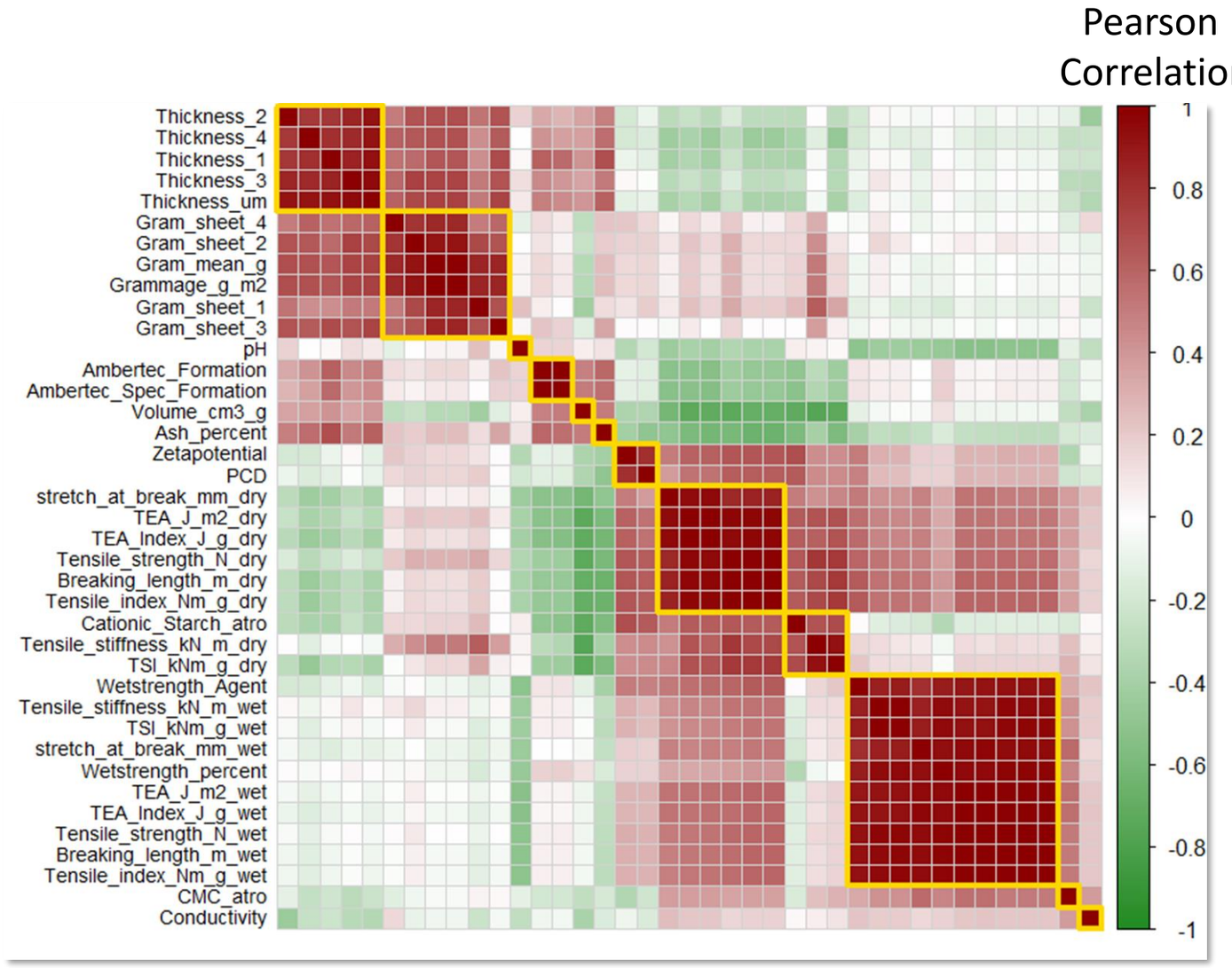


Immer Scatterplot ansehen!
Korrelationskoeffizient allein kann täuschen



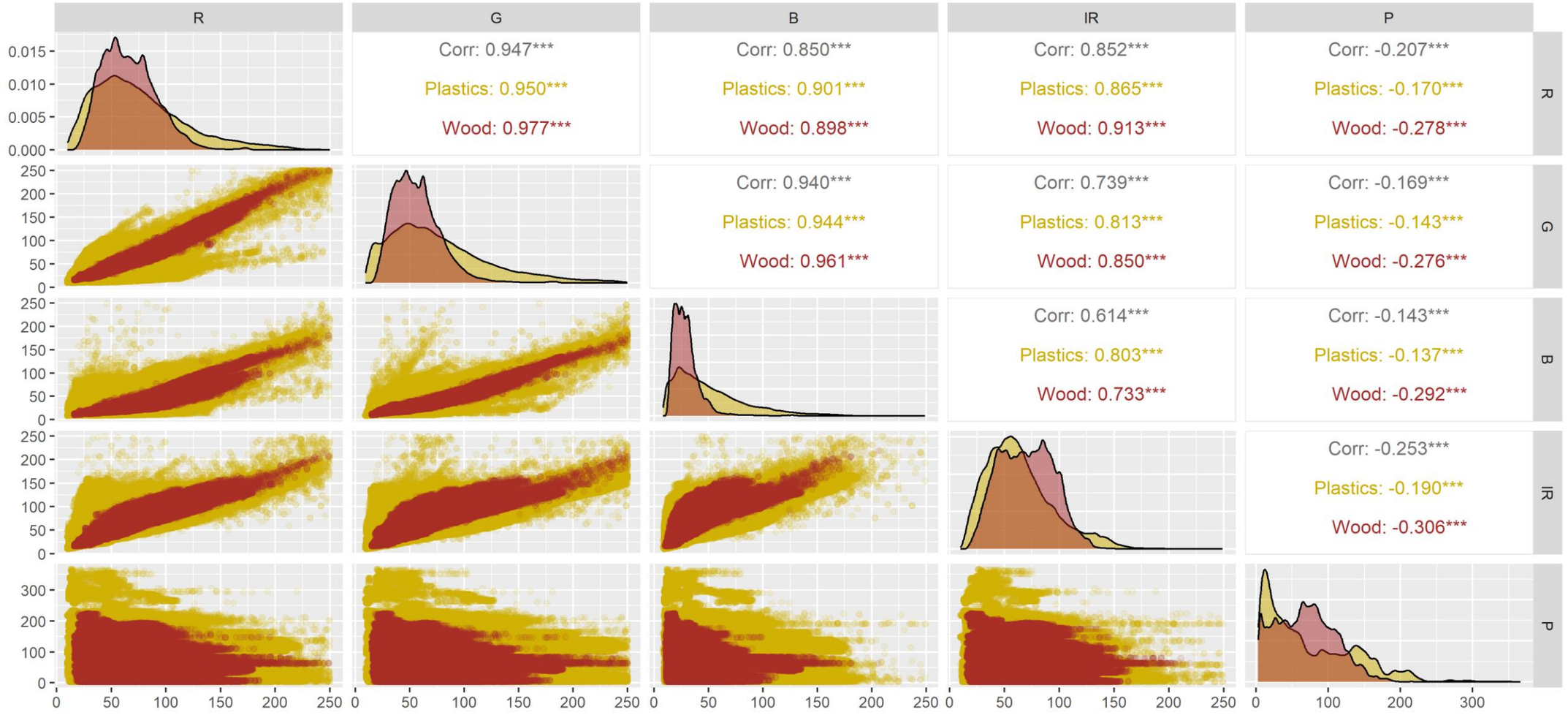
Visualisierung von Korrelationen ➔ Korrelationsmatrix als Plot (Heat Map)



