Unüberwachtes Lernen

Prof. Dr. Karsten Lübke

SoSe 2017

Unüberwachtes Lernen

Unüberwachtes Lernen (engl.: unsupervised learning): Es gibt *keine* bekannte abhänige Variable y, die modelliert werden soll

Methoden (u. a.):

- Hauptkomponentenanalyse (engl.: Principal Component Analysis): Finde (wenige) Linearkombinationen der Variablen: Zusammenfassung von Variablen, Dimensionsreduktion
- Clusteranalyse (engl.: Cluster Analysis): Finde Gruppen (Cluster) von Beobachtungen, die innerhalb der Cluster homogen, zwischen den Clustern heterogen sind[1]

[1] Clustern von Variablen analog

Hauptkomponentenanalyse

Idee: Fasse korrelierte Variablen (linear) zusammen. Die resultierenden Komponenten sind unkorreliert und beinhalten einen möglichst großen Anteil der (multivariaten) Gesamtvariation.

Analyse Extraversion

Extraversionstest nach Dr. Satow, angepasst von Prof. Dr. Sebastian Sauer.

Fragebogen: `http://bit.ly/1HBhKWU`

Analyse Extraversionsdaten: Alternative Einlesen

Menü RStudio:

`File -> Import Dataset-> From Excel ...`

Datei “Extraversion.xlsx” auswählen (`Browse`) `-> Import`

Analyse Extraversionsdaten: Einlesen

```r
# Gggfs. Einmalig vorab installieren
# install.packages("readxl")

# Paket zum Einlesen von Excel Dateien laden
library(readxl)

# Daten einlesen

# Daten "Extraversion.xls" einlesen
# und als Datensatz "Extraversion" in R speichern
# Achtung: Pfad zur Datei anpassen
Extraversion <- read_excel("Extraversion.xlsx")
```

Vorbereitung: Big-5 Extraversionsitems

```
library(mosaic)

Big5_Extra <- Extraversion[,1:10] %>% # Variablen wählen
  na.omit() %>% # Fehlende Werte löschen
  scale() %>% # Skalieren
  data.frame() # Als Datensatz definieren
```

Übung 1: Korrelation/ Kovarianz Big-5

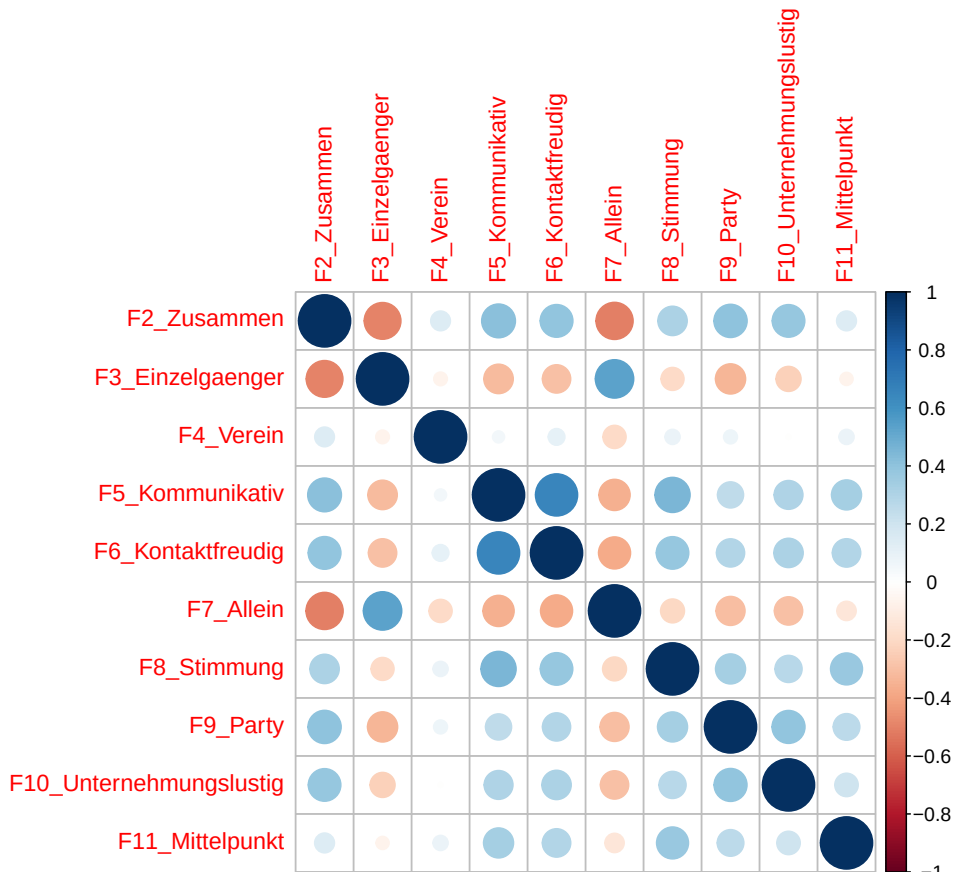
Stimmt die Aussage: Bei skalierten/ standardisierten[2] Variablen ist die Kovarianzmatrix identisch zu der Korrelationsmatrix?

