# SBWL Tourismusanalyse und Freizeitmarketing

## Vertiefungskurs 4: Multivariate Verfahren 2

## Teil 2: Explorative multivariate Analyse & Clusteranalyse

Achim Zeileis & Thomas Rusch

Institute for Statistics and Mathematics

# Inhalt

- Einheit 8: Explorative Grafik
- Einheit 9: Hauptkomponentenanalyse
- Einheit 10: Multidimensionale Skalierung
- Einheit 11: Hierarchisches Clustern
- Einheit 12: $k$-Means

# Notation

Bei der (explorativen) multivariaten Analyse werden in aller Regel $p$ Variablen untersucht, die alle gleichberechtigt in die Analyse eingehen – wo also $nicht$ nur eine Variable die abhängige Größe ist und alle anderen Erklärungsvariablen sind.

Basis der Analyse sind dann die Beobachtungen dieser $p$ Variablen an $n$ Merkmalsträgern. Jede Beobachtung lässt sich als $p$-dimensionaler Vektor schreiben

$$x_i \quad = \quad (x_{i1}, \ldots, x_{ip})^\top.$$

# Notation

Der gesamte Datensatz läßt sich dann wie gehabt als Matrix schreiben

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Dabei kann es sein, daß $n \gg p$ (wie typischerweise in der Regressionsanalyse), aber auch $n \ll p$. Es gibt auch Fälle, wo die Rollen von $n$ und $p$ vertauschbar sind, d.h. also auch $X^\top$ anstatt $X$ betrachtet werden kann.

Wir werden zunächst davon ausgehen, daß alle Beobachtungen in $X$ quantitativ sind.

# Notation

**Beispiel:**

Für die Touristen aus dem `GSA` Datensatz betrachten wir einen aggregierten Teildatensatz: Für jede der betrachteten Sommeraktivitäten wird der Anteil der Touristen jedes Landes berechnet, die angegeben haben, diese Aktivität in ihrem Urlaub betrieben zu haben. Damit erhalten wir eine Datensatz mit $n = 15$ Ländern und verwenden $p = 8$ verschiedene Sommeraktivitäten.

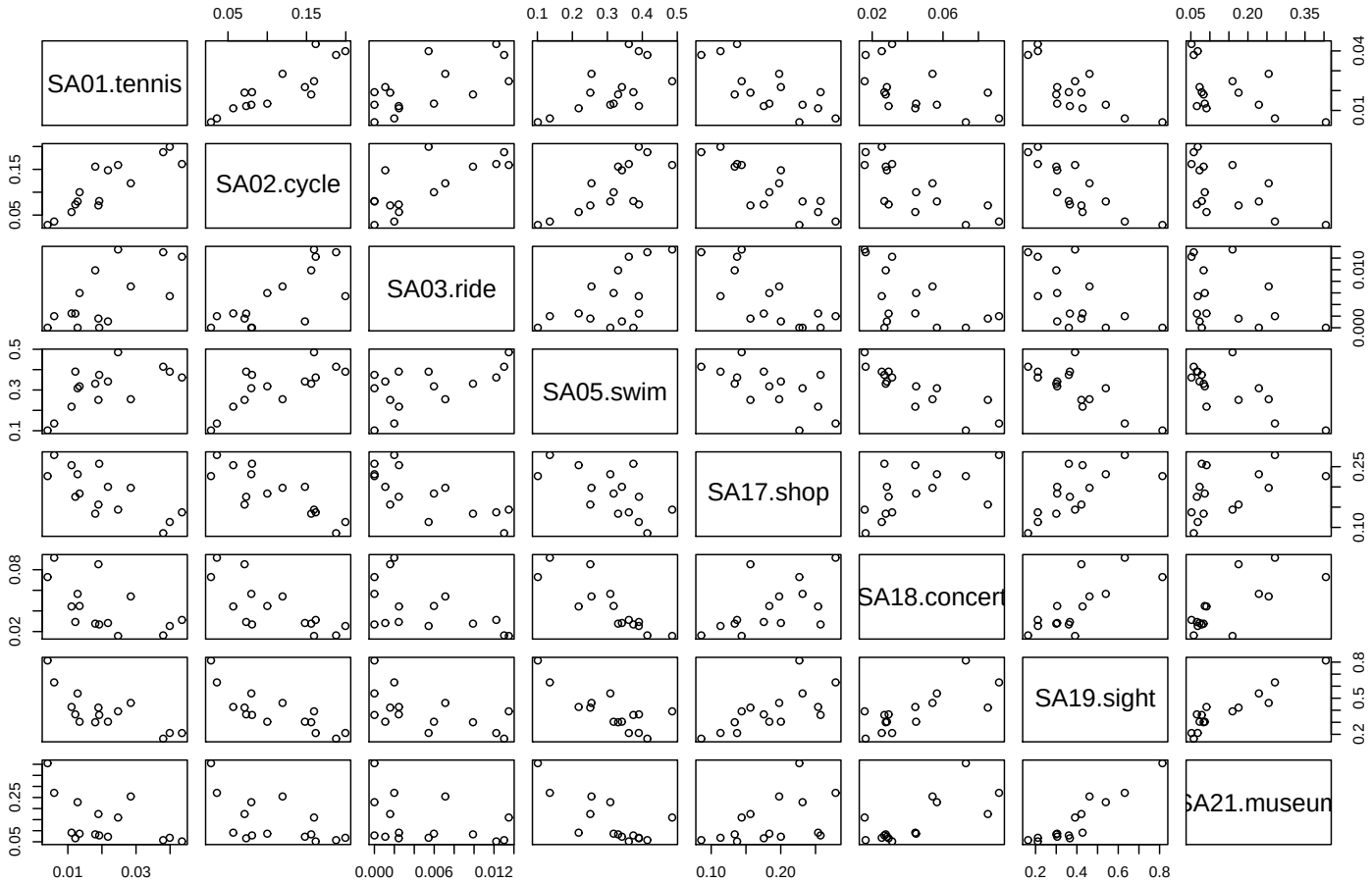
# Explorative Grafik

## Paarweise Streudiagramme

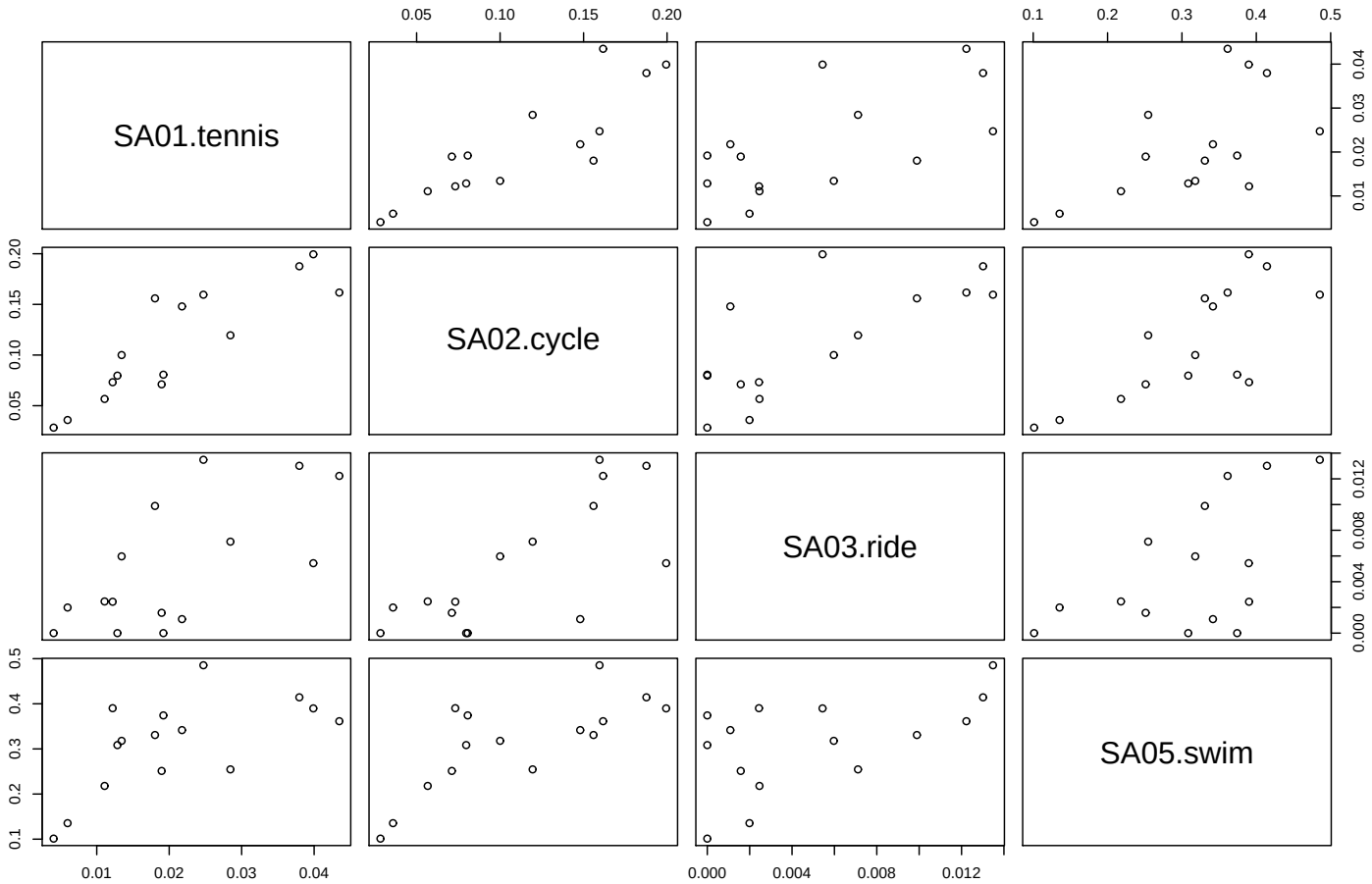
Eine sehr einfache Möglichkeit, sich einen Überblick über eine Datenmatrix zu verschaffen, sind paarweise Streudiagramme. Diese werden manchmal auch Streudiagramm-Matrix genannt.

Dabei wird einfach für jede mögliche paarweise Kombination der $p$ Variablen ein Streudiagramm (mit $n$ Beobachtungen) in ein Matrix-Schema eingetragen. Formal gesprochen heißt das, daß die Daten aus dem $\mathbb{R}^p$ in den $\mathbb{R}^2$ projiziert und dann visualisiert werden.