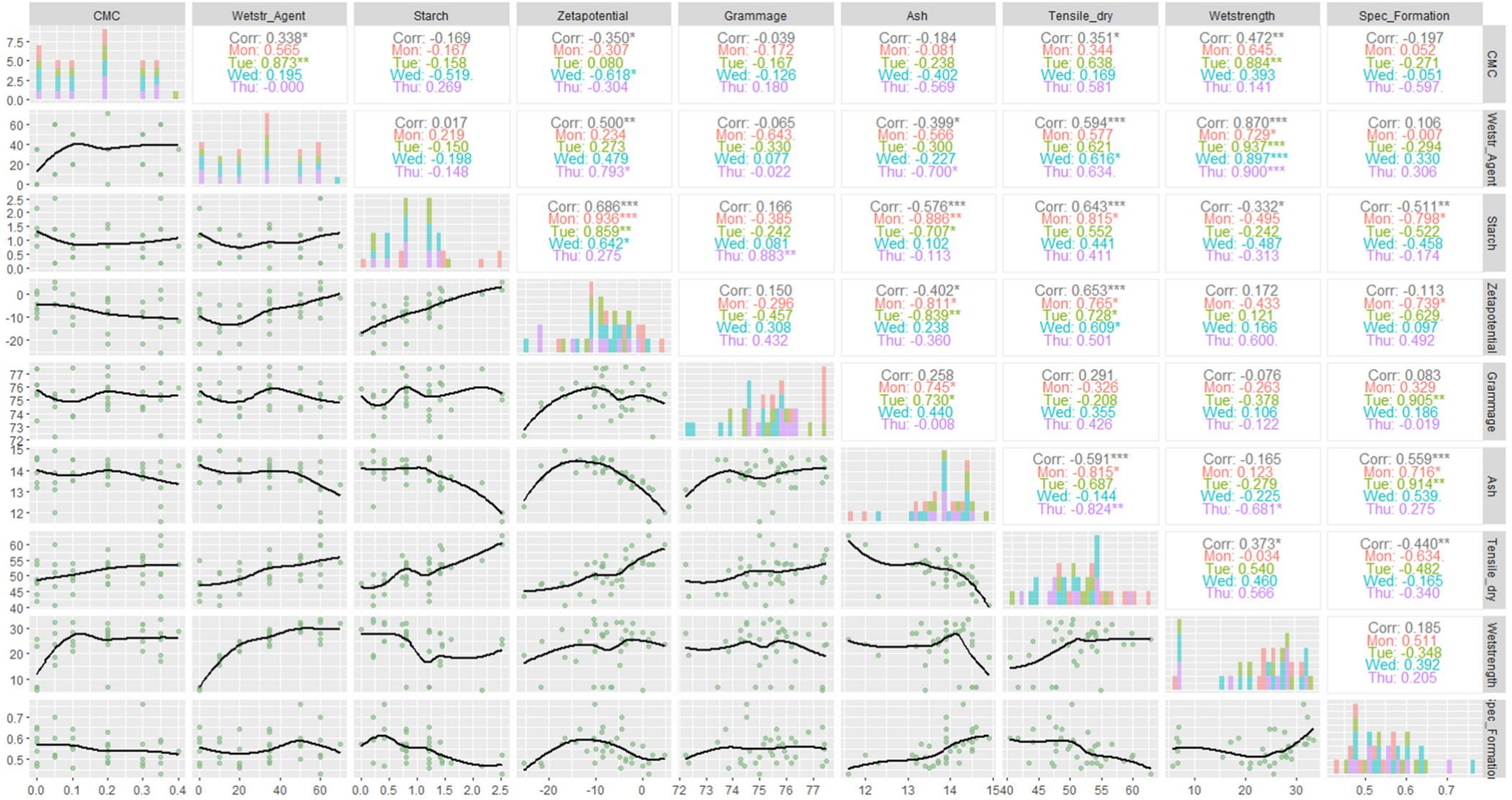


JOANNEUM RESEARCH POLICIES

Visualisierung von Korrelationen ➔ Scatterplot-Matrix



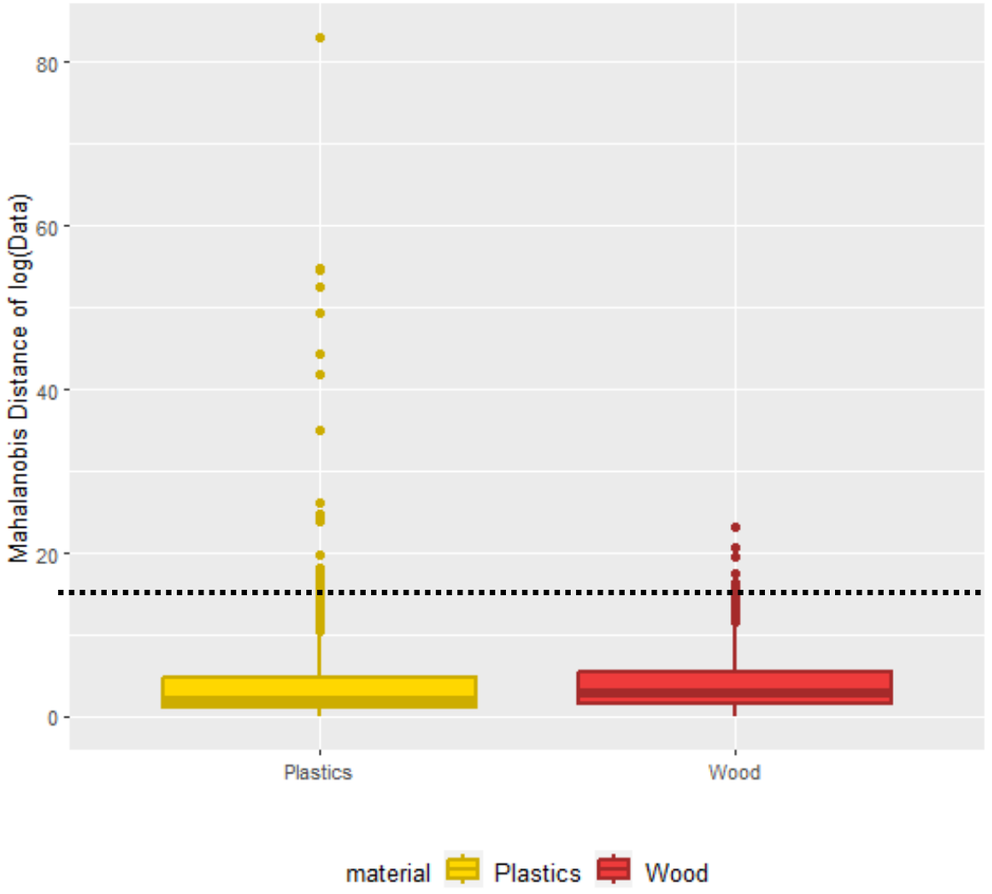
Visualisierung von Korrelationen ➔ Scatterplot-Matrix



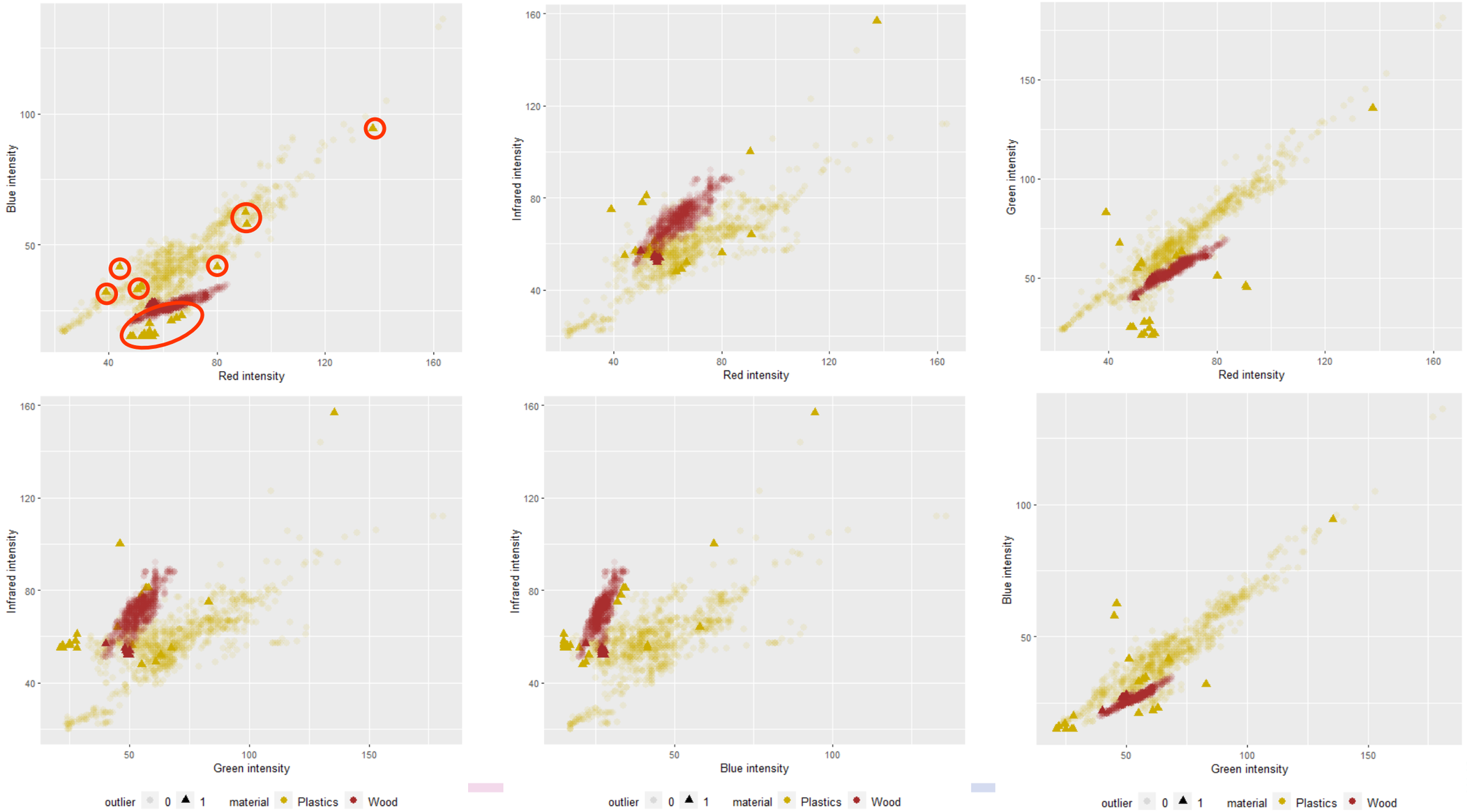
JOANNEUM RESEARCH POLICIES

Mahalanobis-Distanz

- Distanz zwischen 2 Punkten im multivariaten „Datenraum“
  - Meist wird Abstand zu multivariatem Mittelwert berechnet
  - Berücksichtigt Varianz der verschiedenen Merkmale und Kovarianz von je 2 Merkmalen ➔ „Kovarianzmatrix“
- Wozu?
  - Kann zur Identifikation von Ausreißern im multivariaten Sinn verwendet werden ➔ mit Hilfe von statistischem Test
  - Klassifikation und/oder Clustering der Beobachtungen basierend auf Mahalanobis-Distanz möglich
- Annahme einer multivariaten Normalverteilung der Daten
- Wenn keine NV:
  - „Robuste“ Varianten von Mittelwert- und Kovarianz-Schätzung anwenden
  - Daten transformieren