- Ja.
- Nein.

[2] d. h. $\bar{x} = 0, sd = 1$

Big-5 Extraversion: Korrelation

```
library(corrplot) # ggfs. einmalig vorab installieren
cor(Big5_Extra) %>%
  corrplot()
```



Übung 2: Korrelation Big-5

Welche der folgenden Aussagen stimmt?

- A: Die Variable `F4_Verein` korreliert hoch mit allen anderen Variablen
- B: Es gibt eine hohe Korrelation zwischen `F5_Kommunikativ` und `F9_Party`
- C: Die Variablen `F3_Einzelgaenger` und `F7_Allein` korrelieren negativ mit den anderen Variablen
- D: Es gibt eine negative Korrelation zwischen `F5_Kommunikativ` und `F6_Kontaktfreudig`

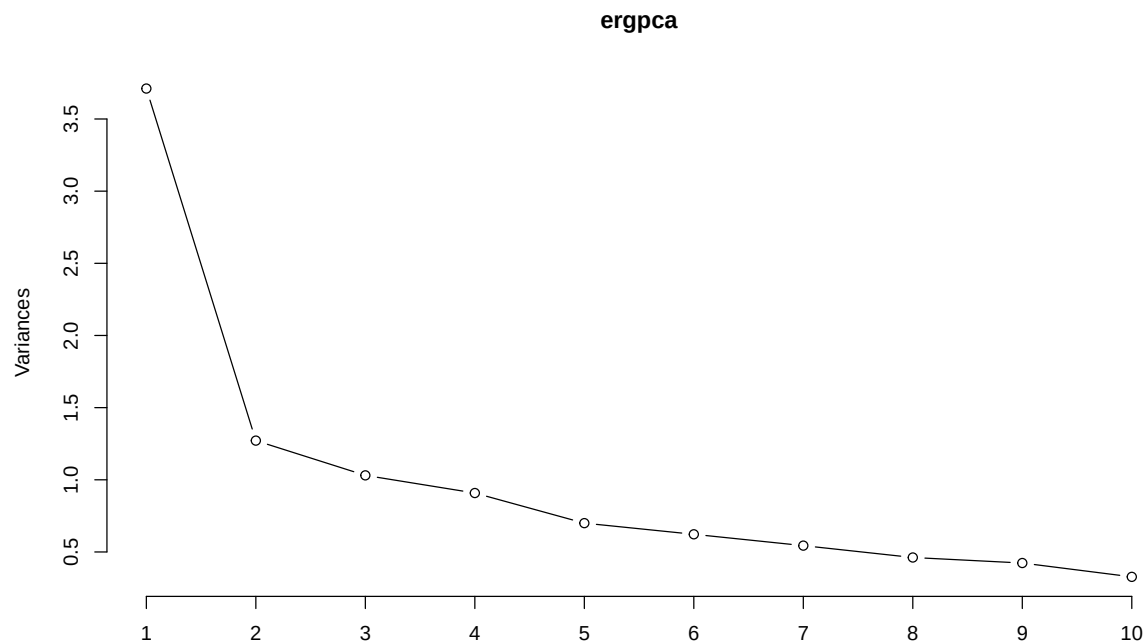
Hauptkomponentenanalyse Big-5

```
ergpca <- prcomp( ~., data=Big5_Extra)
```

Screeplot Big-5

Ein Screeplot stellt die Varianz der Hauptkomponenten dar.

```
plot(ergpca, type="l")
```



Übung 3: Screeplot

Welche der folgenden Aussagen stimmt?