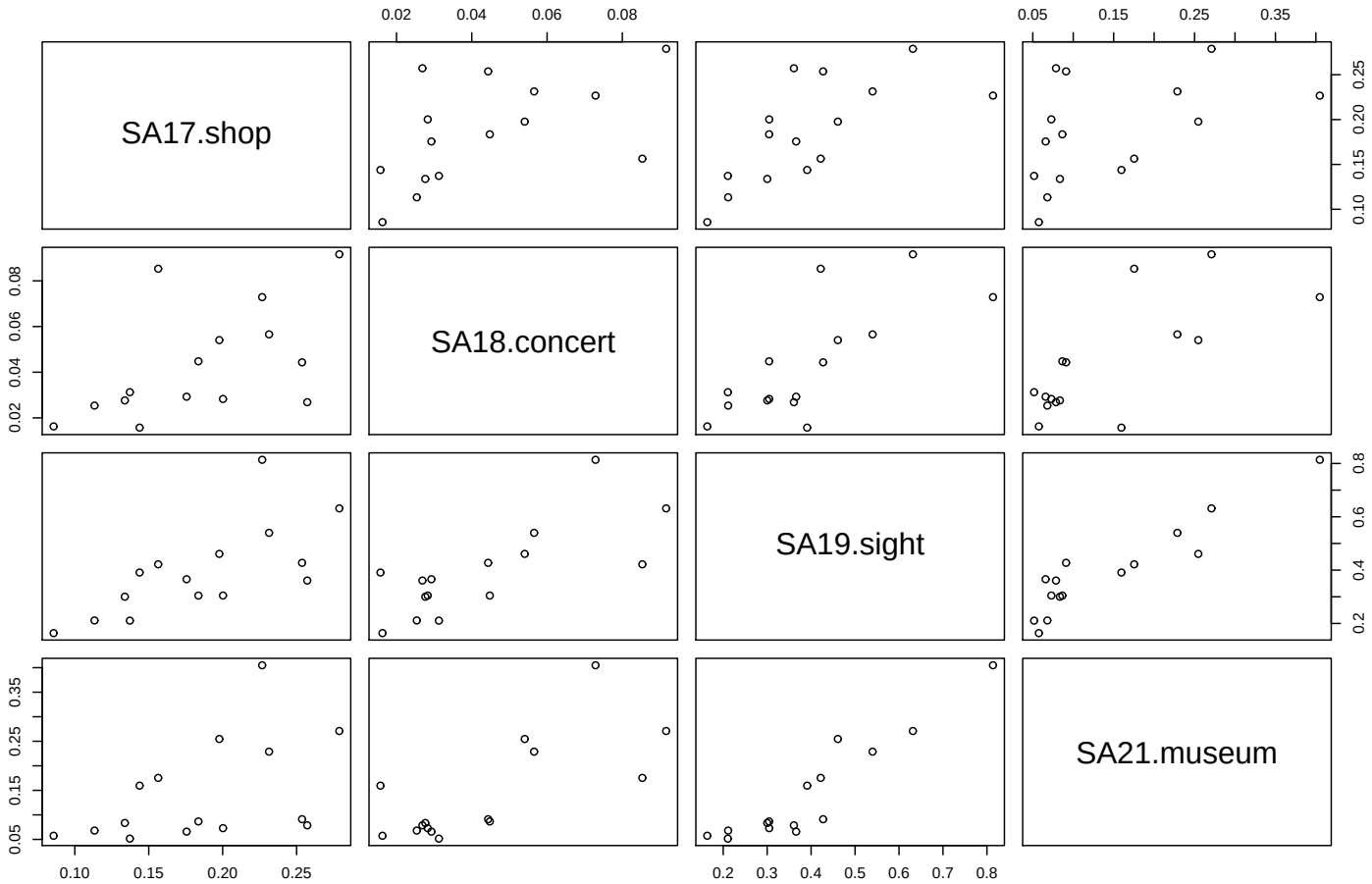
# Explorative Grafik



# Explorative Grafik



# Explorative Grafik



# Explorative Grafik

Während die Verwendung von Variablen mit unterschiedlichen Spannbreiten in paarweisen Streudiagrammen kein Problem ist, so benötigen andere Visualisierungsmethoden Beobachtungen auf einer standardisierten Skala. Dafür wird in der Regel das Einheitsintervall $[0, 1]$ verwendet.

Um eine beliebige Matrix $X$ in eine Matrix $\tilde{X}$ zu transformieren, die nur Beobachtungen aus $[0, 1]$ enthält, skaliert man üblicherweise jede Spalte so, daß das Minimum bei $0$ und das Maximum bei $1$ liegt. Die transformierten Werte $\tilde{X}$ können dann als Anteile zwischen Minimum und Maximum interpretiert werden.

# Explorative Grafik

Formal heißt das, daß für jede Spalte das Minimum $min_j$ und das Maximum $max_j$ $(j = 1, \ldots, p)$ ausgerechnet wird. Damit ist dann $\tilde{x}_{ij}$ definiert als:

$$
\begin{aligned}
min_j &= \min_{i=1,\ldots,n} x_{ij} \\
max_j &= \max_{i=1,\ldots,n} x_{ij} \\
\tilde{x}_{ij} &= \frac{x_{ij} - min_j}{max_j - min_j}
\end{aligned}
$$

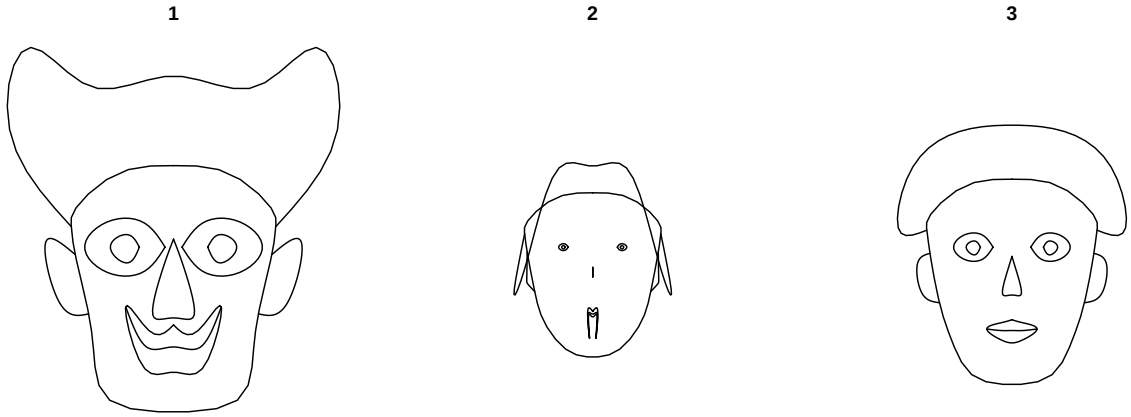
# Explorative Grafik

## Chernoff-Gesichter

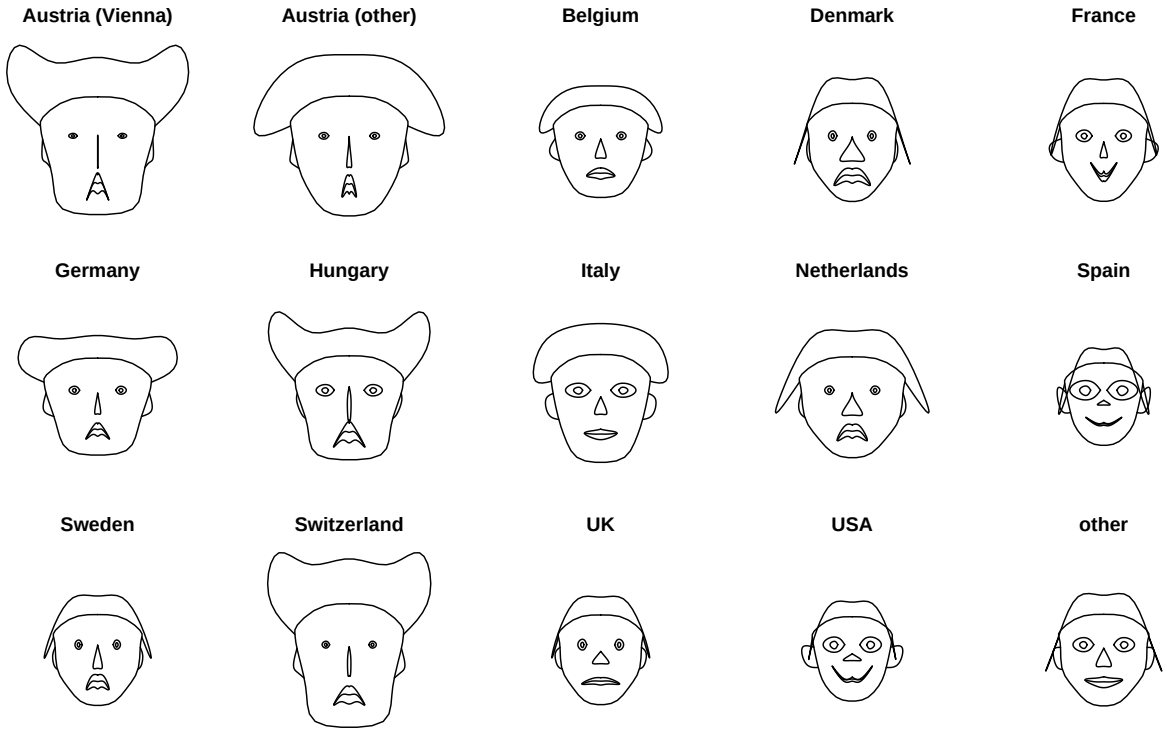
Eine eher unterhaltsame als wirklich informative Darstellung sind die Chernoff vorgeschlagenen und Flury & Riedwyl verbesserten Gesichter. Dabei werden verschiedene Attribute eines Gesichts mit Variablen belegt und gemäß $\tilde{X}$ visualisiert.

Die hier verwendete Implementierung kann bis zu 15 verschiedene Attribute variieren: 1 Höhe des Gesichts, 2 Breite des Gesichts, 3 Form des Gesichts, 4 Höhe des Munds, 5 Breite des Munds, 6 Form des Lächelns, 7 Höhe der Augen, 8 Breite der Augen, 9 Höhe der Haare, 10 Breite der Haare, 11 Styling der Haare, 12 Höhe der Nase, 13 Breite der Nase, 14 Breite der Ohren, 15 Höhe der Ohren.

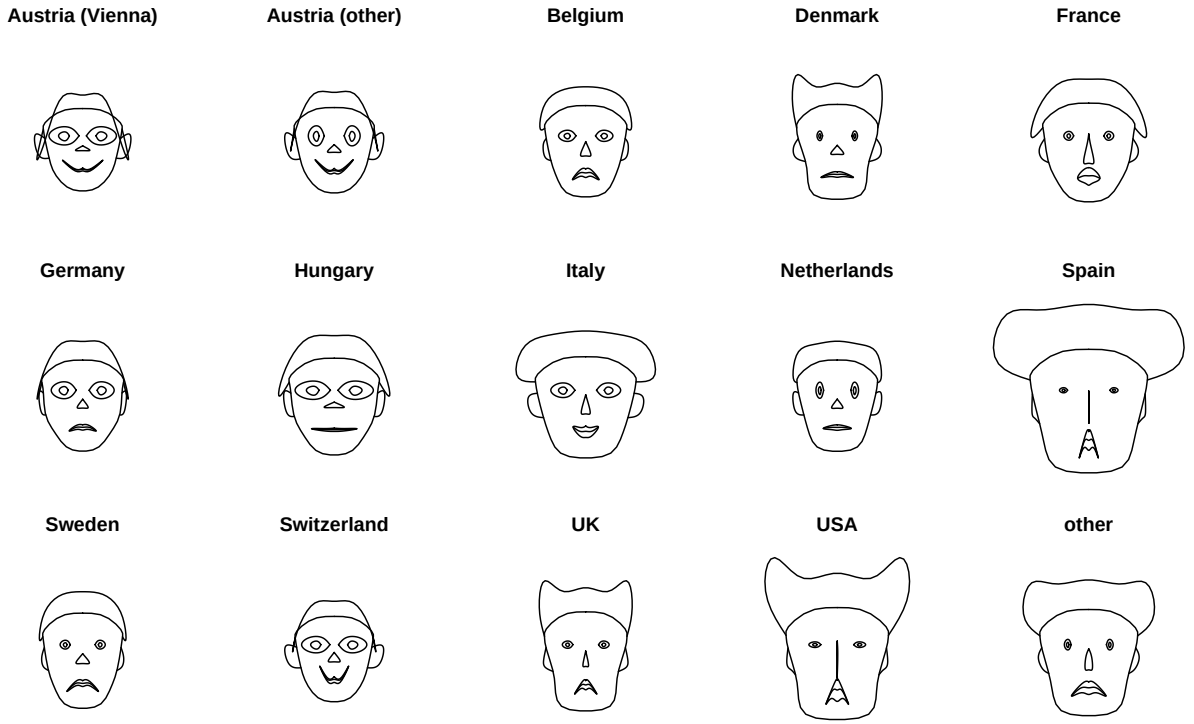
# Explorative Grafik



# Explorative Grafik



# Explorative Grafik



# Explorative Grafik

## Andrews-Kurven

Eine deutlich wissenschaftlichere wenn auch nicht immer informativere Form der Visualisierung sind die Andrews-Kurven. Hierbei wird jede der multivariaten Beobachtungen $x_i$ durch einen kompletten Funktionsverlauf visualisiert.