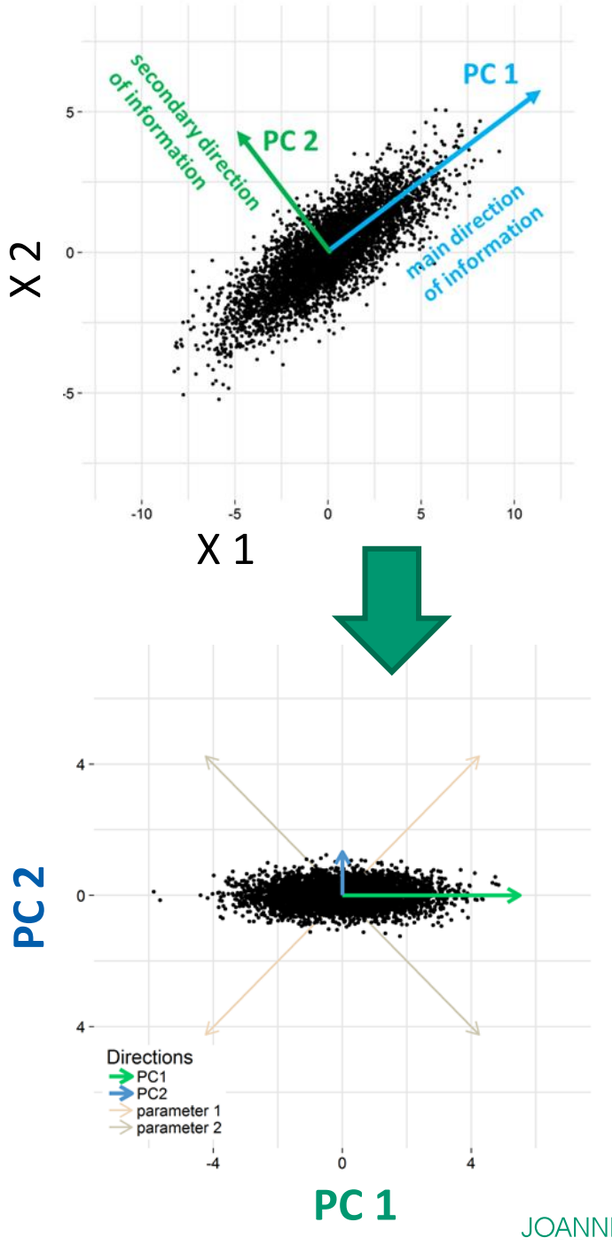


Mahalanobis-Distanz zur Ausreißer-Identifikation

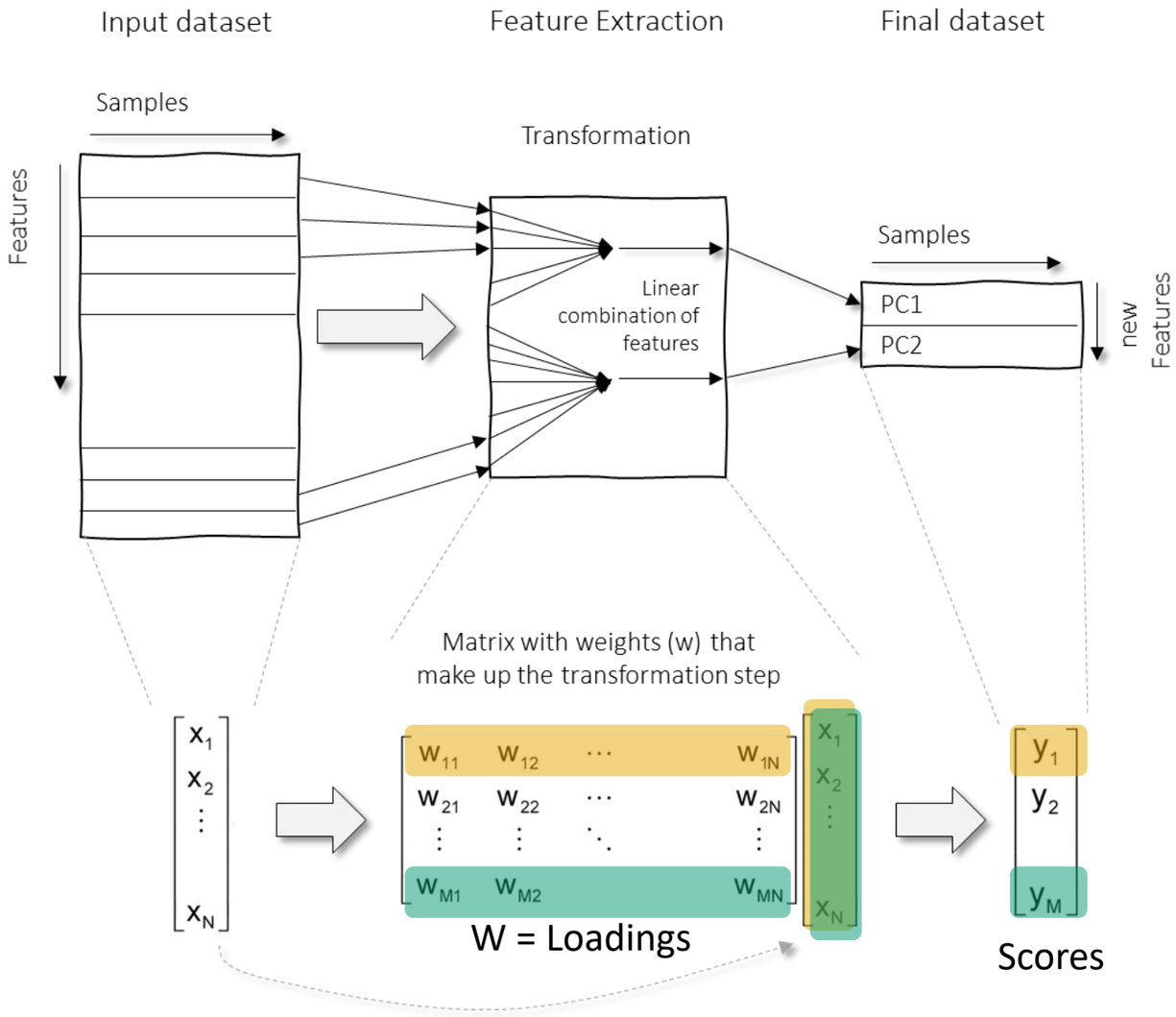


Principal Component Analysis (PCA)

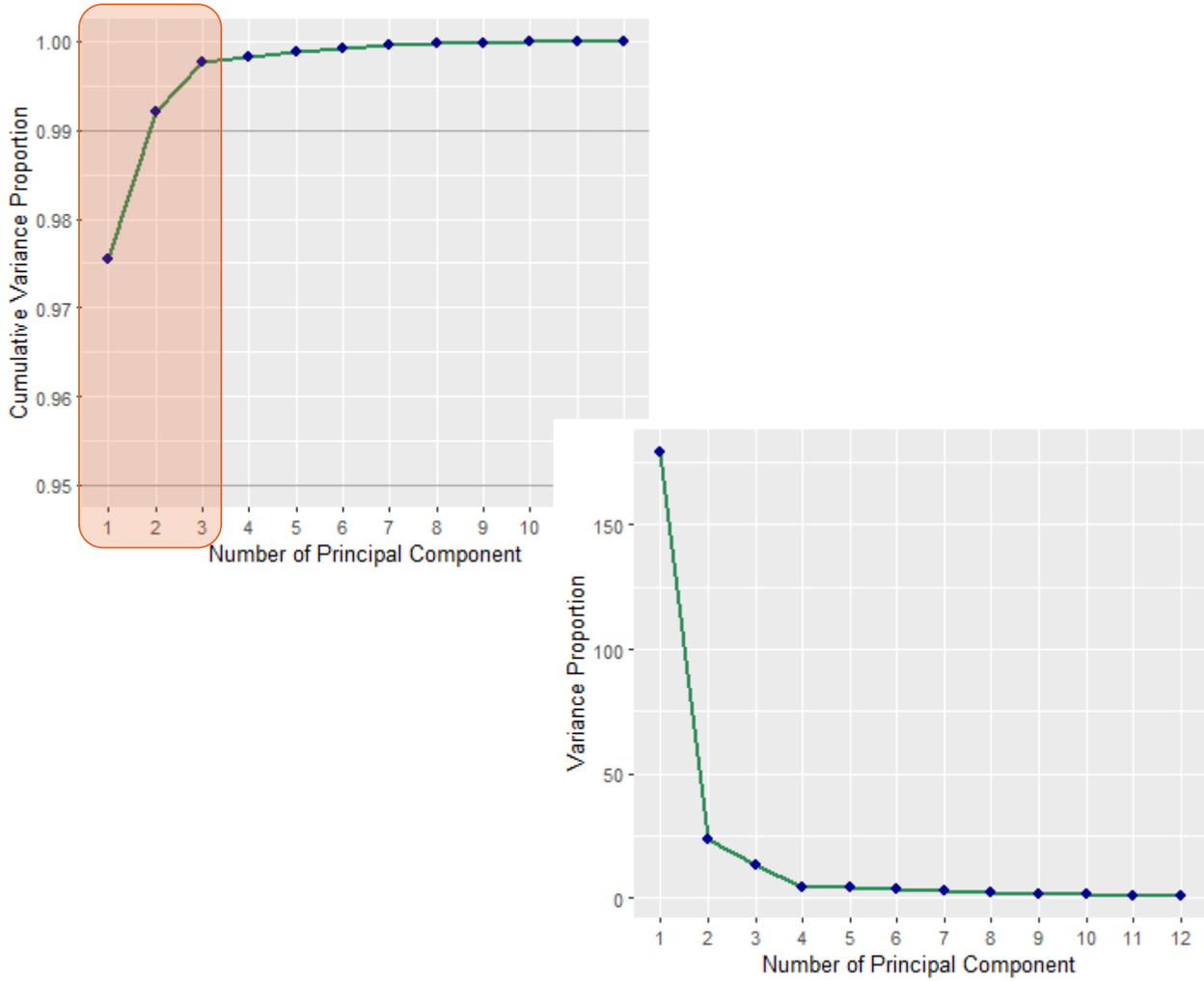
- Methode zur Dimensionsreduktion
  - Reduziert hochdimensionalen Datensatz auf wesentliche Informationen
  - Orthogonale Transformation in einen neuen Raum (bzw. anderes Koordinatensystem)
- Transformierte Daten / Hauptkomponenten (PCs)
  - Linear unabhängig
  - Haben maximale Variabilität
- Vorteil / Nutzen
  - Bei korrelierten Merkmalen kann durch wenige PCs ein hoher Anteil der Gesamtvariabilität „erklärt“ werden
  - ➔ Anzahl der relevanten PCs ist (meist deutlich) geringer, als Anzahl der ursprünglichen Merkmale
- Nachteil
  - PCs meist schwer interpretierbar
- Anwendung
  - Basis für Regression, Klassifikation, Clustering