- A: Alle Hauptkomponenten haben in etwa die gleiche Varianz
- B: Ungefähr nach $k = 5$ Hauptkomponenten gibt es einen Abfall in der Varianz
- C: Die erste Hauptkomponente hat eine deutlich höhere Varianz als die anderen

Big-5: Multivariate Varianz

```
summary(ergpca)
```

```
## Importance of components%s:
##                                PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.9264 1.1276 1.0152 0.95305 0.83645
## Proportion of Variance 0.3711 0.1272 0.1031 0.09083 0.06996
## Cumulative Proportion  0.3711 0.4982 0.6013 0.69214 0.76211
##                            PC7     PC8     PC9    PC10
## Standard deviation     0.73753 0.67945 0.65066 0.57230
## Proportion of Variance 0.05439 0.04616 0.04234 0.03275
## Cumulative Proportion  0.87875 0.92491 0.96725 1.00000
```

Übung 4: Multivariate Varianz

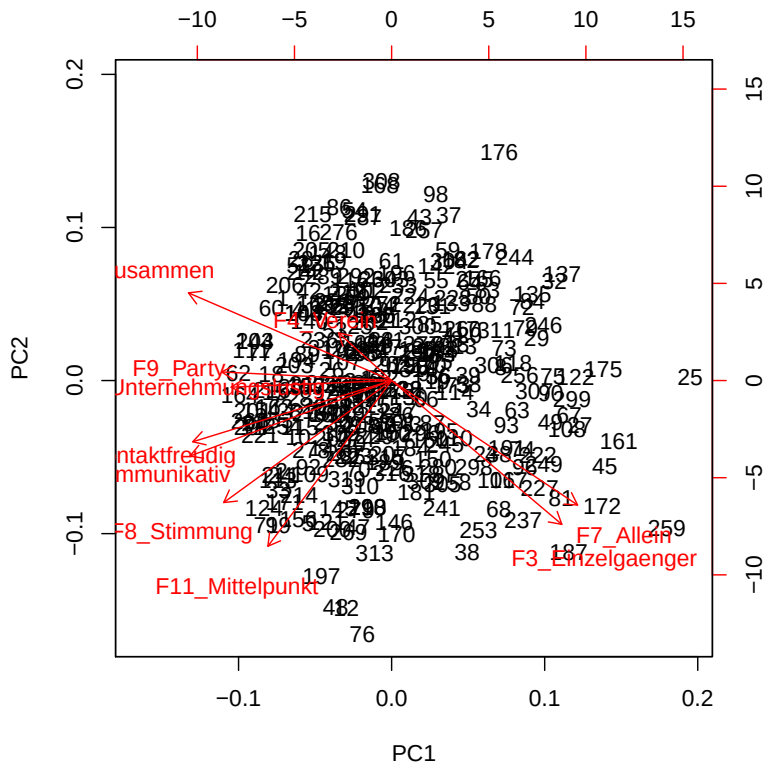
Welche der folgenden Aussagen stimmt?

- A: Die erste Hauptkomponente enthält mehr als die Hälfte der Gesamtvarianz
- B: Die ersten drei Hauptkomponenten haben eine Varianz größer als 1
- C: Die Varianz der Hauptkomponenten nimmt zu
- D: Mit 20% der Hauptkomponenten kann 80% der Gesamtvarianz erfasst werden

Extraversion: Biplot

Ein Biplot visualisiert die Ladungen der Variablen sowie die Werte der Beobachtungen auf den ersten Hauptkomponenten ("Scores").

```
biplot(ergpca)
```



Übung 5: Biplot

Welche der folgenden Aussagen stimmt?