Die Funktion ist definiert als

$$\begin{aligned} f_{x_i}(t) &= \frac{1}{\sqrt{2}} \cdot x_{i1} + x_{i2}\sin(t) + x_{i3}\cos(t) + \\ &\quad x_{i4}\sin(2 \cdot t) + x_{i5}\cos(2 \cdot t) + \ldots + x_{ip}\sin(\lfloor \frac{p}{2} \rfloor \cdot t) \end{aligned}$$

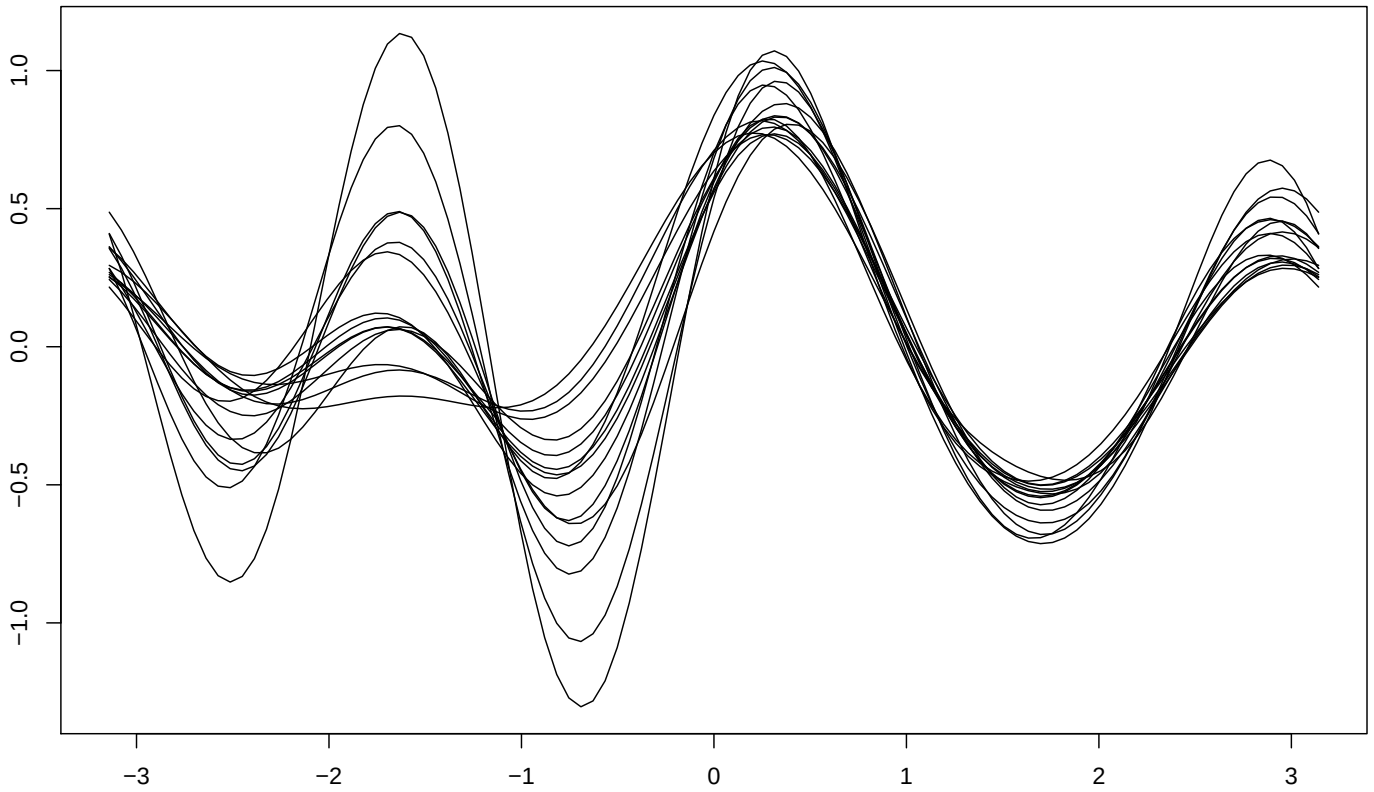
bzw. $\cos(\cdot)$ falls $p$ ungerade.

# Explorative Grafik

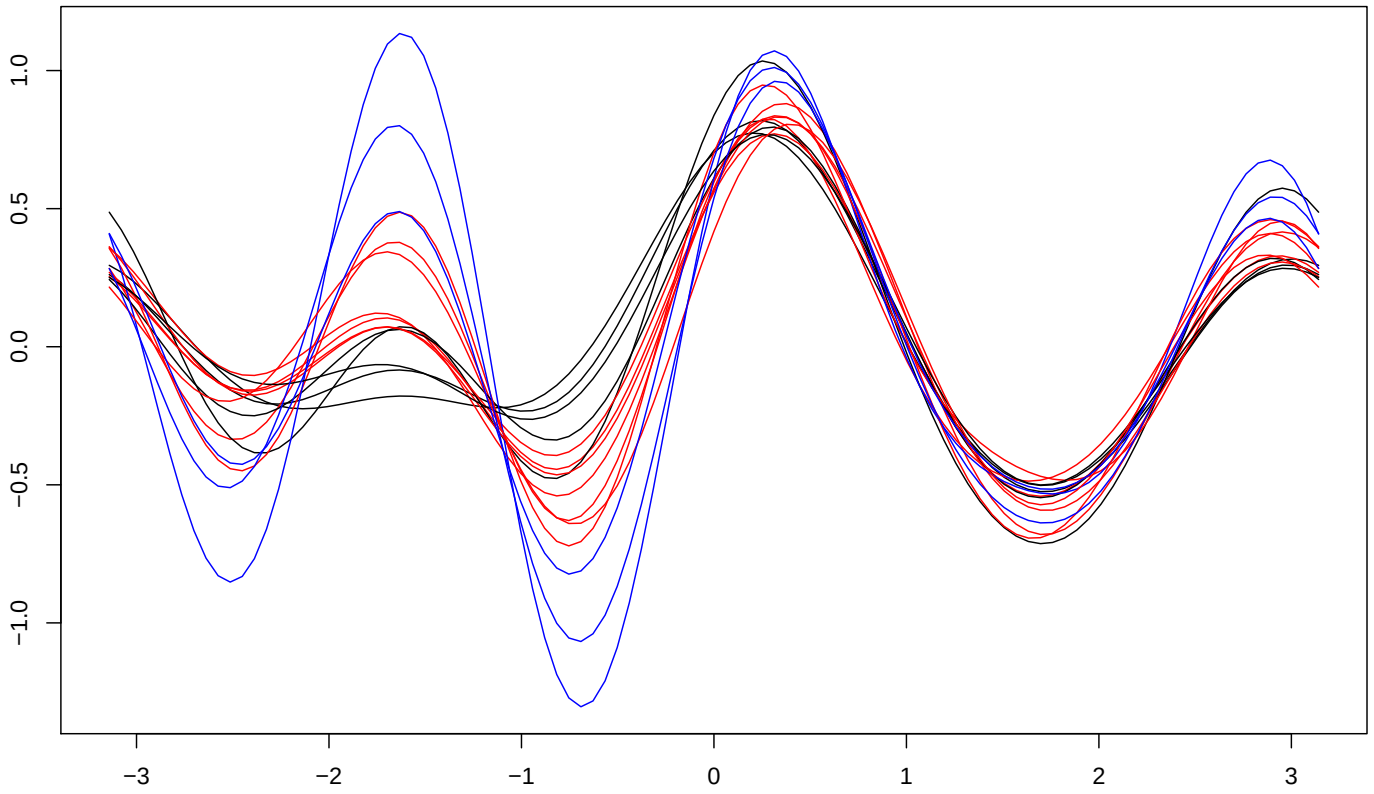
Die Funktion wird dann auf dem Intervall $-\pi < t < \pi$ abgetragen.

Diese Andrews-Kurven haben zwar die angenehme Eigenschaft, daß sie sowohl Mittelwerte als auch euklidische Distanzen beibehälten, jedoch hängt die Darstellung wieder stark von der Reihenfolge der Variablen ab und ist nicht immer sehr informativ.