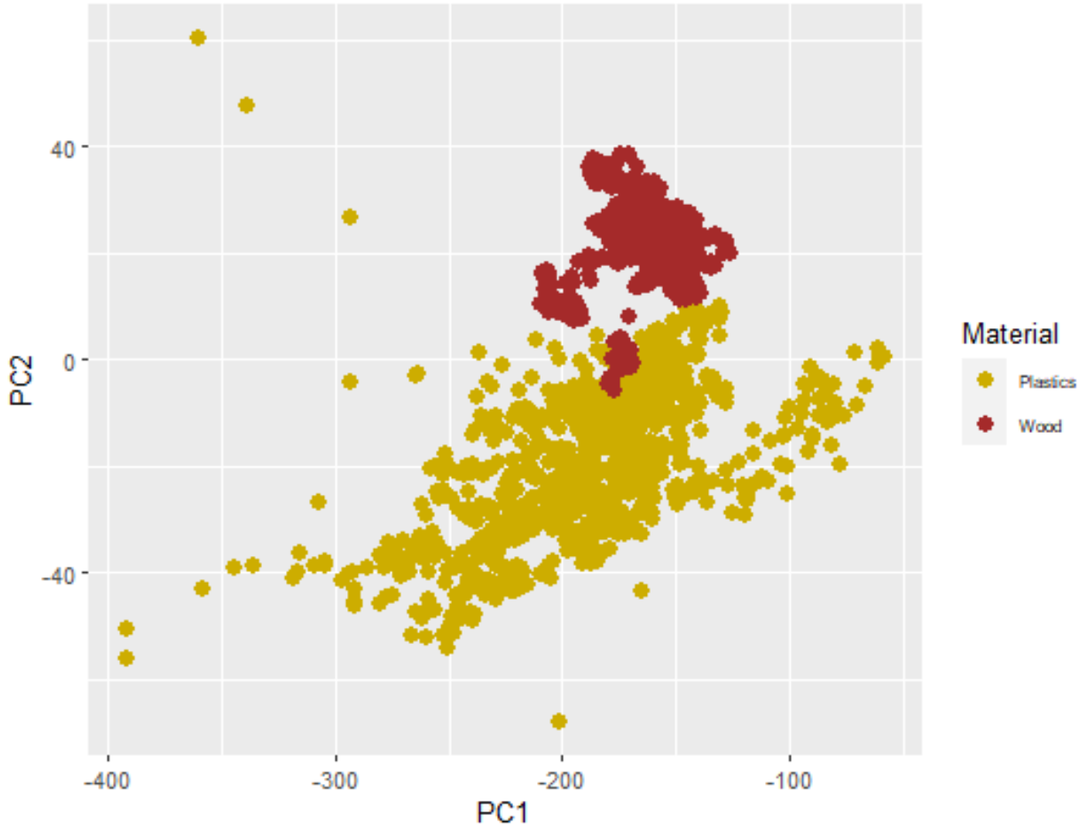
Wie funktioniert PCA?

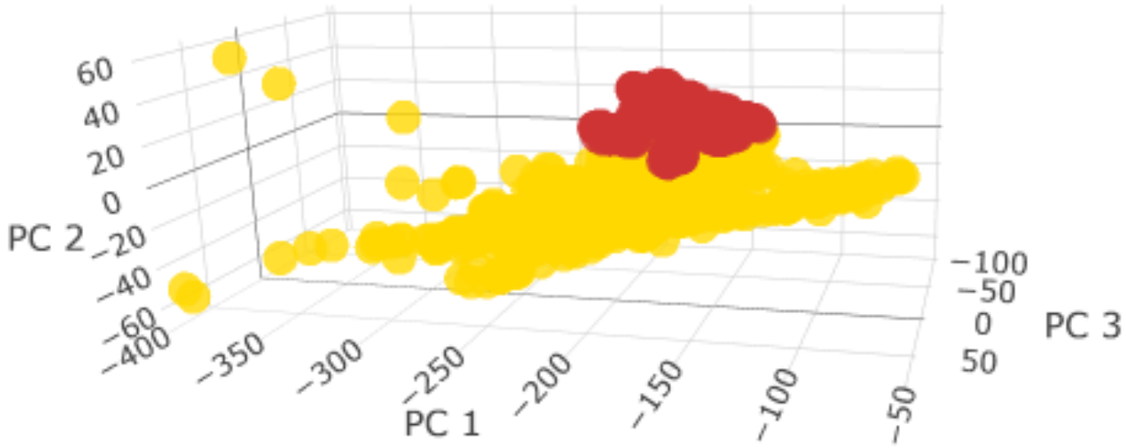


Wie viele PCs werden benötigt?



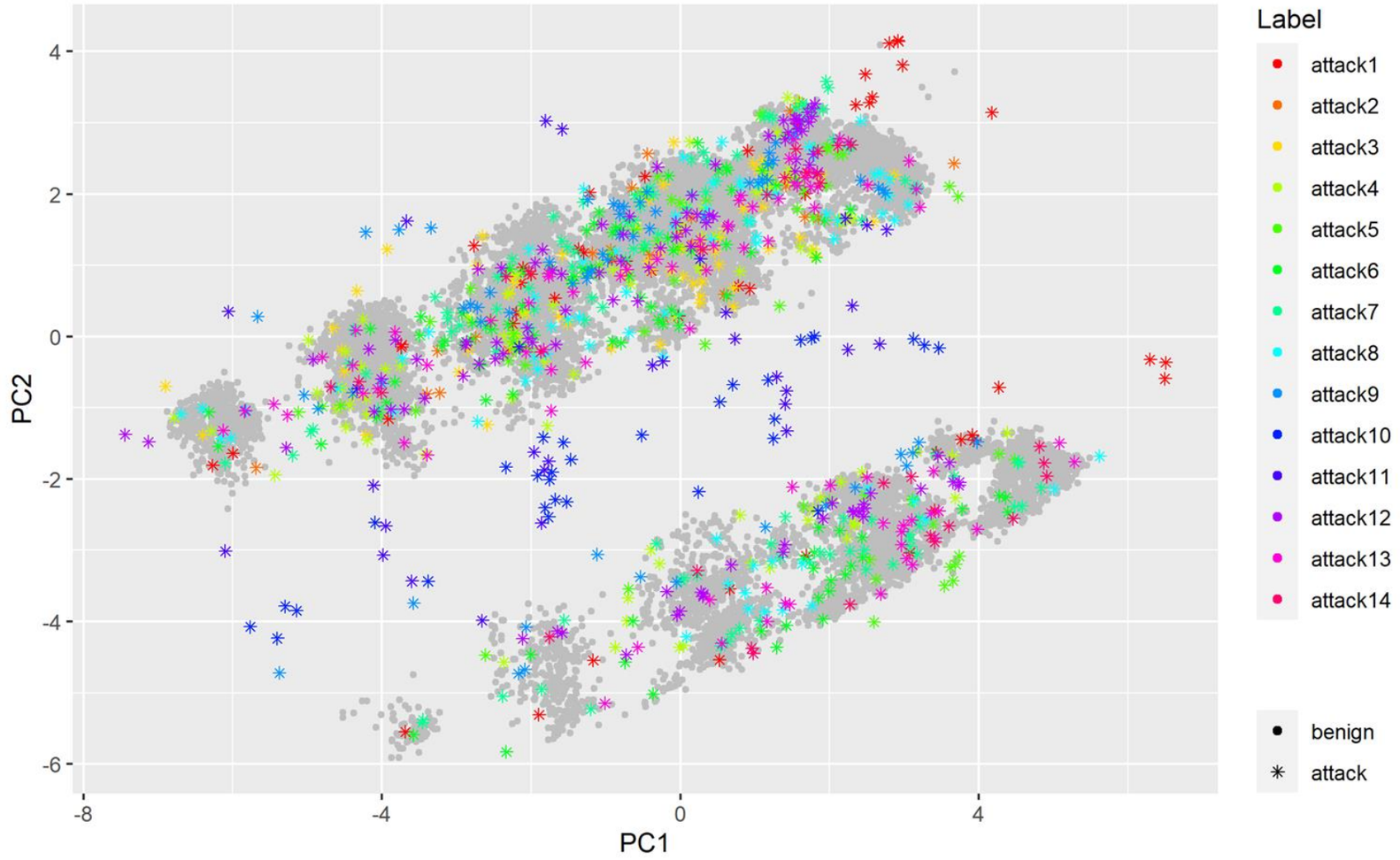
Visualisierung von PCA Ergebnissen





pics\pca_3d.html

Visualisierung von PCA Ergebnissen

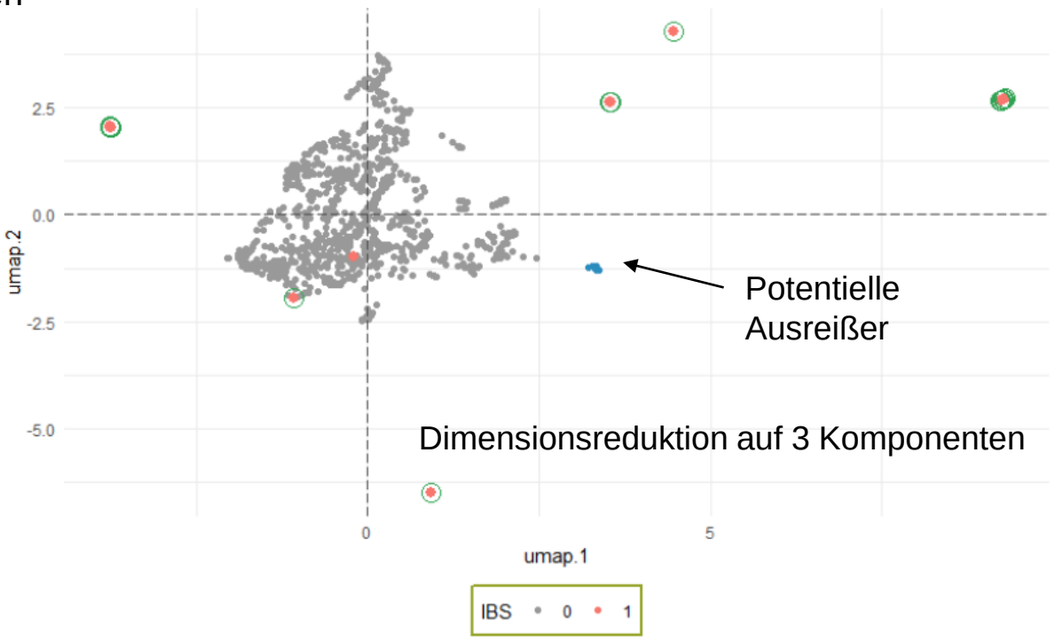


Uniform Manifold Approximation and Projection (UMAP)