- A: Beobachtung 25 steht gerne im Mittelpunkt
- B: Beobachtung 25 ist gerne allein
- C: Weiß nicht

Ladungen

Die Ladungen geben das Gewicht der einzelnen Variablen für die jeweilige Hauptkomponente an.

```
ergpca$rotation[,1:3]
```

```
##                                PC1         PC2         PC3
## F2_Zusammen          -0.37991170  0.28000997 -0.04890728
## F3_Einzelgaenger      0.31830517 -0.45934930  0.08135633
## F4_Verein            -0.09927139  0.15189388  0.86798448
## F5_Kommunikativ      -0.37696270 -0.24099774  0.04304854
## F6_Kontaktfreudig    -0.37188345 -0.19554849  0.08241711
## F7_Allein             0.34772008 -0.39761602 -0.11608462
## F8_Stimmung          -0.31393570 -0.38962475  0.05636911
## F9_Party             -0.31788786  0.02749680 -0.21533871
## F10_Unternehmungslustig -0.30009253 -0.02114792 -0.37783403
## F11_Mittelpunkt      -0.23135004 -0.52923767  0.15219715
```

Übung 6: Ladungen

Welche der folgenden Aussagen stimmt *nicht*?

- A: Die Variable `F4_Verein` ist für die erste Hauptkomponente unwichtig
- B: Die Variablen `F3_Einzelgaenger` und `F7_Allein` haben auf der ersten Hauptkomponente eine andere Richtung als die anderen Variablen
- C: Mit Ausnahme von `F4_Verein` sind die meisten Variablen für die erste Hauptkomponente annähernd gleich wichtig
- D: Je höher der Wert für `F11_Mittelpunkt` desto höher ist der Wert auf der zweiten Hauptkomponente

Cronbachs Alpha

Cronbachs Alpha ist eine Maßzahl für die interne Konsistenz einer Skala (Reliabilität). Sollte i. d. R. > 0.7 sein.

```
library(psych) # ggfs. einmalig vorab installieren
ca <- alpha(Big5_Extra, check.keys = TRUE)
```

```
## Warning in alpha(Big5_Extra, check.keys = TRUE): Some items
##  This is indicated by a negative sign for the variable name
```

```
summary(ca)
```

```
##
## Reliability analysis
##  raw_alpha std.alpha G6(smc) average_r S/N   ase  mean   sd
##        0.8       0.8    0.81      0.28 3.9 0.017 -0.15 0.59
```

Übung 7: Cronbachs Alpha

Ist die interne Konsistenz der Skala hier akzeptabel?

- Ja.
- Nein.

Offene Übung Freizeitverhalten

Lassen sich die Variablen `F12_Facebook`, `F13_Kater` und `F19_Partybesuche` zusammenfassen?

Finde Gruppen, die sich intern ähnlich sind:

- Agglomerativ/ hierarchisch: Beobachtungen werden sukzessiv zusammengefasst
- Partitionierend: Beobachtungen werden zu k Clustern zusammengefasst

Ähnlichkeit/ Unähnlichkeit wird über Distanzmaße (z. B. Euklidischer Abstand $d(x, y) = \sqrt{\sum_j (x_j - y_j)^2}$ definiert.

Vorbereitung: Freizeitverhalten

```
Freizeit <- Extraversion %>%
  select(F12_Facebook, F13_Kater, F19_Partybesuche) %>% # Var
  na.omit() %>% # Fehlende Werte löschen
  scale() %>% # Skalieren
  data.frame() # Als Datensatz definieren
```

Distanz

```
Freizeit[1:3,] # Ersten drei Beobachtungen
```

```
##   F12_Facebook   F13_Kater F19_Partybesuche
## 1    0.4090813 -0.24245120         0.8852787
## 2    0.0531947  0.91730552         0.8852787
## 3    0.6202668 -0.01049986         1.2862530
```

```
Freizeit[1:3,] %>%
  dist() # Distanz
```

```
##          1         2
## 2 1.2131327
## 3 0.5090984 1.1589539
```

Übung 8: Distanz

Welche der folgenden Aussagen stimmt?