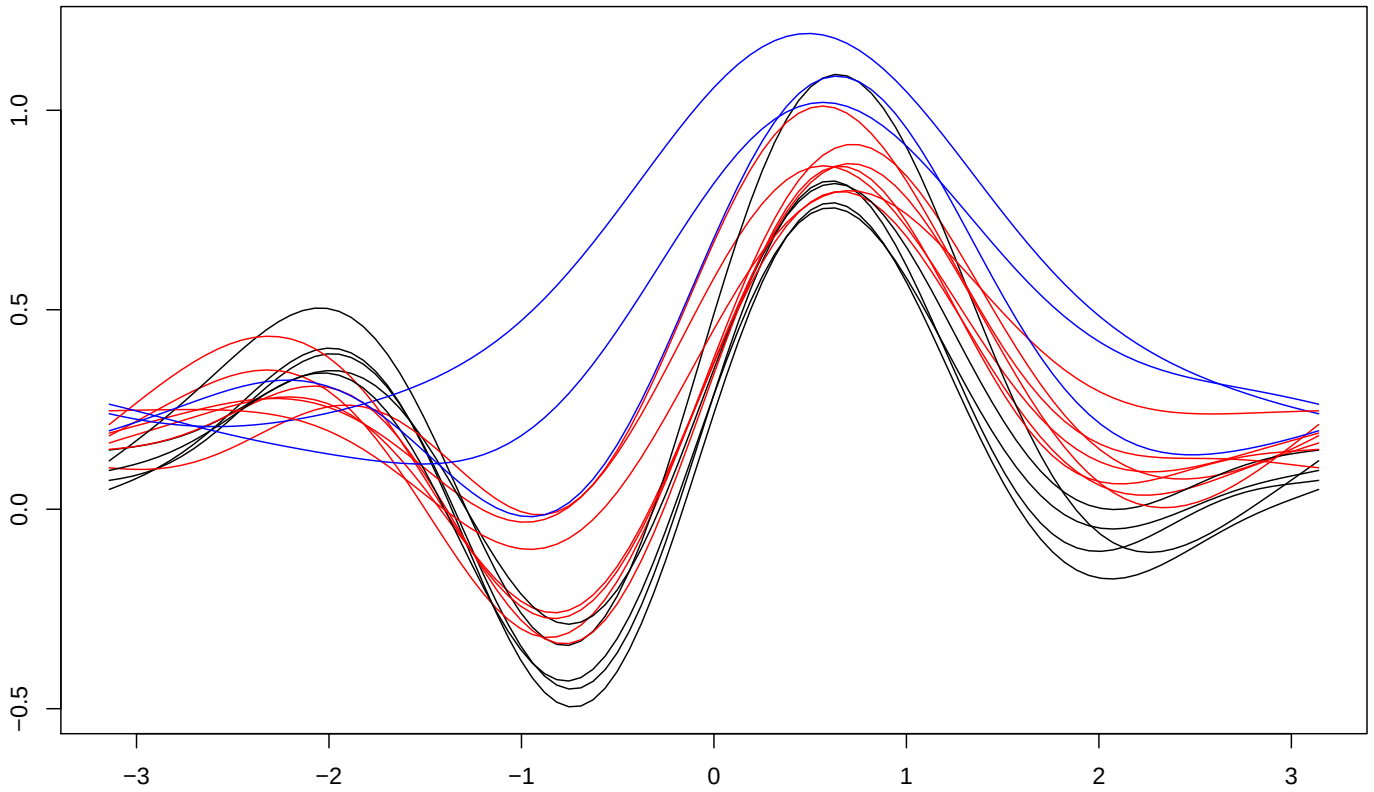
# Explorative Grafik



# Explorative Grafik



# Explorative Grafik



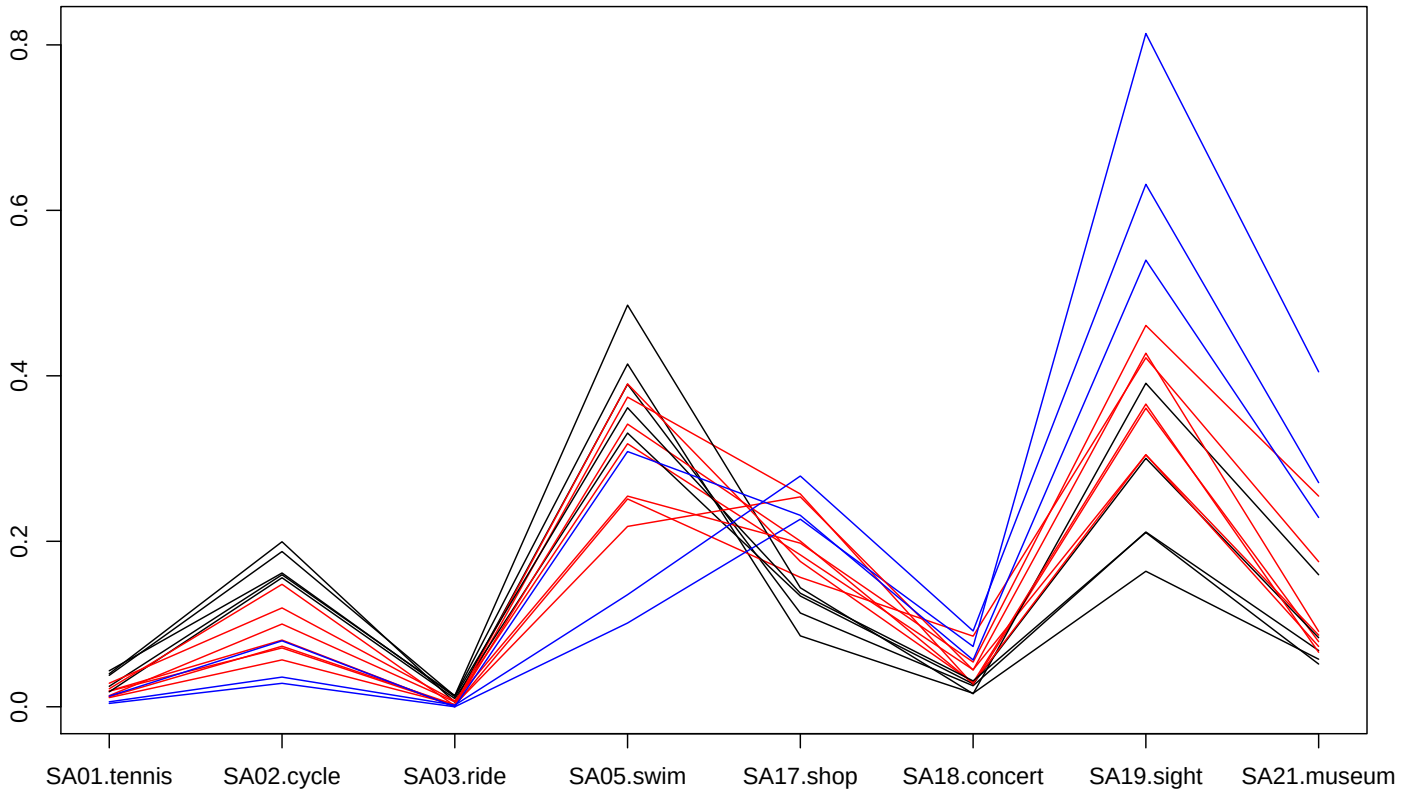
# Explorative Grafik

## Parallele Koordinaten

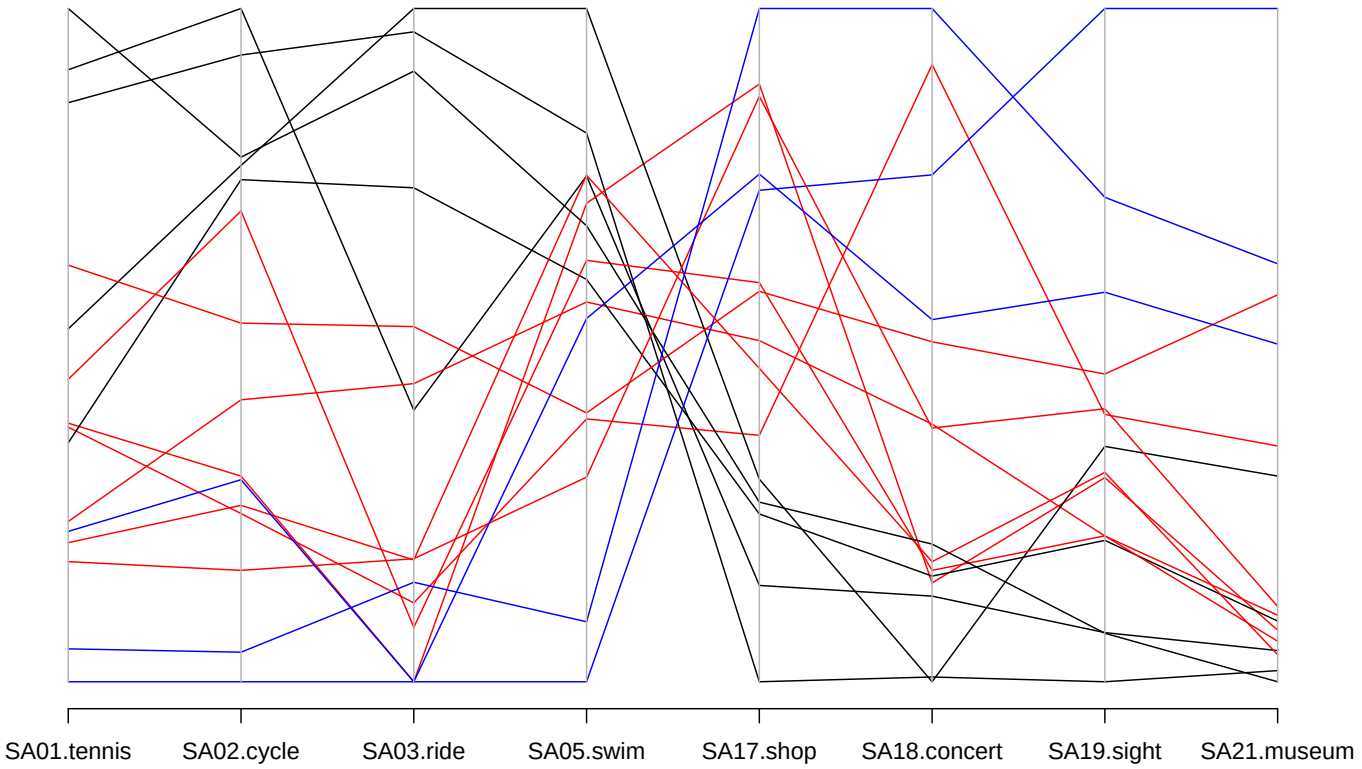
Eine sehr einfache Art der Visualisierung ist es für jede Beobachtung $\tilde{x}_i$ einen Polygonzug zu zeichnen, diese nennt man auch parallele Koordinaten.

Manchmal werden auch die Originalbeobachtungen $x_i$ visualisiert.

# Explorative Grafik



# Explorative Grafik



# Explorative Grafik

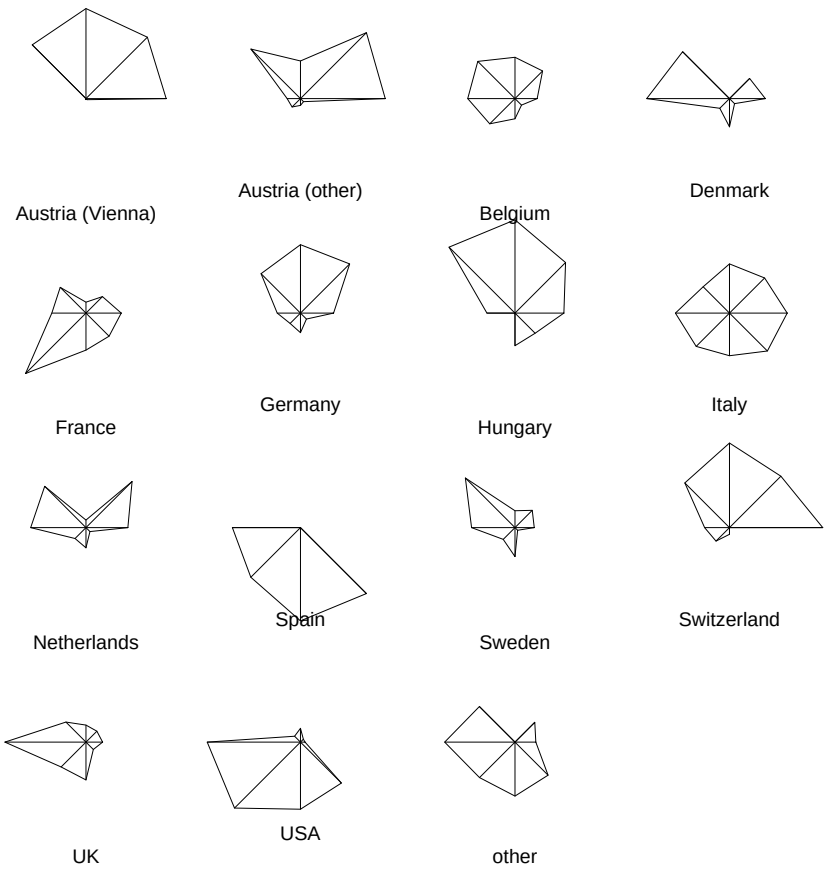
## Sterne

Wenn die Anzahl der Beobachtungen $n$ nicht zu groß ist, sind Sterne ein sehr gutes Mittel der Visualisierung von multivariaten Daten $\tilde{X}$ (oder ggf. auch $X$).

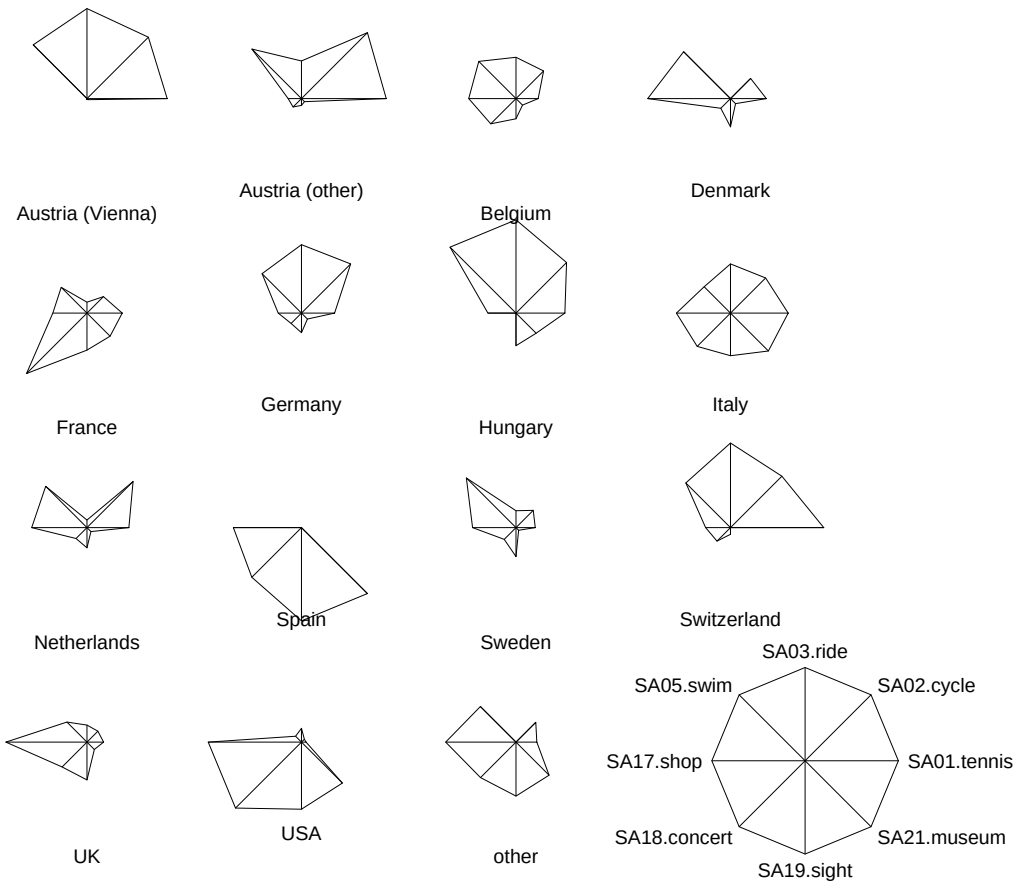
Dabei wird ein Kreis in $p$ gleich große Sektoren eingeteilt und jeder Wert $\tilde{x}_{ij}$ wird in einem der Sektoren abgetragen.