- Methode zur nichtlinearen Dimensionsreduktion (McInnes et al., 2018)
  - Reduktion eines hochdimensionalen Datensatzes in eine niedrigere Dimension (ohne relevante Informationen über die Datenstruktur zu verlieren)
  - Keine Interpretierbarkeit der resultierenden Komponenten
  - Vorteil: große Datenmengen und geringe Rechenzeit
  - Basiert auf komplexen, theoretischen Grundlagen

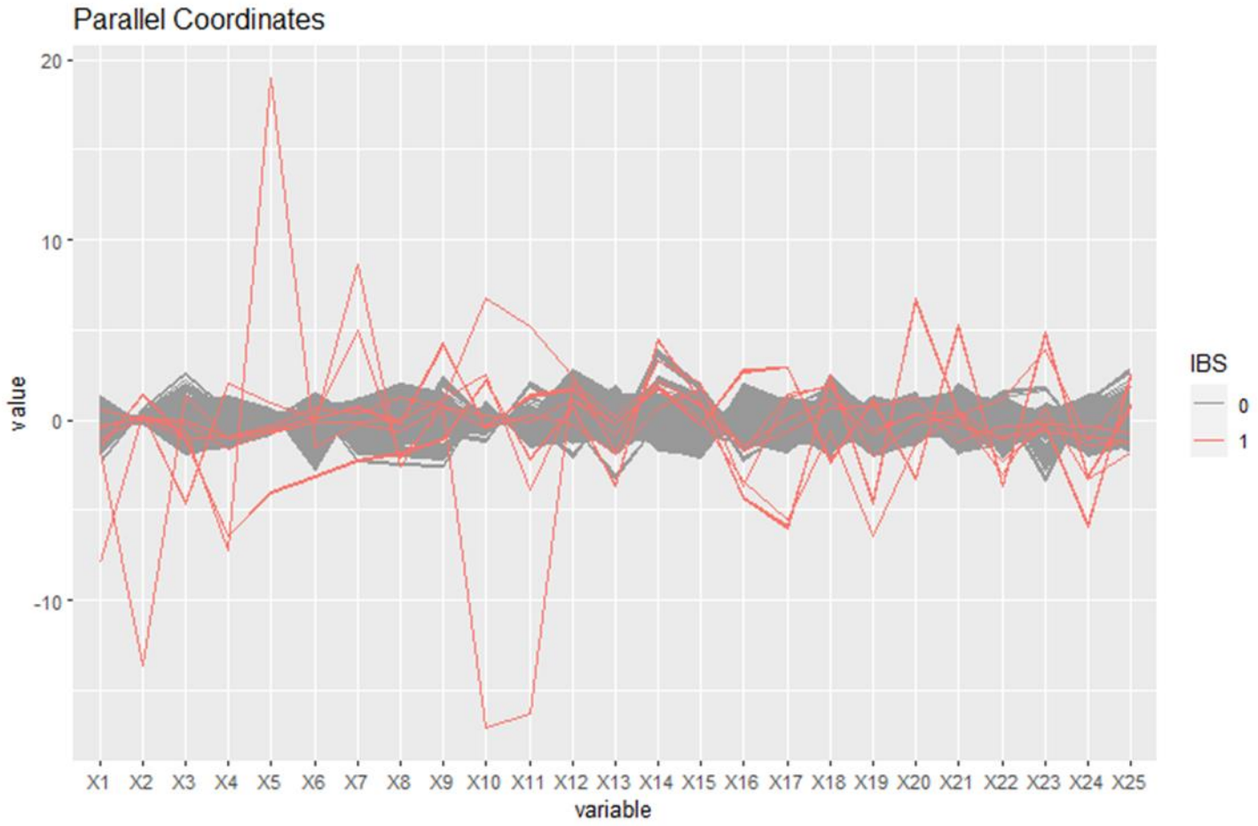
    → Repräsentation der Daten durch einen hochdimensionalen Graphen
    → Optimierung eines niederdimensionalen Graphen, der strukturell möglichst ähnlich ist

- Grundlage für weiterführende Methoden
  (z.B. Clustering, Identifikation von Ausreißern, etc.)
- Visualisierung der Ergebnisse als Andrew‘s Curves
- Vorteile gegenüber t-SNE (t-Distributed Stochastic Neighbor Embedding)
  - Dimensionsreduktion nur in einen 2- oder 3-dimensionalen Raum



Dimensionsreduktion auf 3 Komponenten

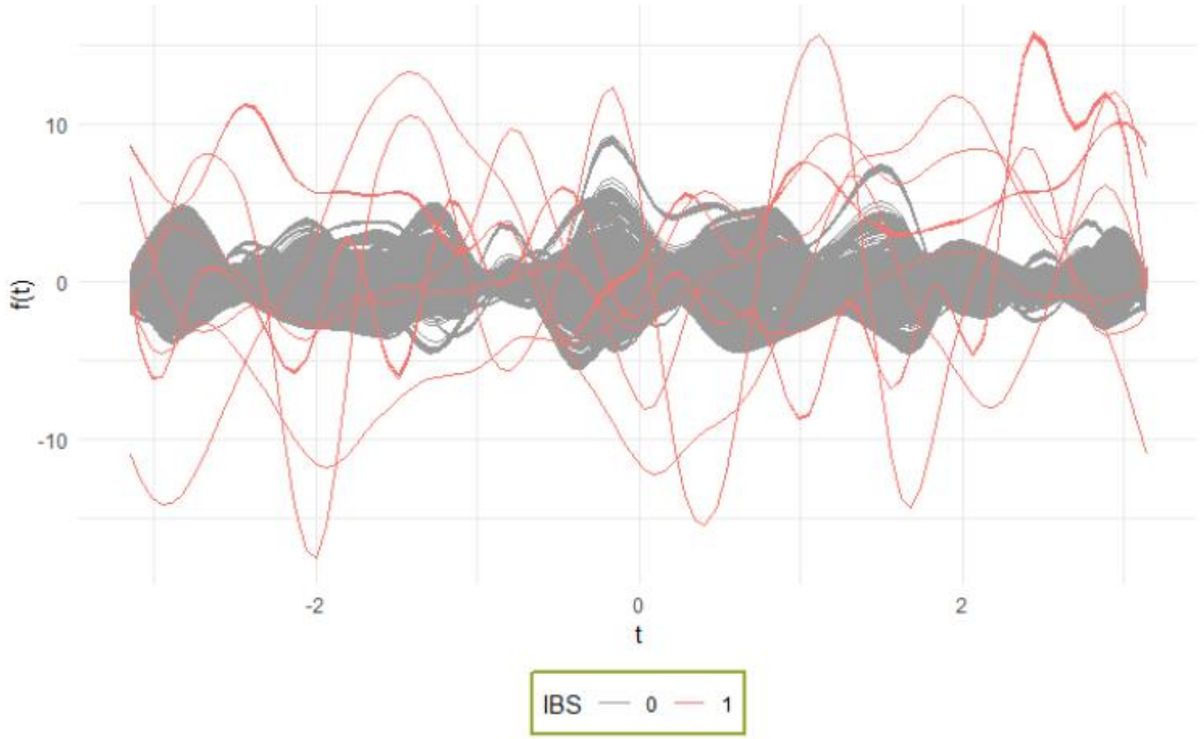
Parallele Koordinaten (nach UMAP)



- Dimensionsreduktion mit UMAP
  - 248 Merkmale → 25 Merkmale
- Visualisierung: Parallele Koordinaten

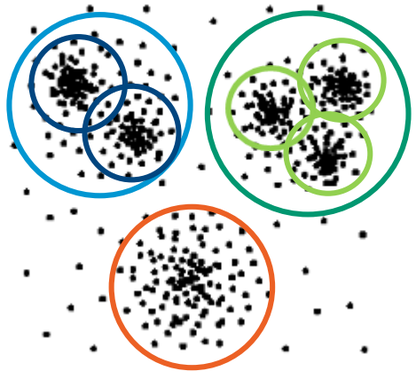
→ Unterschiede zwischen den Objekten deutlich erkennbar

Andrew‘s Curves (nach UMAP)



- Dimensionsreduktion mit UMAP
  - 248 Merkmale → 25 Merkmale
- Transformation und Visualisierung: Andrew‘s Curves

Clustering



- Ziel: Erkennen von „ähnlichen“ (homogenen) Objekten (Cluster) in Daten
  - Objekte innerhalb eines Clusters sollen möglichst ähnlich sein
  - Objekte aus verschiedenen Clustern sollen möglichst unähnlich zueinander sein
- Dimensionsreduktion durch Strukturierung der Daten
- „Unsupervised Classification“ (keine vordefinierten Klassen)

Hierarchische Clusteranalyse

- Distanz- bzw. Ähnlichkeitsmaße
- Hierarchie von Clustern

Agglomerative Verfahren

Divisive Verfahren

Partitionierende Clusteranalyse

- Anzahl an Cluster k zu Beginn festgelegt (vorgegeben)
- Cluster durch Cluster-Zentren repräsentiert
- Minimieren einer Zielfunktion

