Multivariate Verfahren

Einführung und grafische Darstellung

Annika Hoyer

Sommersemester 2020

Einführung - Inhalt

Organisatorisches

Beispiele multivariater Datensätze

Beschreibung und Darstellung multivariater Datensätze

Möglichkeiten in R

Organisatorisches

Organisatorisches

Vorlesung:
Mo, 12:15 – 13:45 Uhr, Hauptgebäude, A021
Di, 10:15 – 11:45 Uhr, Hauptgebäude, A025
Annika Hoyer (annika.hoyer@stat.uni-muenchen.de)

Übung:
Do, 12:15 – 13:45 Uhr, Hauptgebäude, A014
Marc Schneble (marc.schneble@stat.uni-muenchen.de)

Tutorium: (freiwillig)
Mo, 14:15 – 15:45 Uhr, Hauptgebäude, A015
Di, 12:15 – 13:45 Uhr, Hauptgebäude, A015
Cornelia Gruber

Organisatorisches

Klausuren:

- In der letzten Vorlesungswoche oder später
- Nacholklausur: Ende des Semesters (geplant)

Formale Aufteilung

- **Multivariate Verfahren (9 ECTS):** betrifft z.B. Bachelor Statistik nach neuer/alter Prüfungsordnung.
- **Grundlagen der multivariaten Verfahren (6 ECTS):** betrifft z.B. Statistik als Nebenfach von Bachelor- und Masterstudiengängen nach neuer Prüfungsordnung. Ebenso z.B. Master Biostatistik / Wirtschafts- und Sozialstatistik nach neuer Prüfungsordnung.
- **Fortgeschrittene multivariate Verfahren (3 ECTS):** betrifft z.B. Master Quereinsteiger nach neuer Prüfungsordnung.

Praktische Aufteilung

- *Eine* Veranstaltung **Multivariate Verfahren**
  - Stoff bis ca. Juni: **Grundlagen**
  - Stoff ab ca. Juni: **Fortgeschrittene multivariate Verfahren**
- Mehrere Klausuren, aber alle parallel (geplant: letzte VL-Woche)
  - Multivariate Verfahren komplett (120 min)
  - Multivariate Verfahren Grundlagen (60 min)
  - Multivariate Verfahren Fortgeschritten (60 min)

Ziele der Veranstaltung

Erlernen multivariater Analysetechniken:

- Darstellung mehrdimensionaler Datensätze
- Auffinden von Assoziationsstrukturen
- Auswertung mithilfe von Clusteranalyse, Diskriminanzanalyse, Hauptkomponentenanalyse, ...

Lernziel: Kompetenzen erwerben ...

... um bei gegebener Datenlage geeignete Auswertungsmethoden zu identifizieren, anzuwenden und sich der Voraussetzungen und Einschränkungen bezüglich der Anwendungen bewusst zu sein.

Literatur

Literatur

- Fahrmeir, Hamerle, Tutz (1996): *Multivariate statistische Verfahren*. DeGruyter, Berlin
- Johnson, Wichern (2007): *Applied Multivariate Statistical Analysis*. Pearson, London
- Everitt, Hothorn (2011): *An Introduction to Applied Multivariate Analysis with R*. Springer, New York
- Schlittgen (2009): *Multivariate Statistik*. Oldenbourg, München
- Handl (2010): *Multivariate Analysemethoden*. Springer, Berlin Heidelberg

Überblick über die Veranstaltung

1. Grafische Darstellungen multivariater Datensätze
2. Multivariate Verteilungen
3. Multivariate Schätz- und Testprobleme
4. Clusteranalyse
5. Diskriminanzanalyse
6. Hauptkomponentenanalyse
7. Faktorenanalyse und Strukturgleichungsmodelle
8. Assoziationsstrukturen
9. Multidimensionale Skalierung
10. Korrespondenzanalyse

Beispiele multivariater Datensätze

Beispiel 1: PISA-Studie

- Veröffentlichung 2001 durch das deutsche PISA-Konsortium
- Erfassung der qualitativen Merkmale Lesekompentenz, Mathematische Grundbildung, Naturwissenschaftliche Grundbildung

$\rightarrow$ grafische Beschreibung der Daten, Hauptkomponentenanalyse

Beispiel 1: PISA-Studie

Mittelwerte der Punkte in den Bereichen Lesekompetenz, Mathematische Grundbildung, Naturwissenschaftliche Grundbildung aus der PISA-Studie, vgl. Deutsches PISA-Konsortium (Hrsg.) (2001)

Land	Lese-kompetenz	Mathematische Grundbildung	Naturwissenschaftliche Grundbildung
Australien	528	533	528
Belgien	507	520	496
Brasilien	396	334	375
Dänemark	497	514	481
Deutschland	484	490	487
...	...	...	...