Verschiedene Varianten dieser Darstellung werden auch Sonnen, Glyphen, o.ä., genannt.

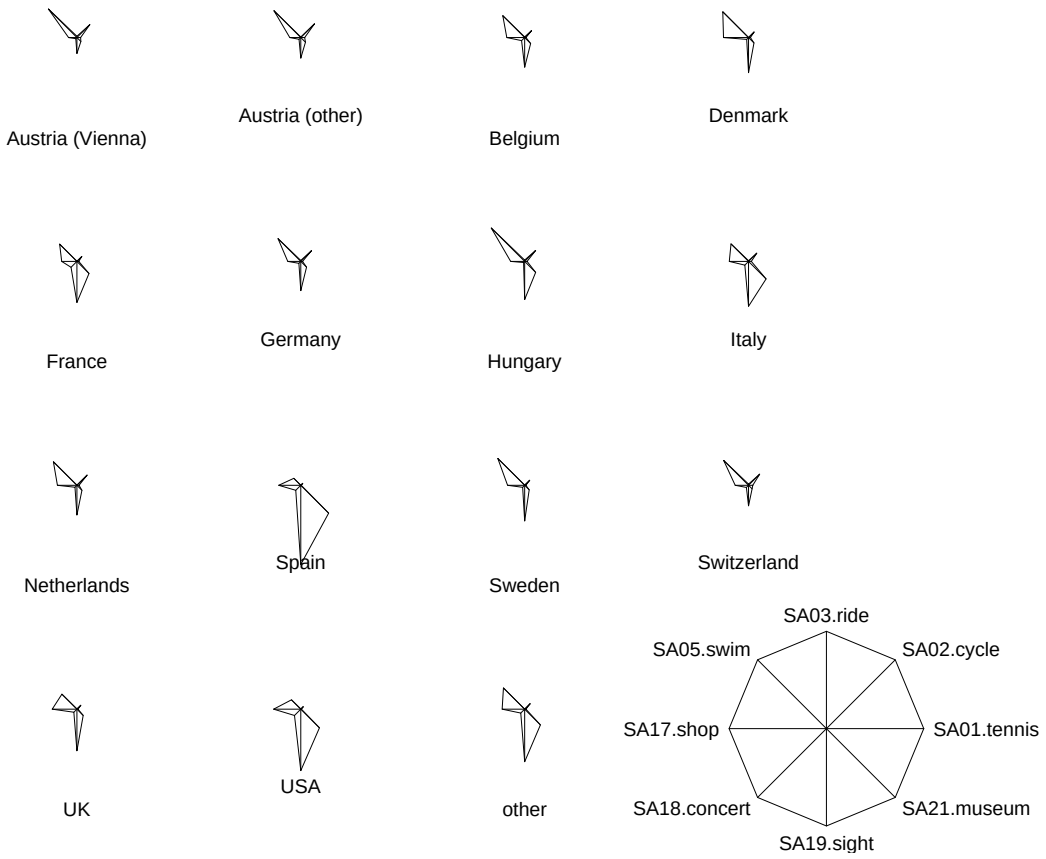
# Explorative Grafik



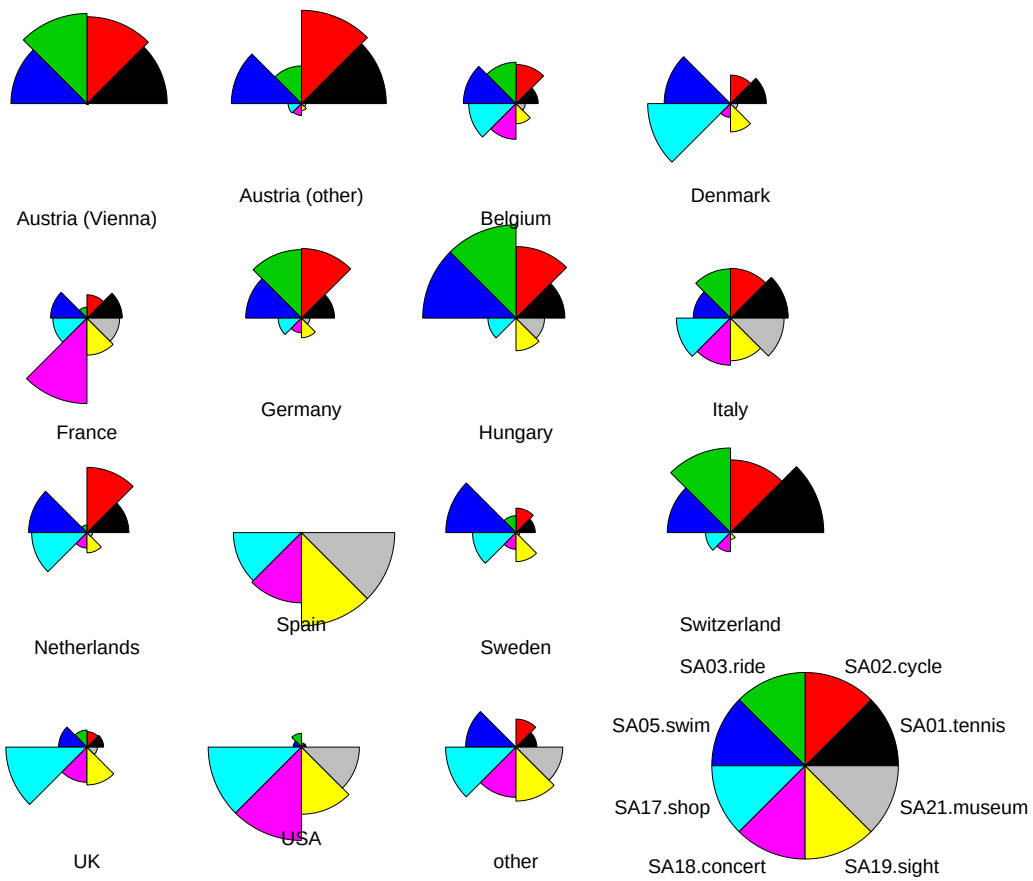
# Explorative Grafik



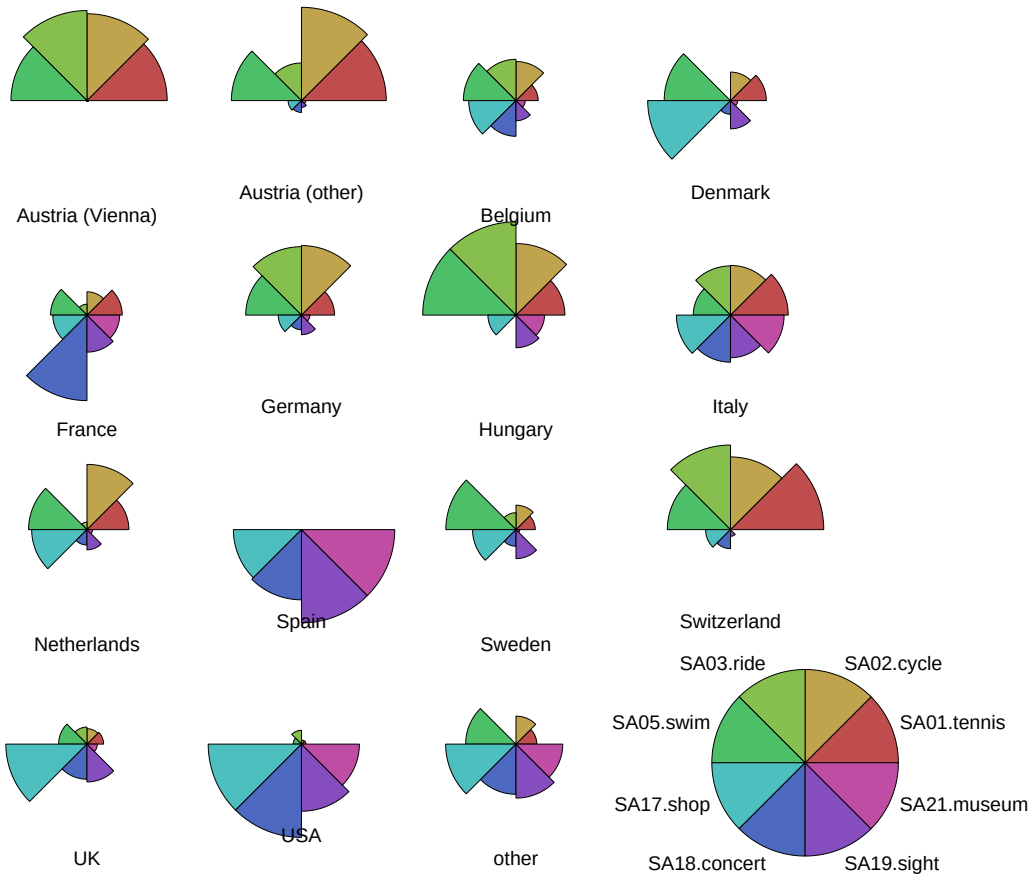
# Explorative Grafik



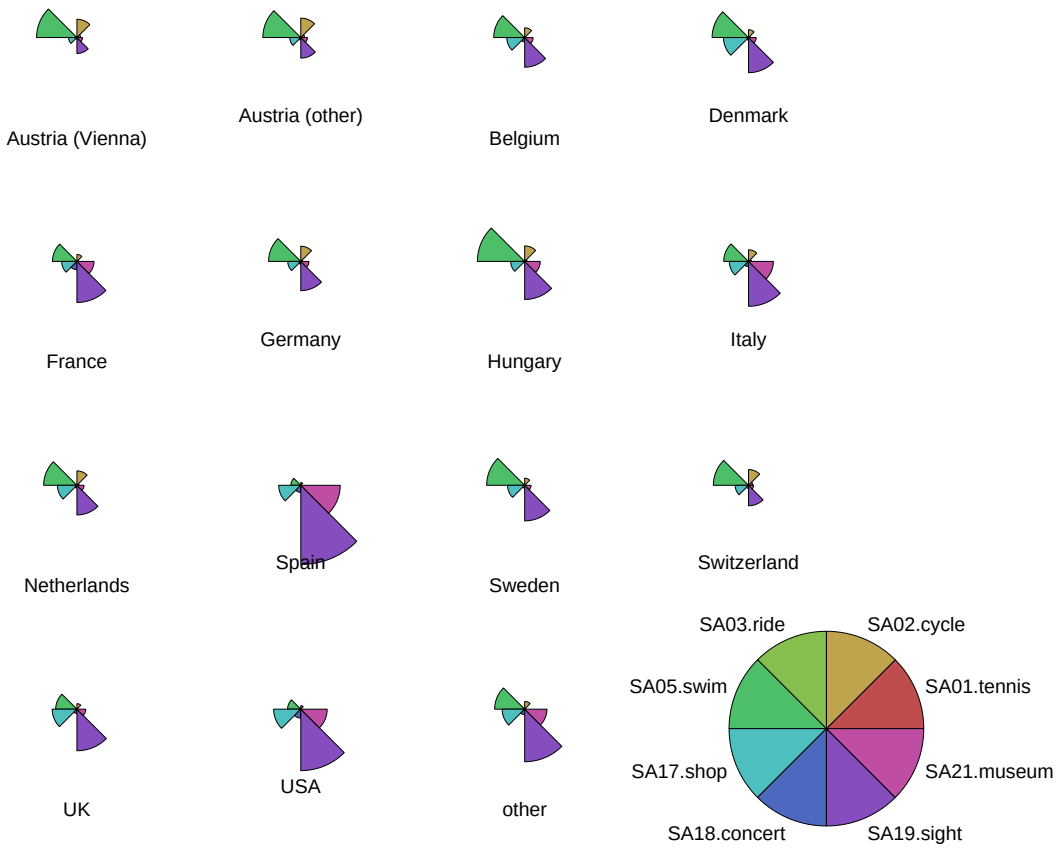
# Explorative Grafik



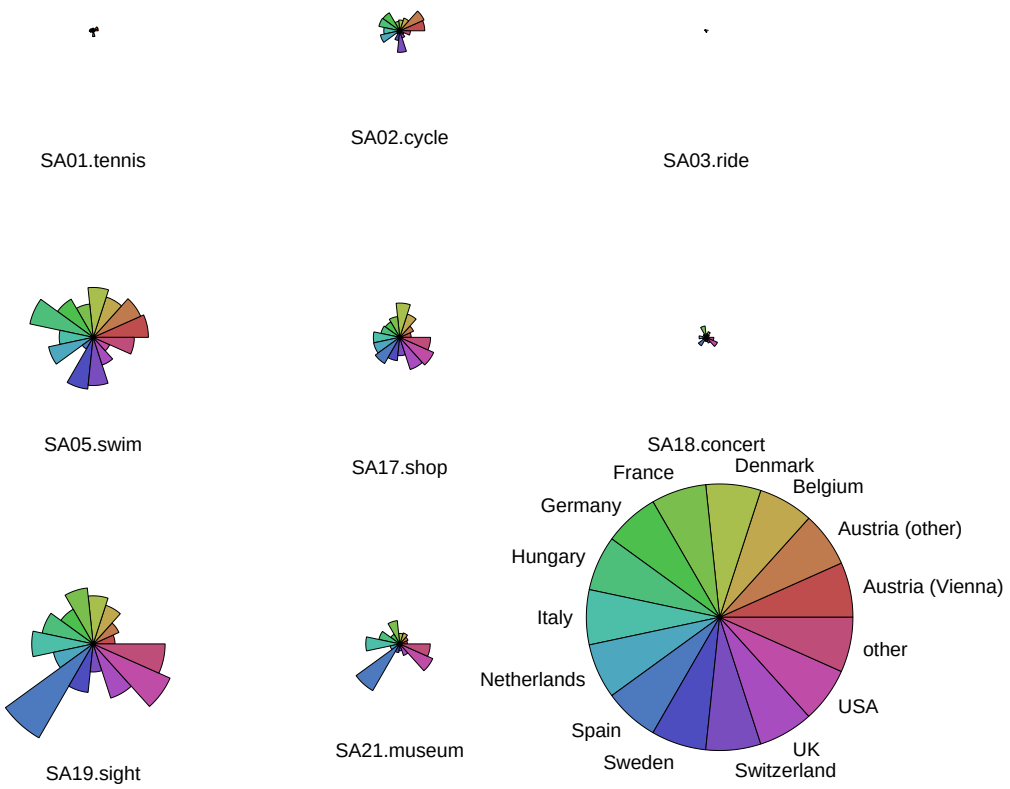
# Explorative Grafik



# Explorative Grafik



# Explorative Grafik



# Tutorium

Explorative multivariate Analyse in R ($MVA.pdf$)

# Übung

**Aufgabe 11:**

102 Kinder haben 11 verschiedene Eissorten (u.a. Magnum, Cornetto, Calippo) getestet und dann jede Eissorte bezüglich 14 Eigenschaften (u.a. 'tastes excellent', 'looks good') bewertet. Dabei konnten sie jede Aussage nur als zutreffend oder unzutreffend einstufen. Der Datensatz $Ice.rda$ enthält aggregierte Daten, die für jede Kombination von Eissorte und Aussage den Anteil der Kinder angeben, die die Aussage als zutreffend eingestuft haben.

- Versuchen Sie die Daten mit Hilfe von verschiedenen Grafiken zu visualisieren.