K-Means

Dichtebasierte Clusteranalyse

- Cluster = Datenpunkte in einer Region mit hoher Datendichte

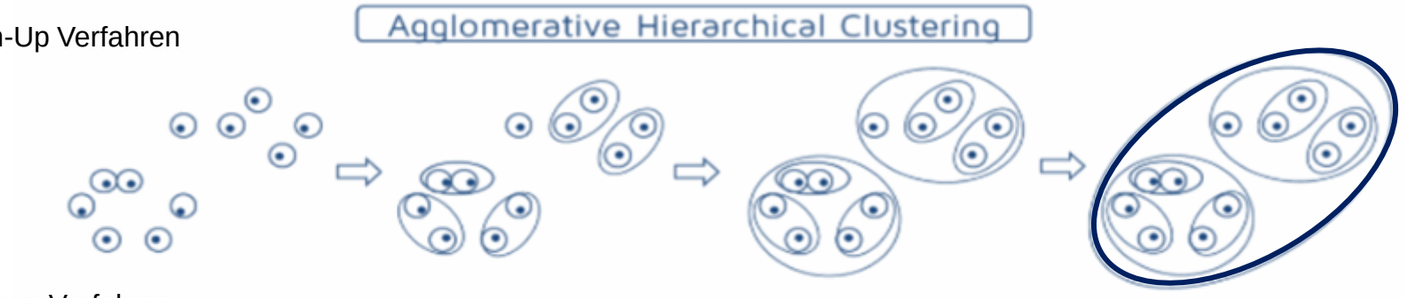
dbscan

hdbscan

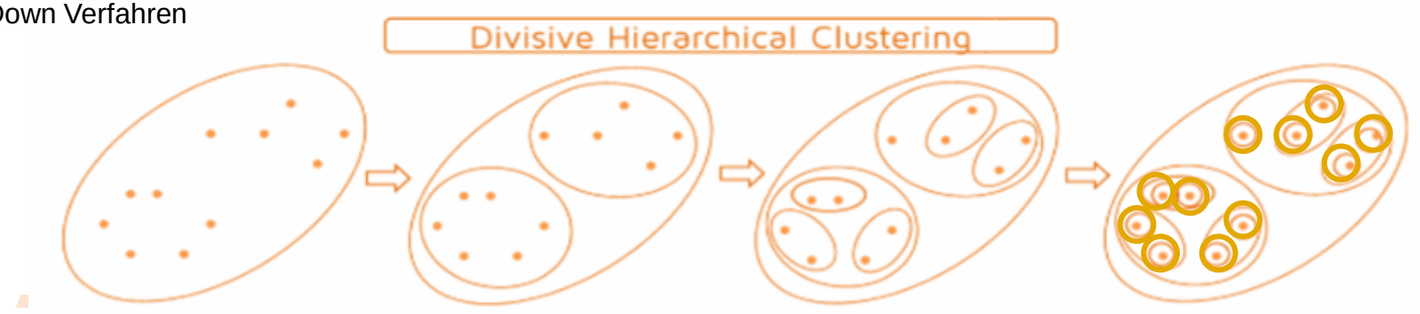
Hierarchisches Clustering

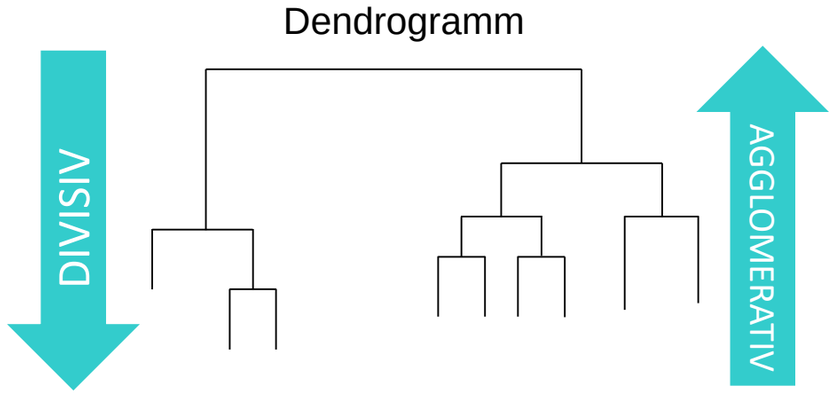
- Ähnlichkeit (bzw. Unähnlichkeit) wird auf Basis von Merkmalen definiert
  - Distanzmaße bzw. Ähnlichkeitsmaße
  - Distanzmatrix basierend auf Distanzmaß (Euklidische Distanz, Manhattan-Distanz, Mahalanobis-Distanz, etc.)
  - Mögliche Probleme: ungleiche Wertebereiche der Merkmale (Standardisierung), unterschiedliches Skalenniveau der Merkmale
- (Strikte) Hierarchische Zuordnung der Daten zu den (untereinander genesteten) Clustern