Beispiel 2: Studienanfänger Wirtschaftswissenschaften

- Test zur Mittelstufenalgebra (26 Fragen) im Wintersemester 1988/89 bei Studienanfängern der Wirtschaftswissenschaften an der FU Berlin
- Variablen:
  - Geschlecht: w/m
  - Besuch Leistungskurs Mathe: j/n
  - Abitur im Jahr 1988: j/n
  - Abinote Mathematik
  - Anzahl der im Test richtig gelösten Aufgaben

$\rightarrow$ Diskriminanzanalyse

Beispiel 2: Studienanfänger Wirtschaftswissenschaften

Ergebnisse der Studienanfänger bei dem Mathetest

Geschlecht	MatheLK	MatheNote	Abitur88	Punkte
m	n	3	n	8
m	n	4	n	7
m	n	4	n	4
m	n	4	n	2
w	n	3	n	6
...	...	...	...	...

Beispiel 3: Erstsemester Bielefeld

- Befragung von 265 Erstsemesterstudenten der Wirtschaftswissenschaften an der Uni Bielefeld im Wintersemester 1996/97
- Variablen:
  - Geschlecht: w/m
  - Gewicht
  - Alter
  - Größe
  - Rauchen: j/n
  - Besitz eines Autos: j/n
  - Geschmack von Cola: 1-5
  - Besuch Leistungskurs Mathe: j/n

$\rightarrow$ Ähnlichkeiten zwischen Studenten finden

Beispiel 3: Erstsemester Bielefeld

Ergebnis der Befragung

Geschlecht	Alter	Größe	Gewicht	Raucher	Auto	Cola	MatheLK
m	23	171	60	n	j	2	j
m	21	187	75	n	j	1	n
w	20	180	65	n	n	3	j
w	20	165	55	j	n	2	j
m	23	193	81	n	n	3	n
...	...	...	...	...	...		

Beispiel 4: Klausuren Bielefeld

- Studienanfänger Wirtschaftswissenschaften Uni Bielefeld Wintersemester 1995/96
- 17 Studierende: alle 16 Klausuren nach vier Semestern im ersten Anlauf bestanden

$\rightarrow$ hochdimensionaler Datensatz

$\rightarrow$ Ziel: Einzelnoten der Studierenden zu Gesamtnote zusammenfassen, also in niedriger Dimension darstellen

$\rightarrow$ Hauptkomponentenanalyse

Beispiel 4: Klausuren Bielefeld

Durchschnittsnoten der Studierenden

Mathematik	BWL	VWL	Methoden
1.325	1.000	1.825	1.750
2.000	1.250	2.675	1.750
3.000	3.250	3.000	2.750
1.075	2.000	1.675	1.000
3.425	2.000	2.400	2.750
...	...	...	...

Beispiel 5: Luftlinienentfernungen zwischen deutschen Städten

Luftlinienentfernungen in Kilometern

	HH	B	K	F	M
HH	0	250	361	406	614
B	250	0	475	432	503
K	361	475	0	152	456
F	406	432	152	0	305
M	614	503	456	305	0

$\rightarrow$ Mithilfe von mehrdimensionaler Skalierung aus Distanzen Konfigurationen gewinnen (z.B. Landkarte Deutschlands)

Beispiel 6: Virtual-Reality

- Möglichkeiten und Grenzen von Virtual-Reality-Technologien auf industriellen Anwendermärkten (Diplomarbeit, Bödeker & Franke (2001))
- Befragung von 508 Unternehmen, um Nutzen zu ermitteln, den Unternehmen von Virtual-Reality-System erwartet
- Bewertung folgender Merkmale auf einer Skala von 1 bis 5:
  - Veranschaulichung von Fehlfunktionen
  - Ermittlung von Kundenanforderungen
  - Qualitätsverbesserung
  - Kostenreduktion
  - Entwicklungszeitverkürzung
- Erklärung der Struktur von Korrelationen durch sogenannte Faktoren (unbeobachtete Variablen)

$\rightarrow$ Faktorenanalyse

Beispiel 6: Virtual-Reality

Korrelationen zwischen Merkmalen

	Fehler	Kunden	Angebot	Qualität	Zeit	Kosten
Fehler	1.000	0.223	0.133	0.625	0.506	0.500
Kunden	0.223	1.000	0.544	0.365	0.320	0.361
Angebot	0.133	0.544	1.000	0.248	0.179	0.288
Qualität	0.625	0.365	0.248	1.000	0.624	0.630
Zeit	0.506	0.320	0.179	0.624	1.000	0.625
Kosten	0.500	0.361	0.288	0.630	0.625	1.000