- A: Beobachtungen 1 und 2 sind sich am ähnlichsten
- B: Beobachtungen 1 und 3 sind sich am ähnlichsten
- C: Beobachtungen 2 und 3 sind sich am ähnlichsten

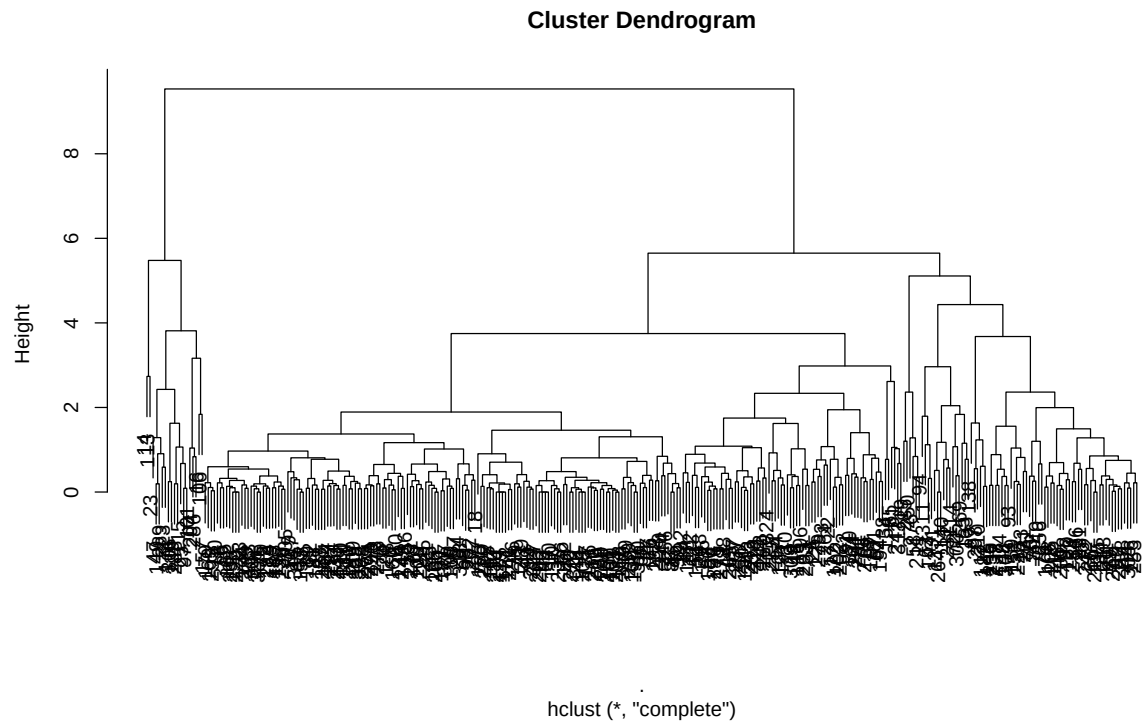
Hierarchische Clusteranalyse

```
erghclust <-Freizeit %>% # Datensatz
  dist() %>% # Distanz
  hclust() # Hierarchische Cluster
```

Dendrogramm

Je höher (`Height`) die Stelle ist, an der zwei Beobachtungen oder Cluster zusammengefasst werden, desto größer ist die Distanz zwischen ihnen.