- Können Sie irgendwelche Strukturen erkennen?

# Übung

Die 14 Aussagen sind im Detail: 1 tastes excellent, 2 looks good, 3 satisfies my hunger, 4 satisfies my thirst, 5 refreshing, 6 for everyday, 7 for special occasions, 8 for adults, 9 my favorite, 10 different, 11 cool, 12 I would never eat that, 13 fun for children, 14 expensive.

# Hauptkomponentenanalyse

Mit den paarweisen Streudiagrammen haben wir bereits eine Visualisierungstechnik kennengelernt, die eine Projektion der gesamten Datenmatrix $X$ aus dem $\mathbb{R}^p$ in den $\mathbb{R}^2$ vornimmt.

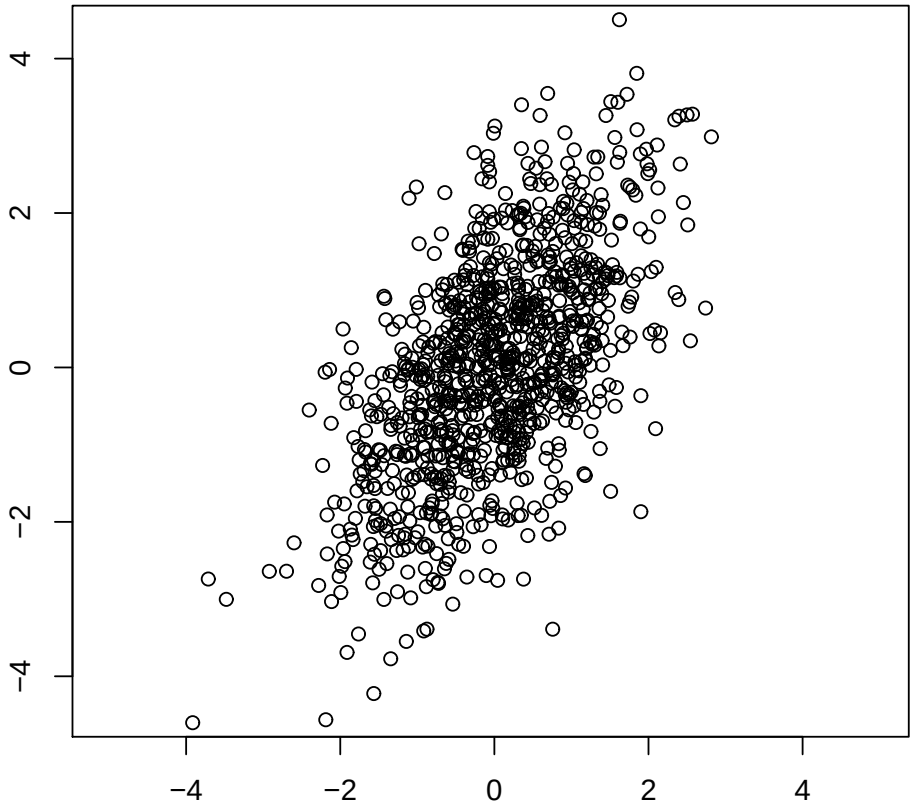
**Problem:**

1. Es gibt eine große Anzahl von möglichen 2-dimensionalen Projektionen ohne eine bestimmte Reihenfolge.
2. Man projiziert nur entlang der Achsen.

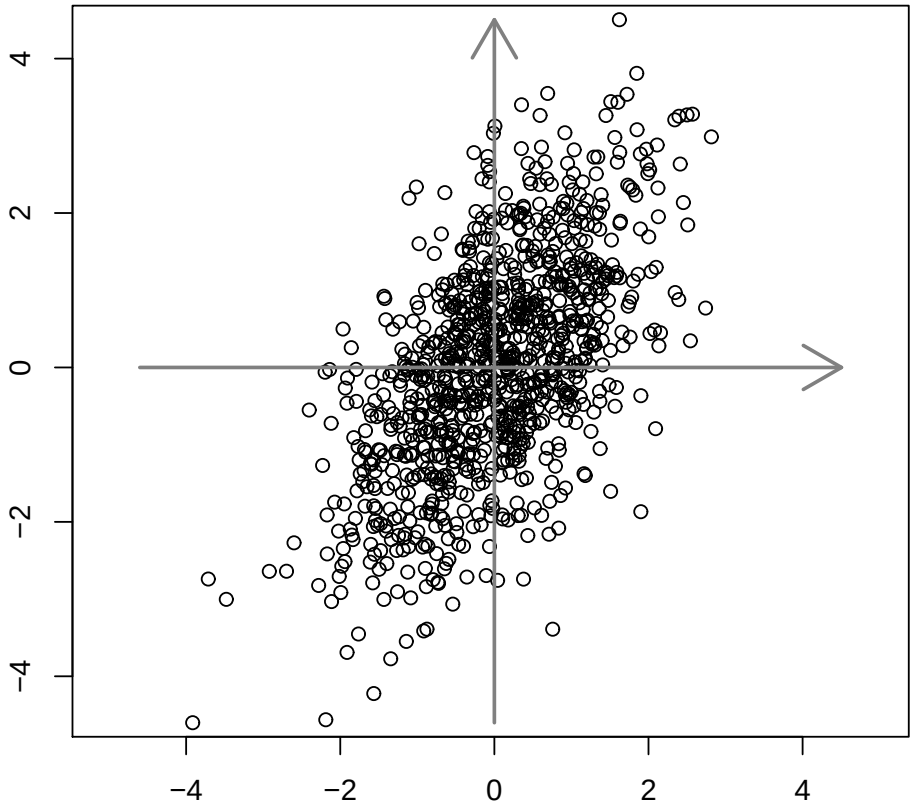
**Lösung:**

Betrachte Linearkombinationen $Xa_j$ $(j = 1, \ldots, p)$ mit beliebigen Koeffizienten $a_j$, so daß $Xa_1$ die "interessanteste" Linearkombination ist und $Xa_2$ die "zweit-interessanteste" usw.