Beispiel 7: Kreditinstitute

- Einteilung von 127 Zweigstellen eines Kreditinstituts anhand verschiedener Merkmale in 9 Gruppen (Lasch & Edel (1994))
- Betrachtung einer Subgruppe von 20 Zweigstellen, die sich in 2 Gruppen einteilen lassen:
  - die ersten 14 haben hohen Marktanteil und ein überdurchschnittliches Darlehens- und Kreditgeschäft
  - die restlichen 6 sind technisch gut ausgestattet, besitzen überdurchschnittliches Einlage- und Kreditgeschäft und hohe Mitarbeiterzahl
- Ziel: Angeben einer Entscheidungsregel, mit der neue Zweigstellen einer dieser Gruppen zugeordnet werden können

$\rightarrow$ Diskriminanzanalyse

Beispiel 7: Kreditinstitute

Eigenschaften der Zweigstellen

Filiale	Einwohner	Gesamtkosten (in tausend DM)
1	1642	478.2
2	2418	247.3
3	1417	223.6
4	2761	505.6
5	3991	399.3

Beispiel 8: Regionen

- Untersuchung von 6 Regionen in Deutschland anhand bestimmter Merkmale (Brühl & Kahl, 2001, Diplomarbeit):
  - Bevölkerungszahl (in tausend Einwohner) in der Region
  - Bevölkerungszahl (in tausend Einwohner) im Oberzentrum (zentraler Ort, Hauptort)
  - Bevölkerungsdichte (in Einwohner je Quadratkilometer) im Umland
  - durchschnittliche Flugzeit zu allen 41 europäischen Agglomerationsräumen (Ballungsräume) in Minuten
  - durchschnittliche PKW-Fahrzeit zu den nächsten drei Agglomerationsräumen in Minuten
  - durchschnittliche PKW-Fahrzeit zum nächsten IC-Systemhalt des Kernnetzes in Minuten
- Ziel: Bildung von Gruppen von Regionen, sodass die Regionen in einer Gruppe ähnlich sind, während sich die Gruppen unterscheiden

$\rightarrow$ Clusteranalyse

Beispiel 8: Regionen

Merkmale der Regionen

	Bev	BevOZ	Luft	PKW	IC	BevUmland
Münster	1524.8	265.4	272	79	24	223.5
Bielefeld	1596.9	323.6	285	87	23	333.9
Duisburg/Essen	2299.7	610.3	241	45	9	632.1
Bonn	864.1	303.9	220	53	11	484.7
Rhein-Main	2669.9	645.5	202	61	15	438.6
Düsseldorf	2985.2	571.2	226	45	16	1103.9

Beispiel 9: Sternzeichen

- Untersuchung des Heiratsverhaltens im Zeichen der Astrologie
- Ermittlung der Sternzeichen von 4219 verheirateten Paaren
- Ziel: Untersuchung, ob sich die Paare anhand von Sternzeichen gefunden haben oder nicht

$\rightarrow$ Korrespondenzanalyse

Beispiel 9: Sternzeichen

Sternzeichen der Paare

	Widder	Stier	Zwill	Krebs	Löwe	...
Widder	41	34	17	41	42	...
Stier	32	35	35	31	28	...
Zwill	34	46	31	22	28	...
Krebs	29	26	25	30	18	...
Löwe	31	43	40	23	37	...
...	...	...	...	...	...	...

Beschreibung und Darstellung multivariater Datensätze

Multivariate Datenstruktur

- Betrachtet werden statistische Erhebungen bei denen n Beobachtungen von p Variablen erhoben wurden.
- Datenmatrix: $\mathbf{X} = (x_{ij})$

$$\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

- Darstellung über Beobachtungen: $\mathbf{x_{(1)}}, \dots, \mathbf{x_{(n)}}$ (Zeilen)
- Darstellung über Variablen: $\mathbf{x_1}, \dots, \mathbf{x_p}$ (Spalten)

Univariate Datenstruktur

- Ansehen einzelner Merkmale vor einer multivariaten Analyse $\rightarrow$ univariate Analyse, Deskription einzelner Variablen
- Qualitative Merkmale (z.B. Geschlecht): absolute und relative Häufigkeiten, Stabdiagramme
- Quantitative Merkmale (z.B. Alter): Histogramme, Boxplots, Kennzahlen nach Tukey (1977) [Minimum, unteres Quartil, Median, oberes Quartil, Maximum]