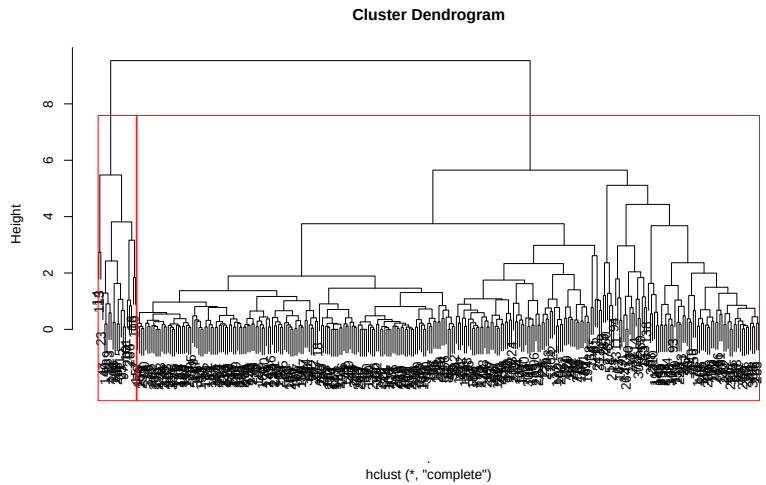
```
plot(erghclust)
```



Anzahl Cluster

Eine mögliche Trennung für $k = 2$ Cluster:

```
plot(erghclust)
rect.hclust(erghclust, k=2, border="red")
```




```
# Zuordnung in Datensatz schreiben
Freizeit$hcclust <- cutree(erghclust, k=2)
```

Cluster Beschreibung

```
mean(F12_Facebook ~  hcclust, data=Freizeit)
```

```
##           1           2
## -0.01216125  0.19522002
```

```
mean(F13_Kater ~  hcclust, data=Freizeit)
```

```
##          1          2
## -0.1925119  3.0903234
```

```
mean(F19_Partybesuche ~  hcclust, data=Freizeit)
```

```
##          1          2
## -0.1526532  2.4504855
```

Übung 9: Cluster Beschreibung

Welche der folgenden Aussagen stimmt?[3]

- A: Im Mittelwert haben Beobachtungen aus Cluster 2 0.2 Facebook-Freunde
- B: Im Mittelwert gehen Personen aus Cluster 1 öfter auf Partys als aus Cluster 2
- C: Personen aus Cluster 2 haben im Mittelwert öfter einen Kater als aus Cluster 1

[3] Beachte: der Datensatz Freizeit ist standardisiert

k-Means Clusteranalyse

Der Ablauf des Verfahrens:

- 1: Zufällige Beobachtungen als k Clusterzentrum
- 2: Zuordnung der Beobachtungen zum nächsten Clusterzentrum
- 3: Neuberechnung der Clusterzentren als Mittelwert der dem Cluster zugeordneten Beobachtungen

Wiederholung bis keine Änderung in (2) oder maximale Iterationsanzahl erreicht.

k-Means Clusteranalyse Freizeitverhalten

Mit z. B. $k = 3$ Zentren:

```
Freizeit <- Freizeit %>%
  select(-hcclust) # Clusterzuordnung löschen

set.seed(1896) # Zufallszahlengenerator setzen

ergkclust <- kmeans(Freizeit, # Datensatz
                    centers = 3, # Anzahl Zentren (k)
                    nstart = 10) # Anzahl Startpartitionen
```

Ergebnis k-Means

```
ergkclust$size # Anzahl Beob. je Cluster
```

```
## [1]  95  25 204
```

```
ergkclust$centers # Cluster Zentren
```

```
##   F12_Facebook  F13_Kater F19_Partybesuche
## 1    1.0836399  0.1302987         0.3527567
## 2    0.1553459  2.6925065         2.2485914
## 3   -0.5236737 -0.3906423        -0.4398366
```

Übung 10: Clustergröße

Welche Aussage stimmt?

- A: Die Anzahl Beobachtungen ist in Cluster 1 am größten
- B: Die Anzahl Beobachtungen ist in Cluster 2 am größten
- C: Die Anzahl Beobachtungen ist in Cluster 3 am größten

Übung 11: Beschreibung Cluster

Für welchen Cluster passt die Beschreibung *Facebook Junkies* am Besten?

- A: Cluster 1
- B: Cluster 2
- C: Cluster 3

Übung 12: Clusterzentren

In welcher Variable unterscheidet sich Cluster 3 am meisten von den anderen?[4]

- A: `F12_Facebook`
- B: `F13_Kater`
- C: `F19_Partybesuche`

[4] Beachte: der Datensatz Freizeit ist standardisiert

Offene Übung Extraversion

Führen Sie eine Clusteranalyse auf den Scores der ersten beiden Hauptkomponenten der Extraversionsitems durch

```
pcascores <- predict(ergpca)[,1:2]
```