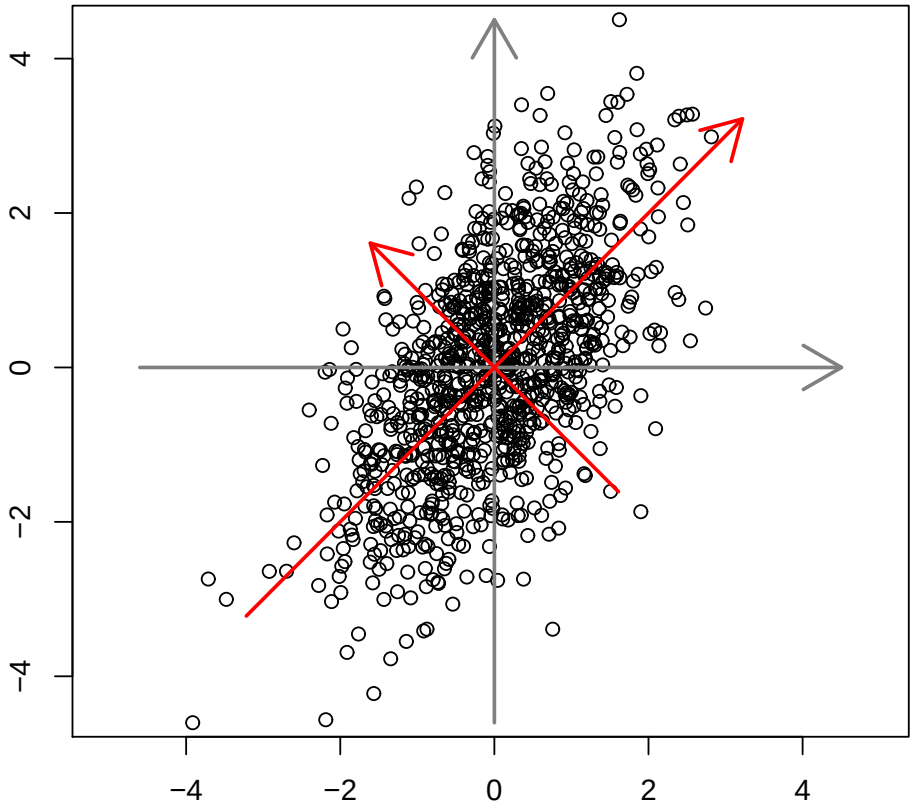
# Hauptkomponentenanalyse



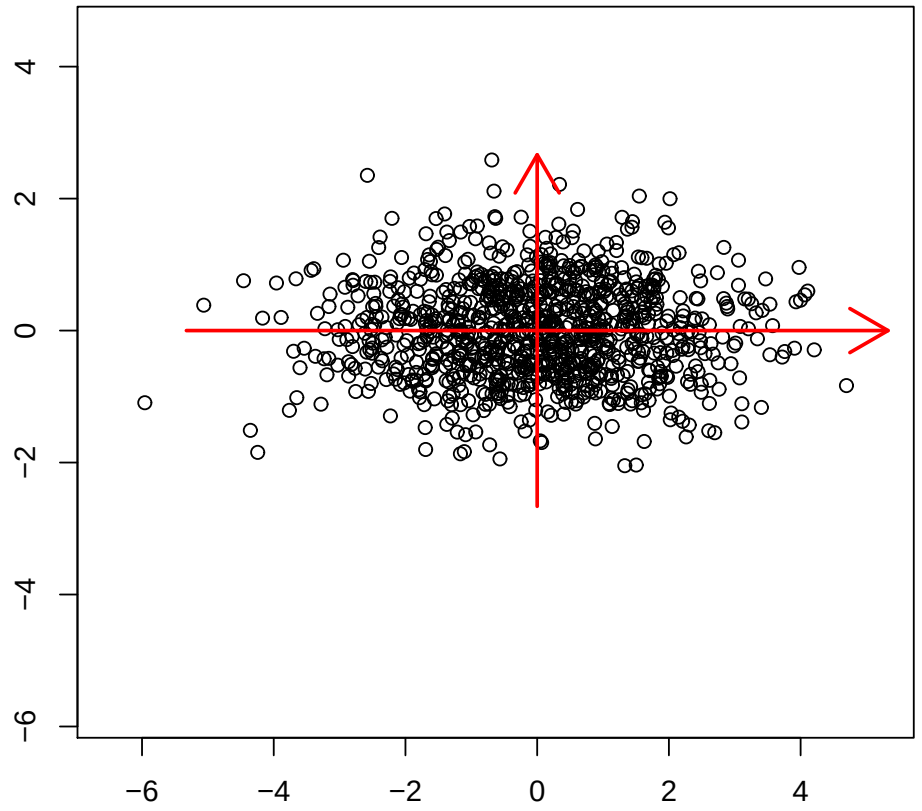
# Hauptkomponentenanalyse



# Hauptkomponentenanalyse



# Hauptkomponentenanalyse



# Hauptkomponentenanalyse

Formal gesprochen heißt "interessant" immer "mit hoher Varianz", da man ja genau versucht die zufällige Variation der Daten zu verstehen.

Man versucht also durch Linearkombinationen $XA$ ein neues Koordinatensystem zu finden, in dem die erste Variable die "meiste Information" diesbezüglich enthält, die zweite die "zweitmeiste Information" usw. Mit etwas Glück enthalten dann die ersten paar (bspw. $2$ oder $3$) fast die gesamte Information und die verbleibenden ($p-2$ bzw. $p-3$) Variablen können bei der weiteren Analyse vernachlässigt werden.

# Hauptkomponentenanalyse

Die durch Linearkombinationen neu konstruierten Variablen nennt man auch Hauptkomponenten (engl.: principal components).

Die Hauptkomponentenanalyse (engl.: principal component analysis, PCA) ist eine Technik zur Dimensionsreduktion.

# Hauptkomponentenanalyse

**Konstruktion:**

Sei $S = V(X)$ die Kovarianzmatrix der Daten $X$. Dann wollen wir eine Linearkombination $Xa_1$ finden, so daß $V(Xa_1) = a_1^\top S a_1$ maximal wird.

Um dieses Problem identifizierbar zu machen, legen wir außerdem die Nebenbedingung $a_1^\top a_1 = 1$ an.

Die Lagrange-Funktion, die hier optimiert werden muß ist daher

$$f(a_1, \lambda_1) \quad = \quad a_1 S a_1 \; - \; \lambda_1 (a_1^\top a_1 - 1)$$

Das heißt $f(a_1, \lambda_1)$ muß nach $a_1$ und nach $\lambda_1$ differenziert und dann gleich $0$ gesetzt werden.

# Hauptkomponentenanalyse

Differenzierung nach $\lambda_1$ ergibt die Nebenbedingung und Differenzierung nach $a_1$

$$\begin{aligned} 2Sa_1 - 2\lambda_1 a_1 &= 0 \\ Sa_1 &= \lambda_1 a_1 \end{aligned}$$

Ein Vektor $a_1$ mit dieser Eigenschaft heißt Eigenvektor der Matrix $S$ und $\lambda_1$ ist der zugehörige Eigenwert.

Daher kann man die Hauptkomponenten aus der Eigenwertzerlegung der Kovarianzmatrix $S$ berechnen.

# Hauptkomponentenanalyse

**Alternative:** Statt mit der Kovarianzmatrix rechnet man mit der Korrelationsmatrix um sich des Problems unterschiedlicher Skalen zu entledigen.

Das entspricht wieder eine Skalierung der Daten $X$ zu $\hat{X}$, so daß jede Spalte von $\hat{X}$ den Mittelwert $0$ und die Varianz $1$ hat:

$$\hat{x}_{ij} \quad = \quad \frac{x_{ij} - \bar{x}_j}{SD_j}$$

# Hauptkomponentenanalyse

**Beispiel:** Hauptkomponentenzerlegung für die aggregierten GSA-Daten für $n = 15$ Länder und $p = 8$ Aktivitäten.

Die Hauptkomponentenanalyse berechnet die $8 \times 8$ Rotationsmatrix $A$, die die Koeffizienten/Eigenvektoren $a_1, \ldots, a_8$ enthält. Diese nennt man auch Ladungen der Hauptkomponenten.

Die zugehörgigen Eigenwerte $\lambda_1, \ldots, \lambda_8$ geben an, welcher Anteil der Gesamtvarianz durch die entsprechende Hauptkomponente eingefangen wird.

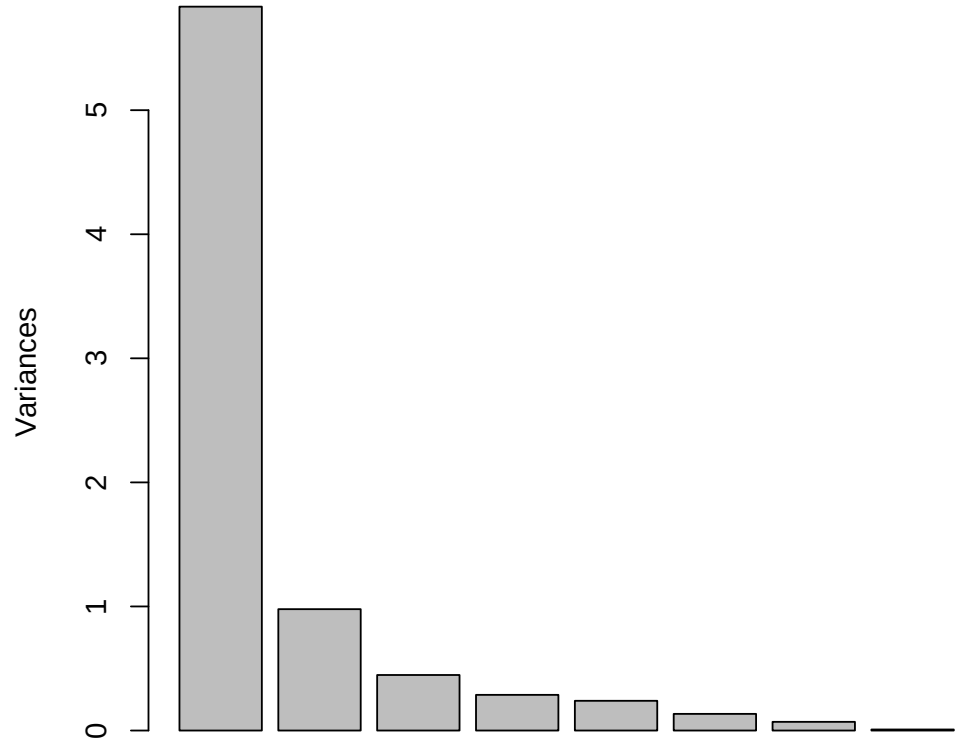
**Bemerkung:** Da die Eigenvektoren nur bis auf einen Faktor definiert sind, sind die Hauptkomponenten nur bis auf das Vorzeichen definiert.

# Hauptkomponentenanalyse

| Aktivität | $PC_1$ | $PC_2$ | ... |
|---|---|---|---|
| Tennis | -0.351 | -0.287 | |
| Radfahren | -0.387 | -0.202 | |
| Reiten | -0.315 | -0.477 | |
| Schwimmen | -0.363 | 0.249 | |
| Shopping | 0.344 | 0.387 | |
| Konzert | 0.351 | -0.303 | |
| Sightseeing | 0.384 | -0.232 | |
| Museum | 0.327 | -0.541 | |

Die erste Hauptkomponente fängt dabei 73% der Varianz ein, die zweite 12%. Einen Plot der Varianzen nennt man auch Screeplot.

# Hauptkomponentenanalyse



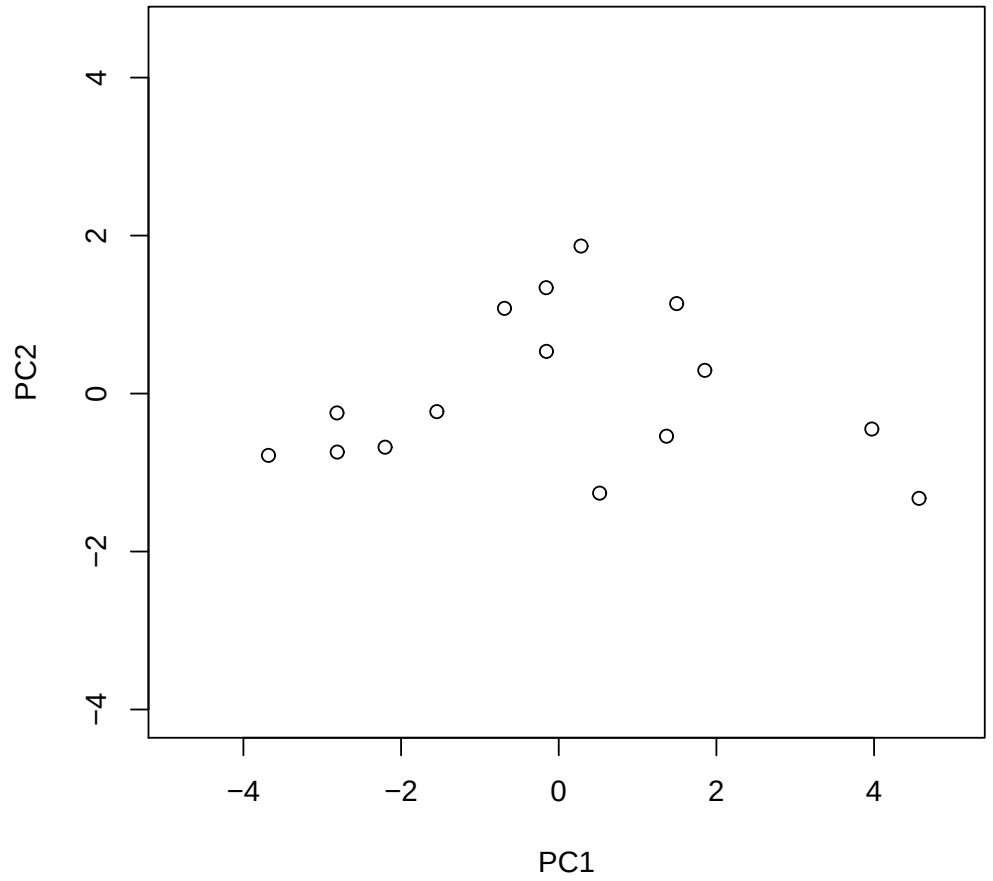
# Hauptkomponentenanalyse

Hier fängt also die erste Hauptkomponente bereits den grössten Teil der Varianz ein.