Multivariate Datenstruktur - Fluch der Dimension

Der Begriff Curse of Dimensionality wurde geprägt von Richard E. Bellmann (1920 – 1984):

- Wenn die Dimension des Beobachtungsraumes steigt (bei uns die Dimension p), dann steigt das Volumen des Beobachtungsraumes so schnell, dass die verfügbaren Daten den Raum möglicherweise nicht ausreichend abdecken.
- **Beispiel:** 100 Beobachtungen reichen aus, um das Einheitsintervall $[0, 1]$ mit Abstand ≤ 0.01 zwischen den Punkten abzutasten. Wie ist das in einem 10-dimensionalen Beobachtungsraum?

Grafische Elemente

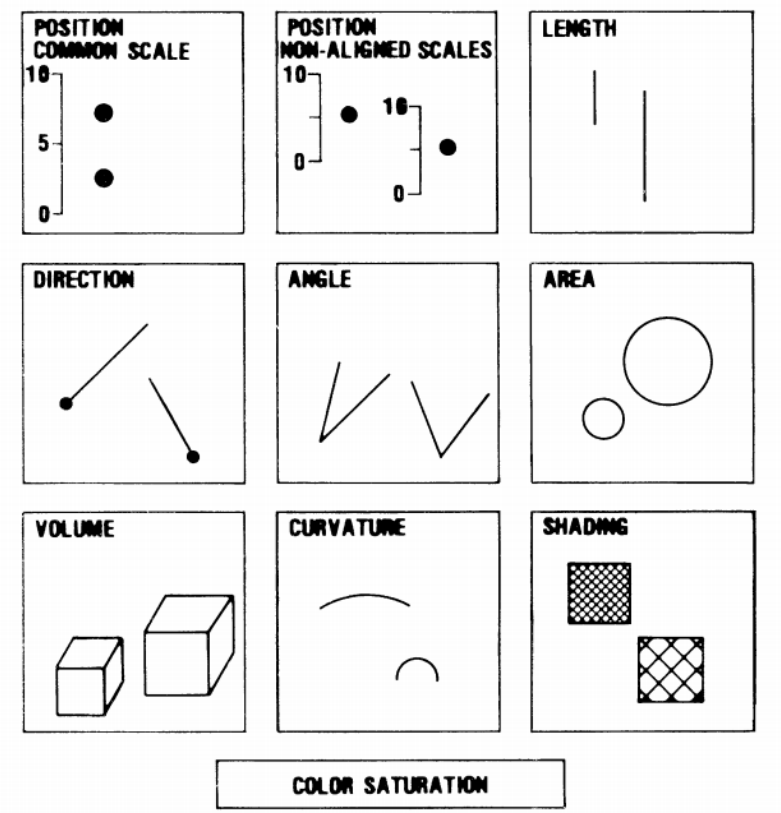


Figure 1. Elementary perceptual tasks.

Quelle: W. S. Cleveland and R. McGill (1984): Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods, *JASA (79)*, 531–554.