Bottom-Up Verfahren



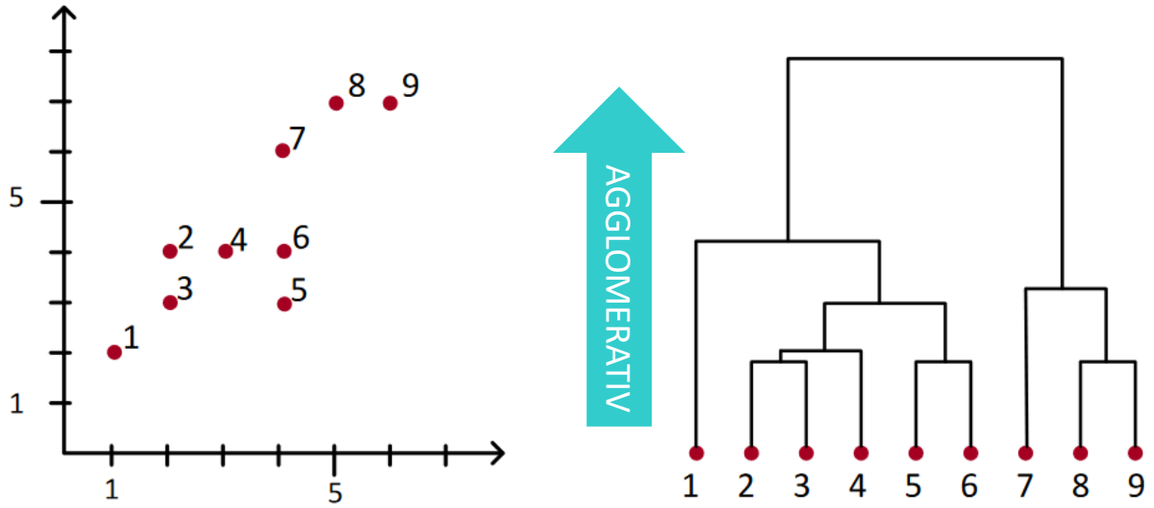
Top-Down Verfahren





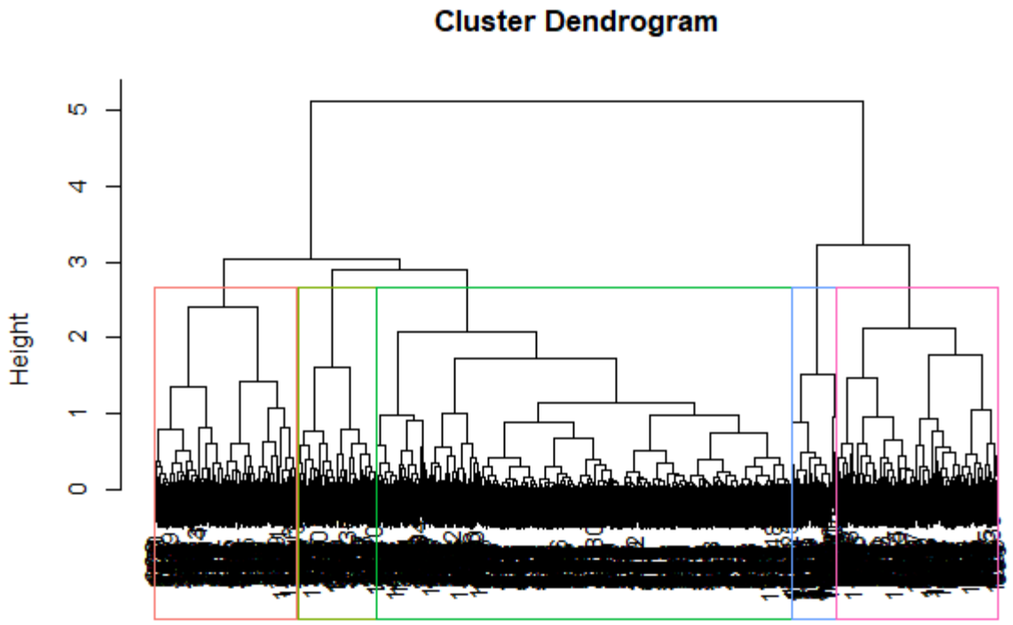
Hierarchisches Clustering

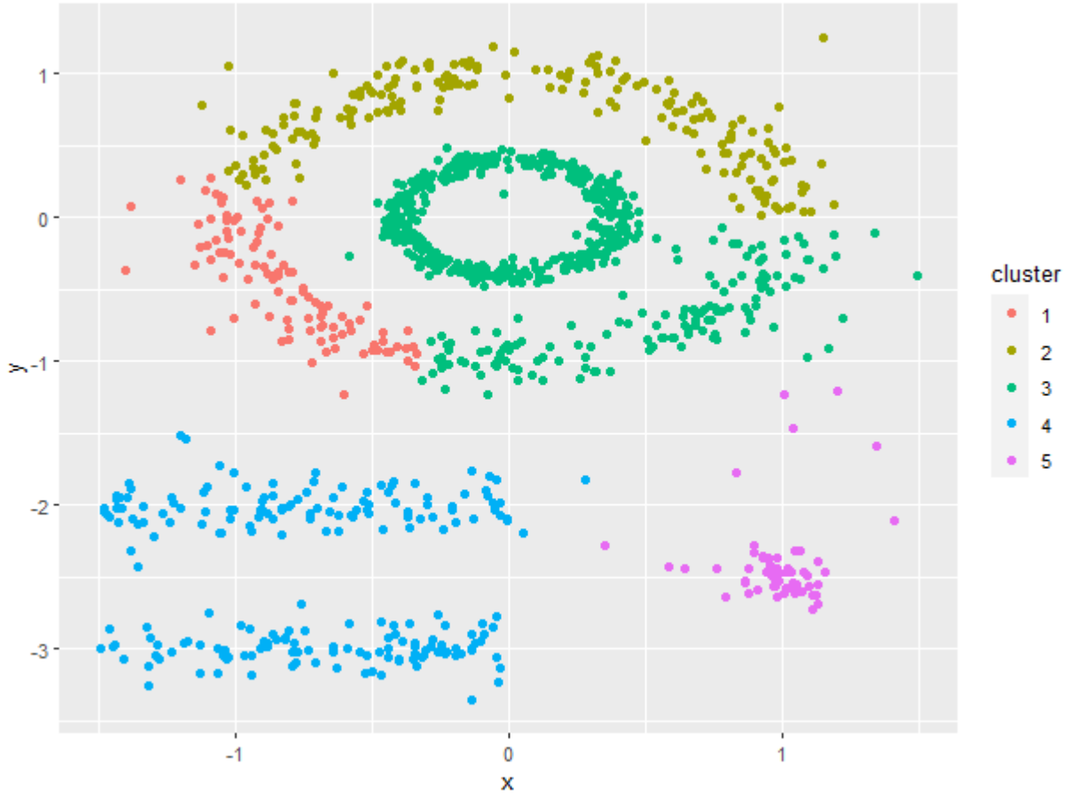
- Vorgehensweise (agglomerativ)
  - Paarweise Distanzen zwischen den Clustern (Objekten=Beginn)
  - Zusammenfassen des engsten Paares (A,B) → C = A ∪ B (d.h., Fusionierung der beiden Cluster, die die geringste Distanz zueinander haben) basierend auf Linkage-Methode (Single-Link, Complete Link, Average Link etc.)
  - Berechnung der Distanz zwischen C und den anderen Objekten
  - STOP: bis C alle Objekte umfasst



Explorativer Ansatz →
„beste“ Anzahl an Clustern von Anwendung abhängig

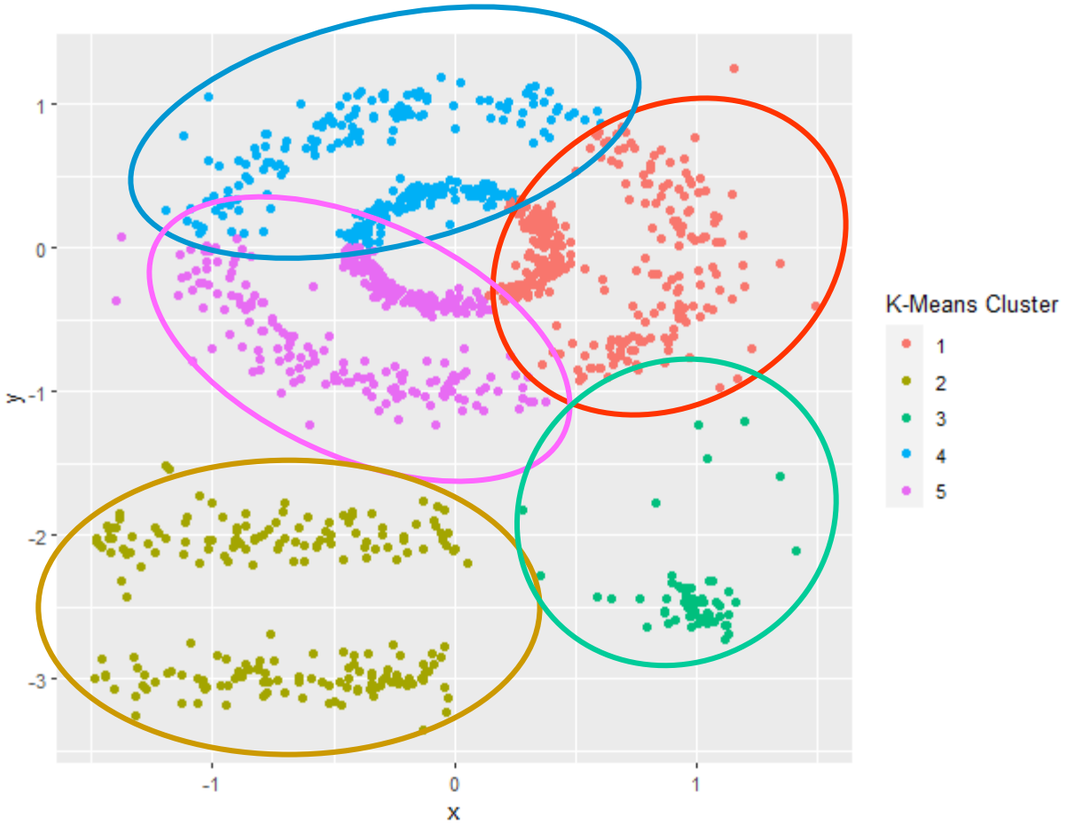
Hierarchisches Clustering





JOANNEUM RESEARCH POLICIES

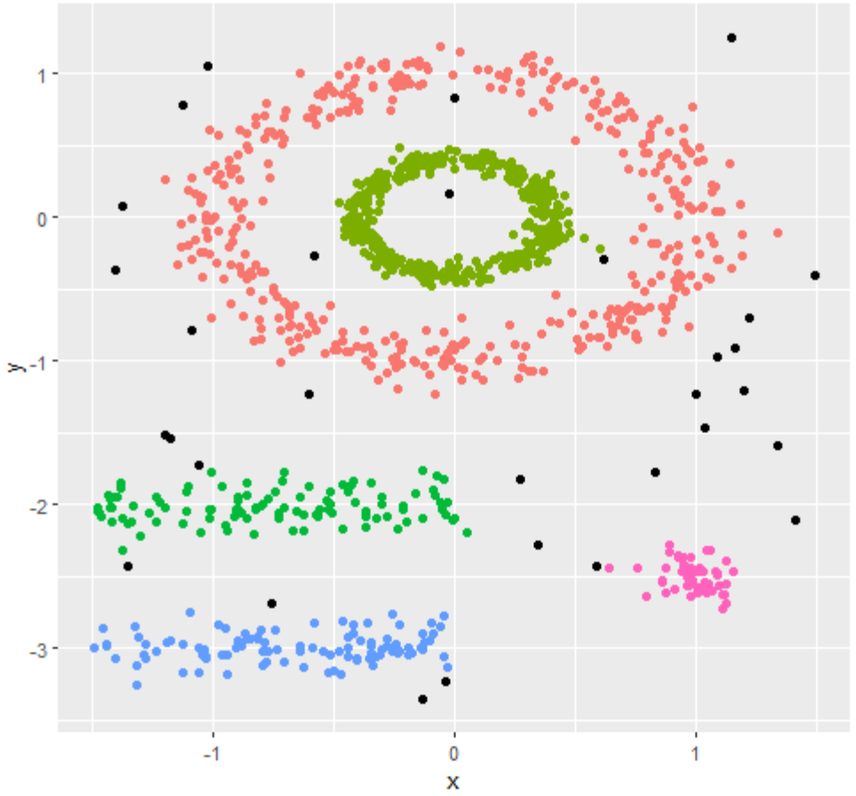
K-Means Clustering



- + Effiziente Methode
- + Einfache Implementierung

- - Anzahl Cluster oft schwer zu bestimmen
- - Sensitive Methode (Ausreißer) (!)
- - Cluster müssen konvexe Form haben
- - Abhängig von Initialisierung

dbscan Clustering



von Wahl der Parameter
ε und MinPts abhängig(!)

Ausreißer
(Cluster 0)

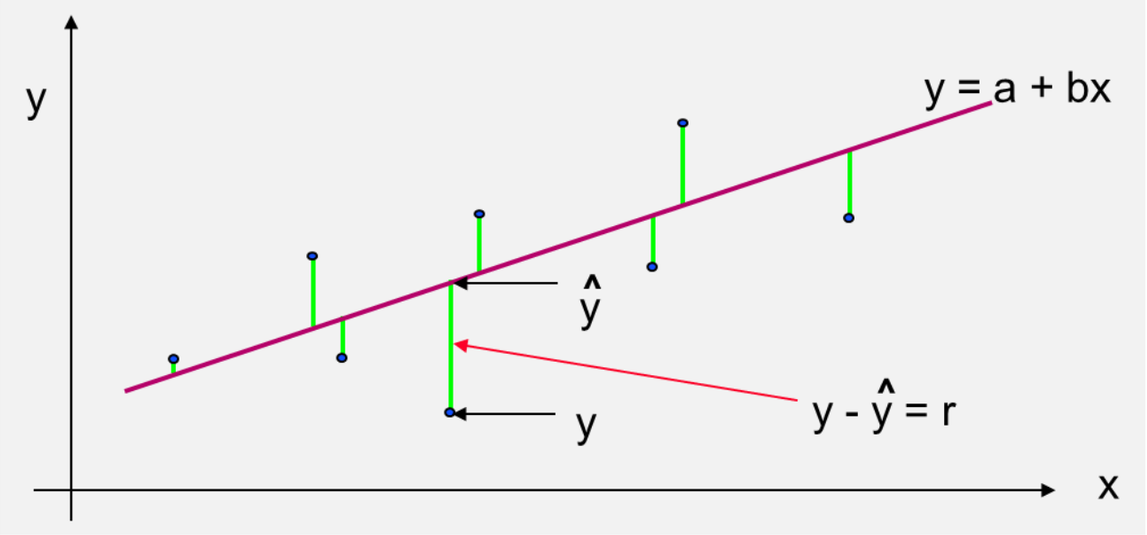
Erweiterung:
→ **hdbscan** (Hierchical Density-Based Spatial Clustering): kein Parameter ε