Grafische Elemente

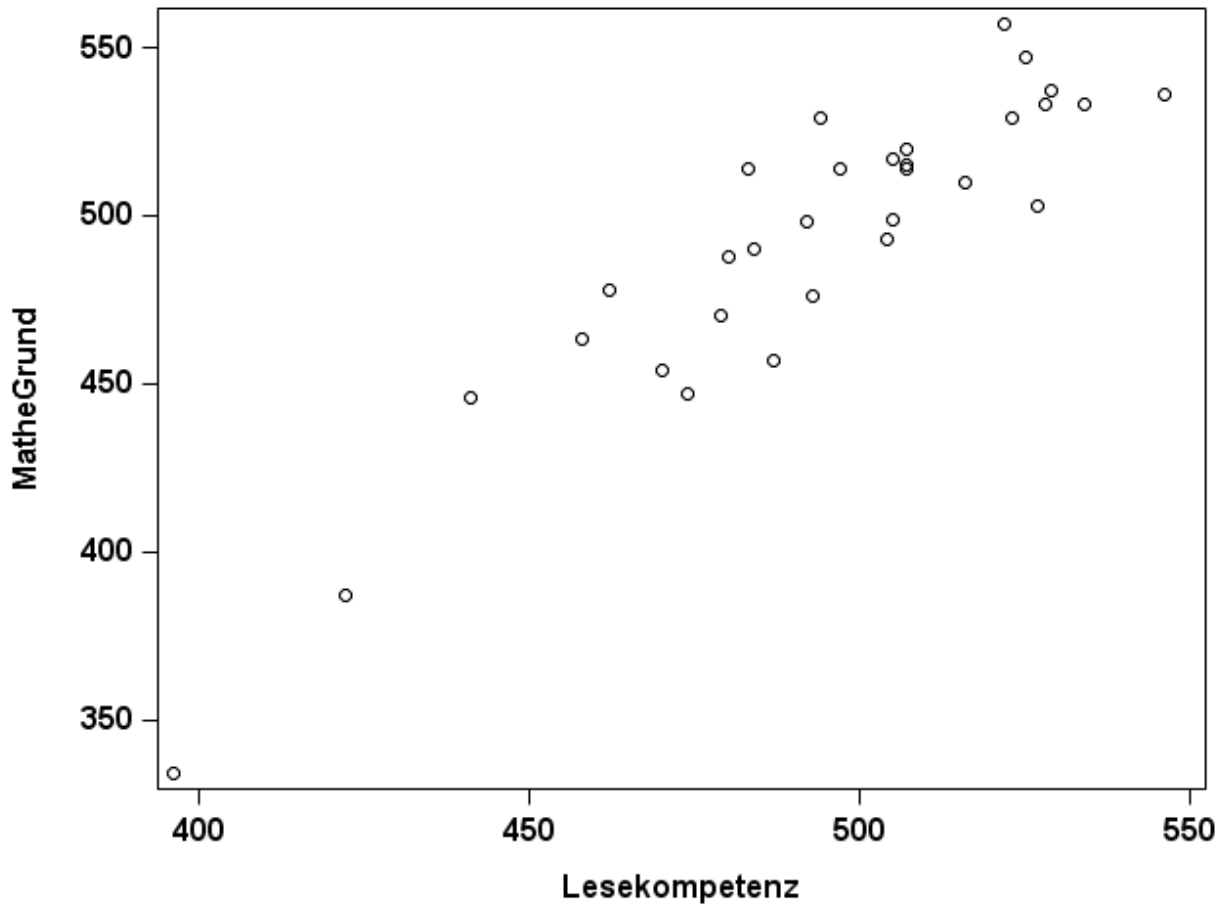
Ordnung der Elemente gemäß ihrer Präzision nach W. S. Cleveland und R. McGill:

1. Position along a common scale
2. Positions along nonaligned scales
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color saturation

$\rightarrow$ Insbesondere im Multidimensionalen ist die Wahl der Darstellung entscheidend, um die wesentliche Botschaft klar herauszuarbeiten!

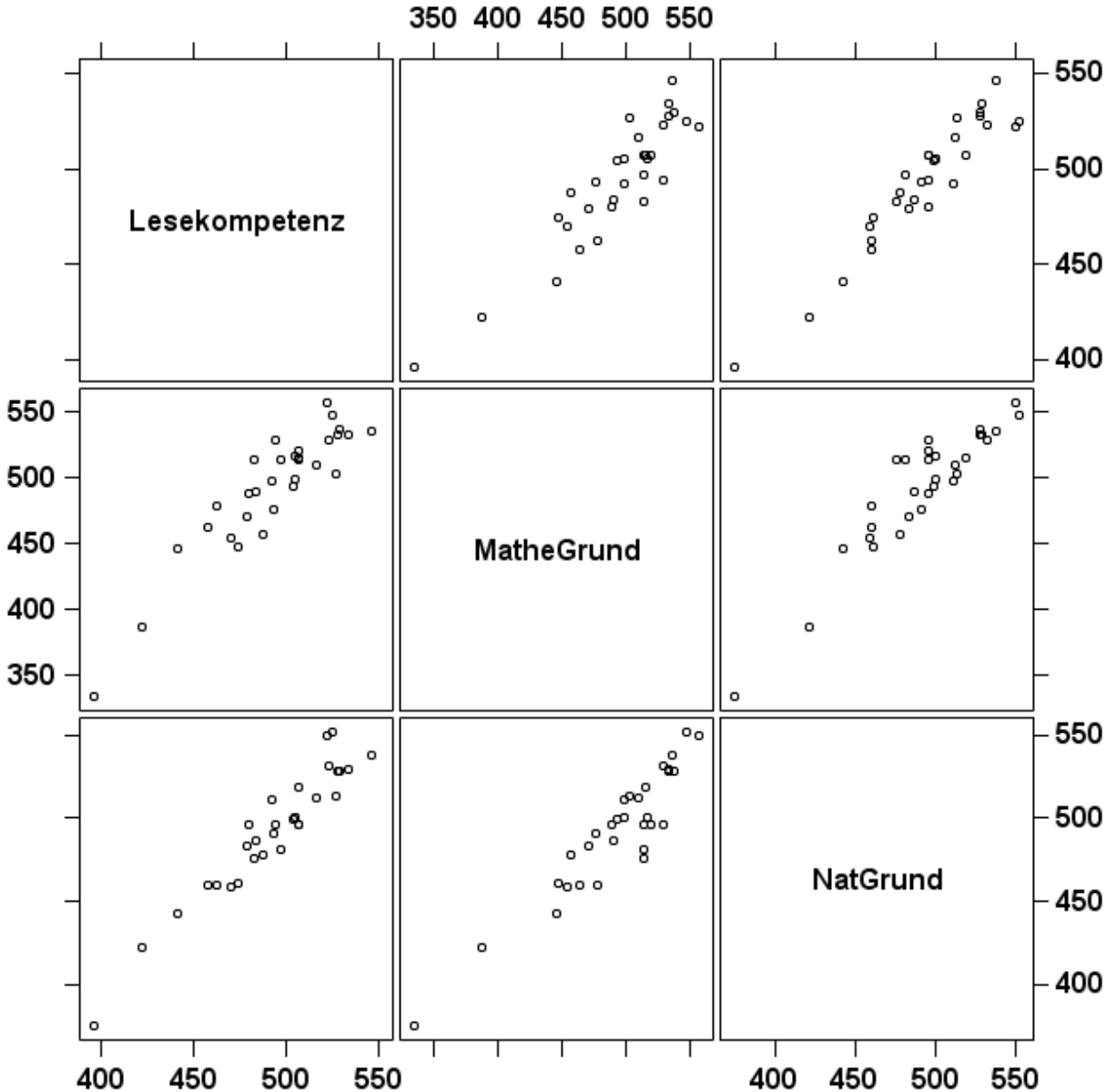
Darstellung multivariater Daten

Scatterplot



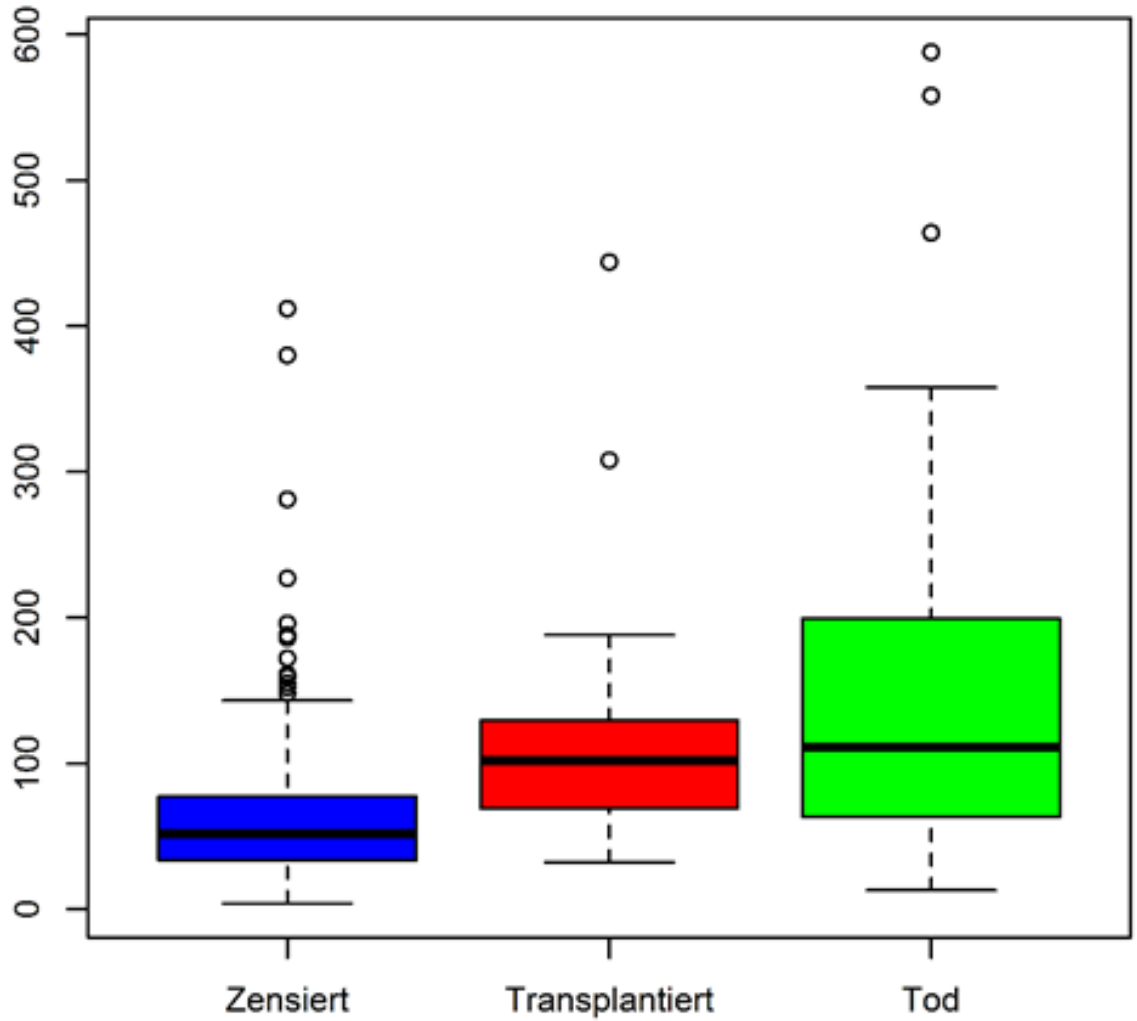
Darstellung multivariater Daten

Scatterplot-Matrix



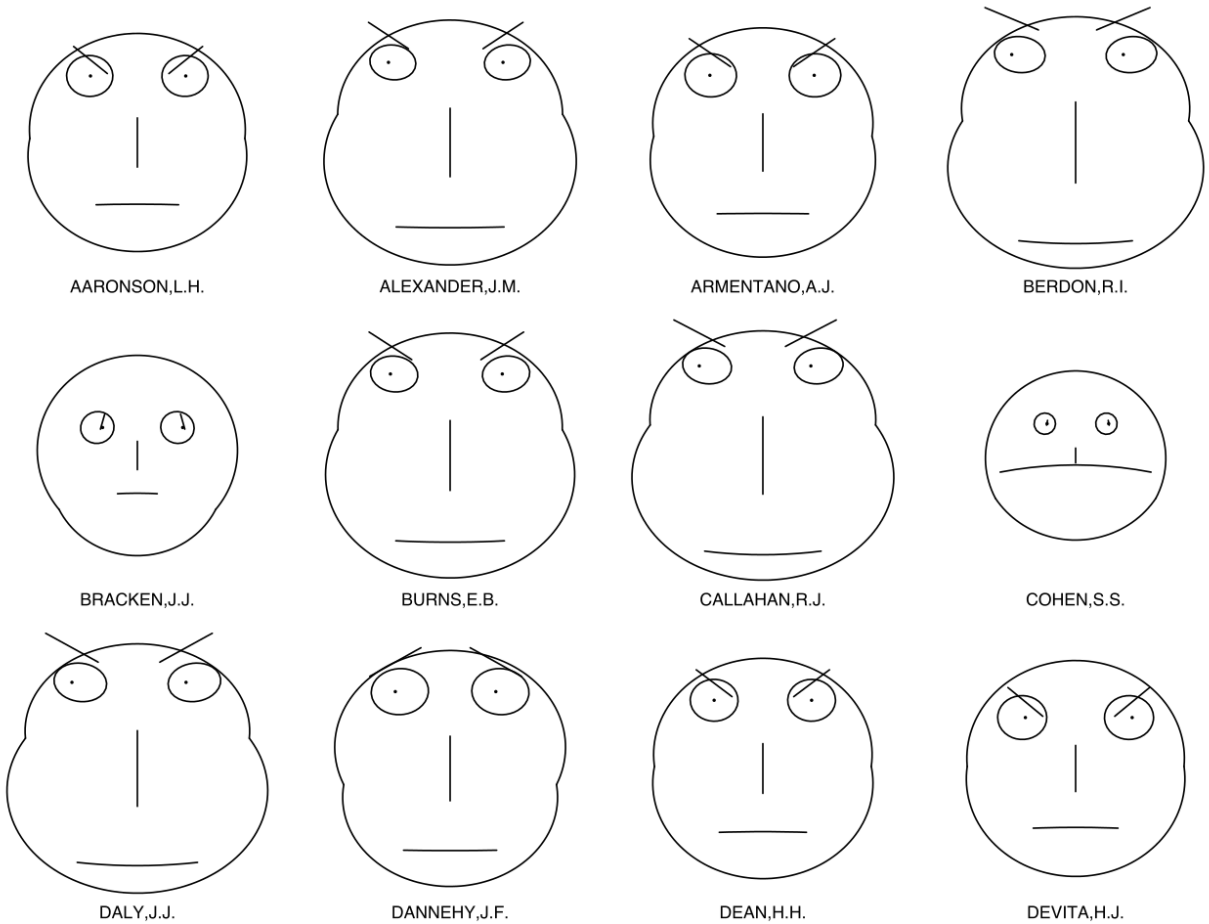
Darstellung multivariater Daten

Erweiterte Boxplots



Darstellung multivariater Daten

Chernoff-Gesichter



Quelle: https://de.wikipedia.org/wiki/Chernoff-Gesichter

Darstellung multivariater Daten - allgemein

- 3-dimensionale Grafiken
- Höherdimensionale Grafiken mit Animierung
- Farbe als weitere Dimension
- Punktgröße als weitere Dimension
- Parallele-Koordinaten-Plots
- Mosaik-Plots
- Interaktive Verbindung von Plots

Darstellung multivariater Daten - Farbwahl

Unterscheidung nach Farbräumen

1. RGB(red, green, blue)

   in R: `rgb(r, g, b, alpha, maxColorValue)`