Modellierung

- Ziele
  - Beschreiben von Zusammenhängen zwischen Input-Merkmalen X und Output Y mit Hilfe eines Rechenmodells
  - Prädiktion / Vorhersage von Output Y basierend auf Input Y

Regressionsanalyse

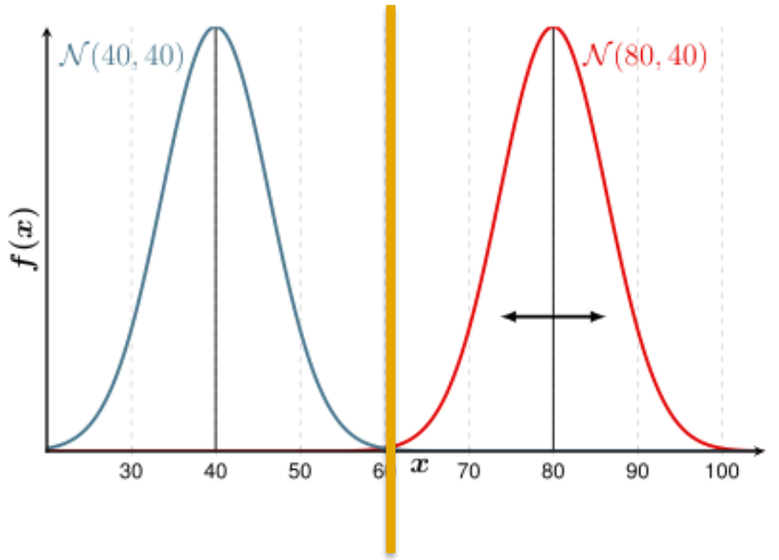
- Frage
  - Gibt es einen funktionalen Zusammenhang zwischen Einflussfaktor x und Zielgröße y?
- Regressionsanalyse
  - Schätzen der Parameter der Funktion
    y = f(x) + e
  - Einfache lineare Regression
    - y = a + bx + e
  - Residuen (Fehler) =
    gemessene Werte – geschätzte Werte
- Schätzmethode der kleinsten Quadrate (Gauss)
  - Bestimme a und b, sodass die quadrierten Residuen des Modells minimal werden

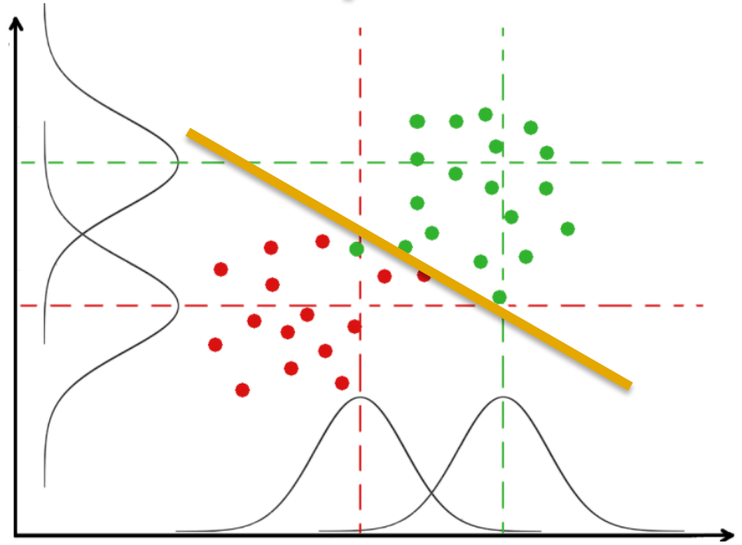


- Maß für die Modellgüte
  - Bestimmtheitsmaß = Erklärungsgrad = R^2 = R-Squared
    - $0 < R^2 < 1$
    - Sehr gute Anpassung R^2 nahe 1
    - Schlechte Anpassung R^2 nahe 0
    - ➔ Faustregel $R^2 > 0.5$

Diskriminanzanalyse (DA)

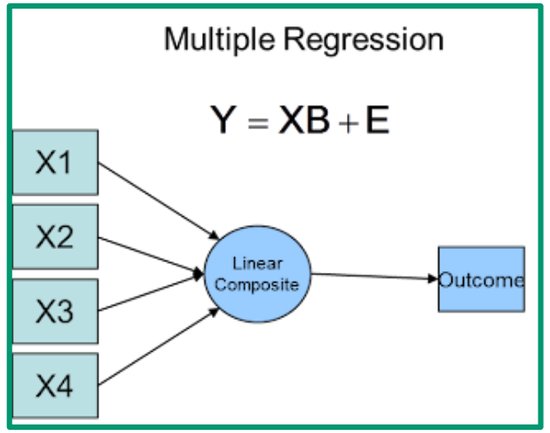
- Verfahren zur Klassifikation von Daten
  - Unterscheidung von 2 oder mehr Gruppen
  - Jedes Objekt gehört genau einer Klasse an
- Verwandt mit PCA
  - DA berücksichtigt Klassenzugehörigkeit!
  - ➔ „supervised learning“
- Varianten
  - Lineare DA (LDA)
  - Quadratische DA (QDA)
  - Kerndichteschätzer
  - …

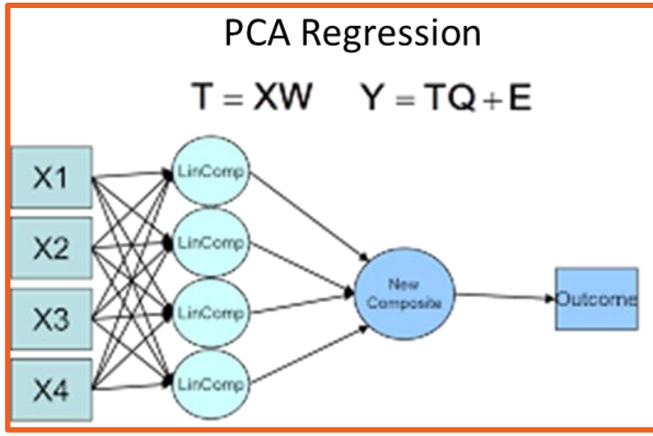


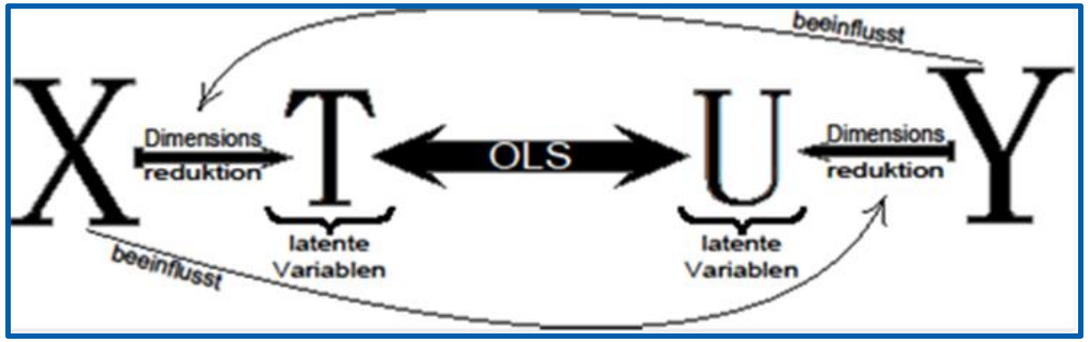


Partial Least Squares (PLS) Regression

- Dimensionsreduktion kombiniert mit Regression
  - Sowohl Input X als auch Zielgröße Y werden in neuen Raum transformiert
  - Klassenzugehörigkeit als Zielgröße ➔ PLS-DA
- Ähnlich PCA Regression
  - ➔ Regression basierend PCs
- Vorteil gegenüber PCA Regression
  - Zielgröße „steuert“ Bestimmung der latenten Variablen
- Anwendung
  - Bei großer Anzahl von Merkmalen, z. B. mehr Variablen als Beobachtungen
  - Bei hoch korrelierten Merkmalen
  - Sehr verbreitet im Bereich Chemometrics
- … auch bekannt als „Projection to Latent Structures“ (PLS)







Diskussion & Fragen

JOANNEUM RESEARCH POLICIES

Use Cases

- Mikrobiom Daten Ortho-Analytic
- ?

Kontakt:

Michaela Dvorzak, Ulrike Kleb

JOANNEUM RESEARCH
Forschungsgesellschaft mbH

POLICIES
Institute for Economic and Innovation Research

Leonhardstrasse 59
8010 Graz

Tel. +43 316 876-1555
ulrike.kleb@joanneum.at

www.joanneum.at/policies

Analyse und Modellierung von Daten

Datenaufbereitung

- Prüfen der Datenqualität
- Plausibilität prüfen
- Imputation von fehlenden Werten
- Korrigieren von Fehlern
- Berechnen neuer Merkmale / Codieren
- Aggregieren
- **Transformieren**

Exploration und Beschreibung

- Beschreibung durch Kennzahlen / deskriptive Methoden
- Korrelationsanalyse
- Visualisierung
  - Scatterplots (2D/3D)
  - Histogramme
  - Boxplots
  - Linienplots

Modellierung und Inferenz

- Statistische Tests
- Half Normal Plot
- Varianzanalyse (ANOVA)
- Lineare Modelle
- Regression / Response Surface Models
- **Validierung / Bestätigungsexperiment ➔ ggf. weitere Versuche und/oder Modellanpassung**

Simulation und Optimierung

- Prädiktion der Zielgröße(n) für gesamte Region of Interest (Versuchsraum)
- Suche nach lokalem Optimum für Zielgröße(n)
- Ggf. Erweiterung des Versuchsraums und neue Versuche