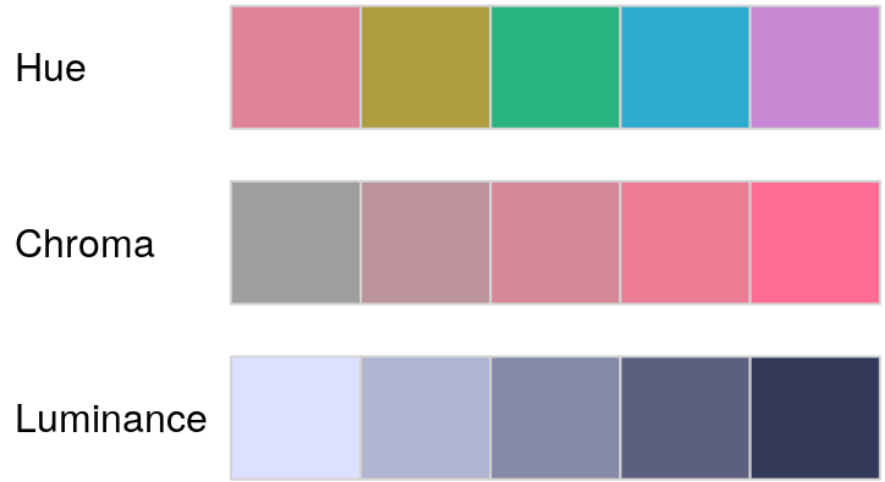
   $\rightarrow$ Intensität zwischen 0 – 1 bzw. 0 – 255

2. HCL(hue, croma, luminance)

   in R: `hcl(h, c, l, alpha)`

   $\rightarrow$ Werte zwischen 0 – 360 (0: rot, 120: grün, 240: blau, etc.)

HCL Farbraum



(Quelle: https://eeecon.uibk.ac.at/ zeileis/news/colorspace/)

$\rightarrow$ R-Paket **colorspace**: A Toolbox for Manipulating and Assessing
Colors and Palettes: http://colorspace.r-forge.r-project.org/

Möglichkeiten in R

Wichtige R-Pakete

1. **ggplot2**: Create Elegant Data Visualisations Using the Grammar of Graphics
   - Insbesondere komplexe Grafiken lassen sich deutlich einfacher erstellen als mit base R
   - Ausgangspunkt ist immer ein Datensatz (`class data.frame`)
   - Es gibt keine Methoden für bestimmte S3 (bzw. S4) Klassen
   - Erweiterung für spezielle Graphiken: **GGally**
   - Erweiterung für interaktive Graphiken: **plotly**

Wichtige R-Pakete

2. **manipulate**: Interactive Plots for RStudio
   - Dynamische Änderung von Werten in der Graphik durch Elemente wie Schieberegler (*slider*), Dropdown-Auswahl (*picker*), Checkboxen (*checkbox*) oder Knöpfe (*button*)
3. **shiny**: Web Application Framework for R
   - Ermöglicht die Erstellung von interaktiven Graphiken in Form einer HTML-Oberfläche

Weitere R-Pakete

4. **corrplot**: Visualization of a Correlation Matrix
5. **iplots**: iPlots - interactive graphics for R
6. **MASS**: Support Functions and Datasets for Venables and Ripley's MASS
   - u.a. Parallelkoordinaten-Plot mit *parcoord()*
7. **scatterplot3d**: 3D Scatter Plot
8. **vcd**: Visualizing Categorical Data
   - u.a. erweiterte Mosaik-Plots mit *mosaic()*