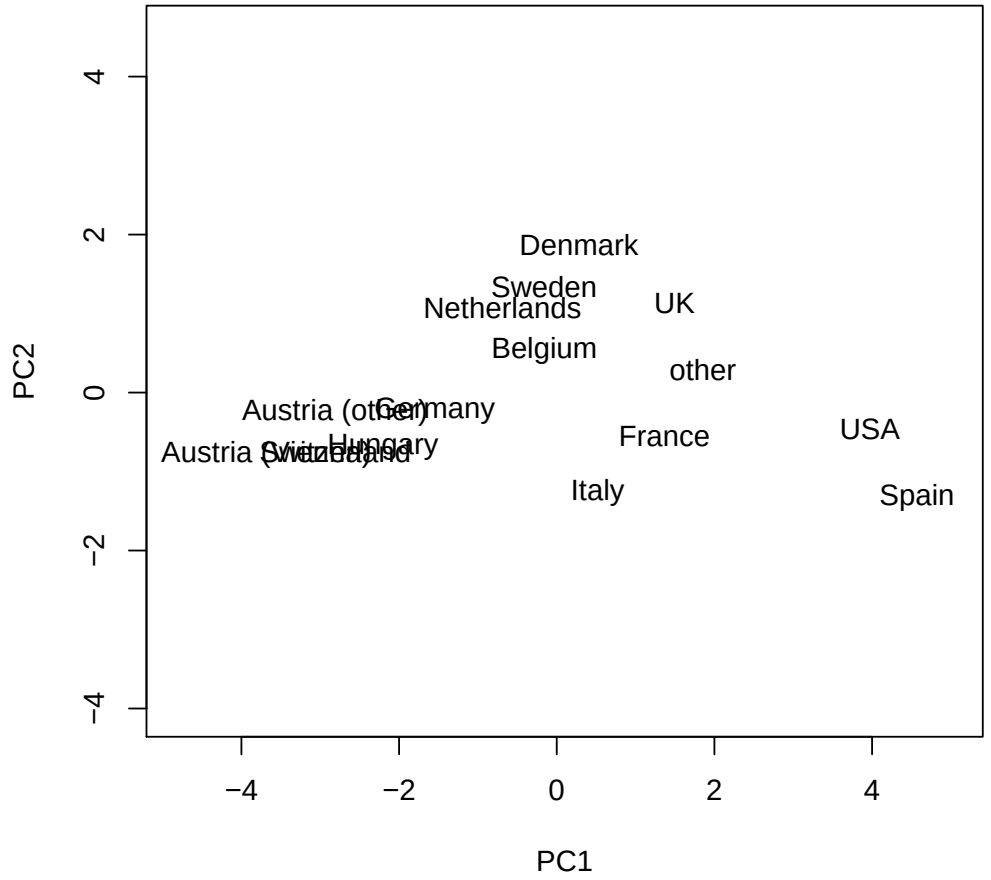
Alle Koeffizienten von $PC_1$ haben in etwa denselben Absolutbetrag, aber unterschiedliche Vorzeichen. Damit kontrastiert $PC_1$ die durchschnittliche sportliche Aktivität (Tennis, Rad, Reiten Schwimmen) mit der durchschnittlichen kulturellen Aktivität (Shopping, Konzert, Sightseeing, Museum). Ein hoher Wert bei $PC_1$ spricht für hohe kulturelle Aktivität, ein niedriger für hohe sportliche Aktivität.

Die zweite Komponente $PC_2$ kontrastiert vor allem die Kombination Shopping/Schwimmen mit den übrigen Variablen.

# Hauptkomponentenanalyse



# Hauptkomponentenanalyse



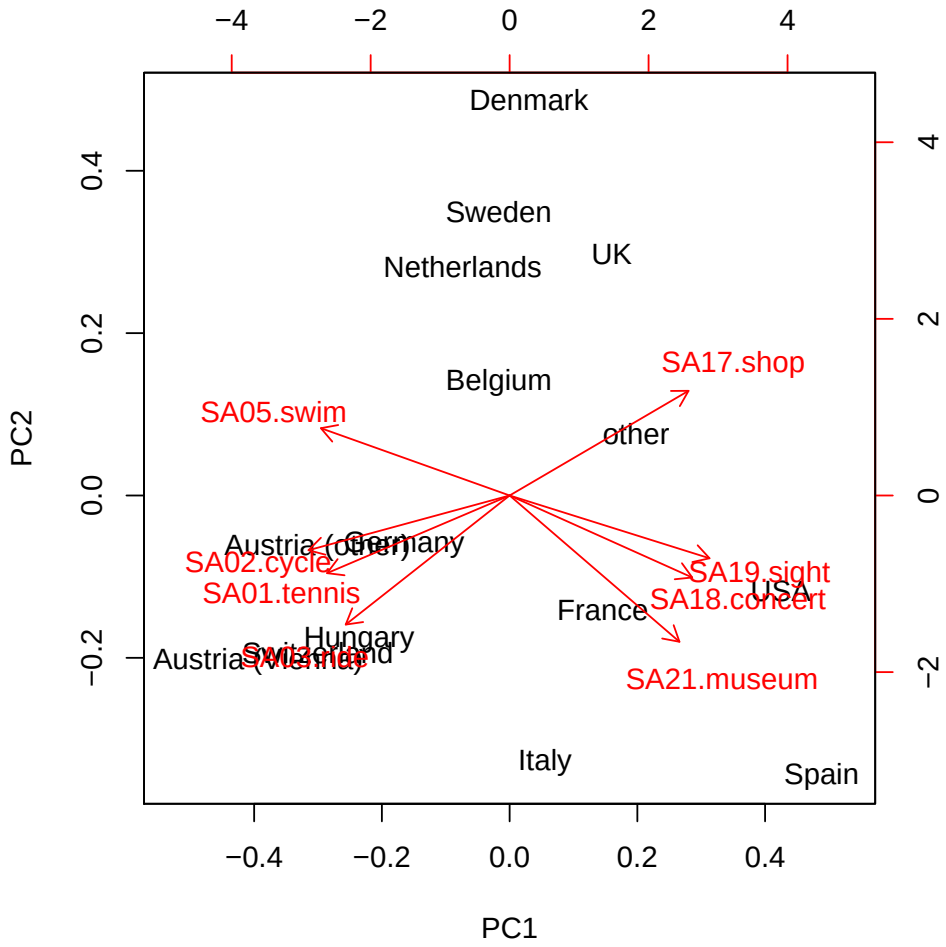
# Hauptkomponentenanalyse

**Biplot:**

Wenn zusätzlich zu den Beobachtungen auch noch die Projektion der ursprünglichen Achsen in die Grafik der (ersten beiden) Hauptkomponenten zeichnet, dann nennt man sie Biplot.

Achsen die in eine ähnliche Richtung zeigen (also nur einen geringen Winkel nach der Projektion haben), messen ähnliche Konzepte. Präziser formuliert entsprechen Winkel Korrelationen zwischen den Variablen.

# Hauptkomponentenanalyse



# Tutorium

Hauptkomponentenanalyse in R ($PCA.pdf$)

# Übung

**Aufgabe 12:**

Führen Sie eine Hauptkomponentenanalyse (mit Skalierung) der `Ice` Daten durch.

- Wie viele Hauptkomponenten muß man bei der Analyse mindestens berücksichtigen?
- Welche Konzepte messen die ersten Hauptkomponenten?
- Visualisieren Sie das Ergebnis geeignet.

# Übung

**Aufgabe 13:**

Der Datensatz `SwissBank` enthält 6 verschiedene physische Abmessungen von 200 Schweizer Banknoten (u.a. Randbreite, Diagonale, Höhe, usw.). Einige dieser Banknoten sind echt, andere Falschgeld. Führen Sie eine Hauptkomponentenanalyse (ohne Skalierung) durch.

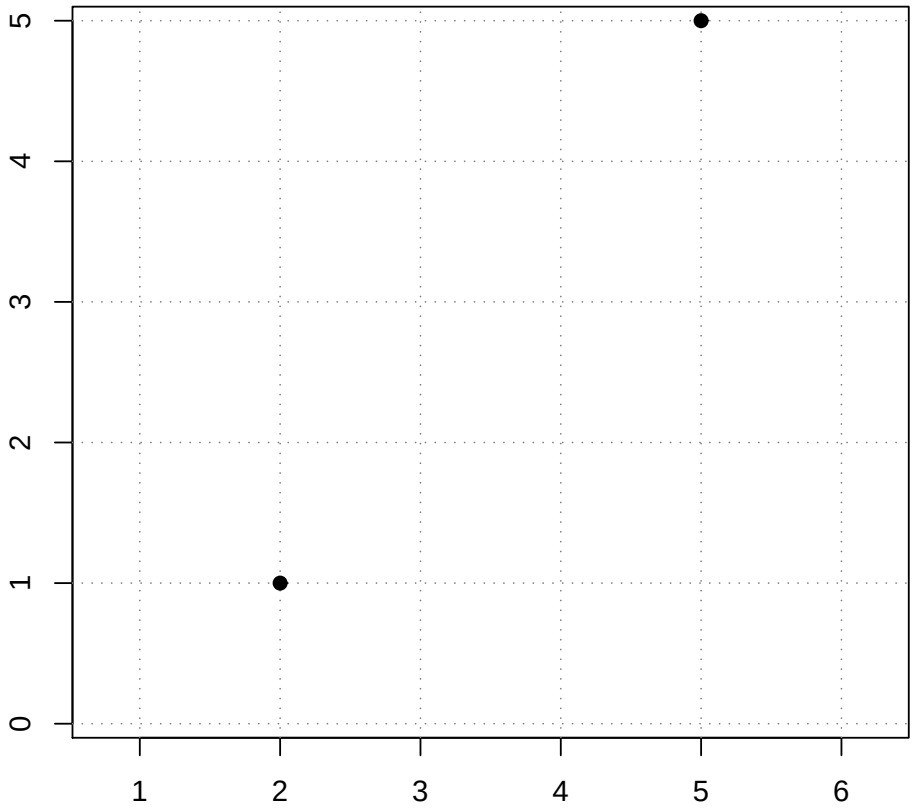
- Wie viele Hauptkomponenten muß man bei der Analyse mindestens berücksichtigen?
- Visualisieren Sie die Daten. Gibt es Gruppen in den Daten?
- Welche Eigenschaften führen zu einer guten Diskriminierung der Geldscheine?

# Distanzen

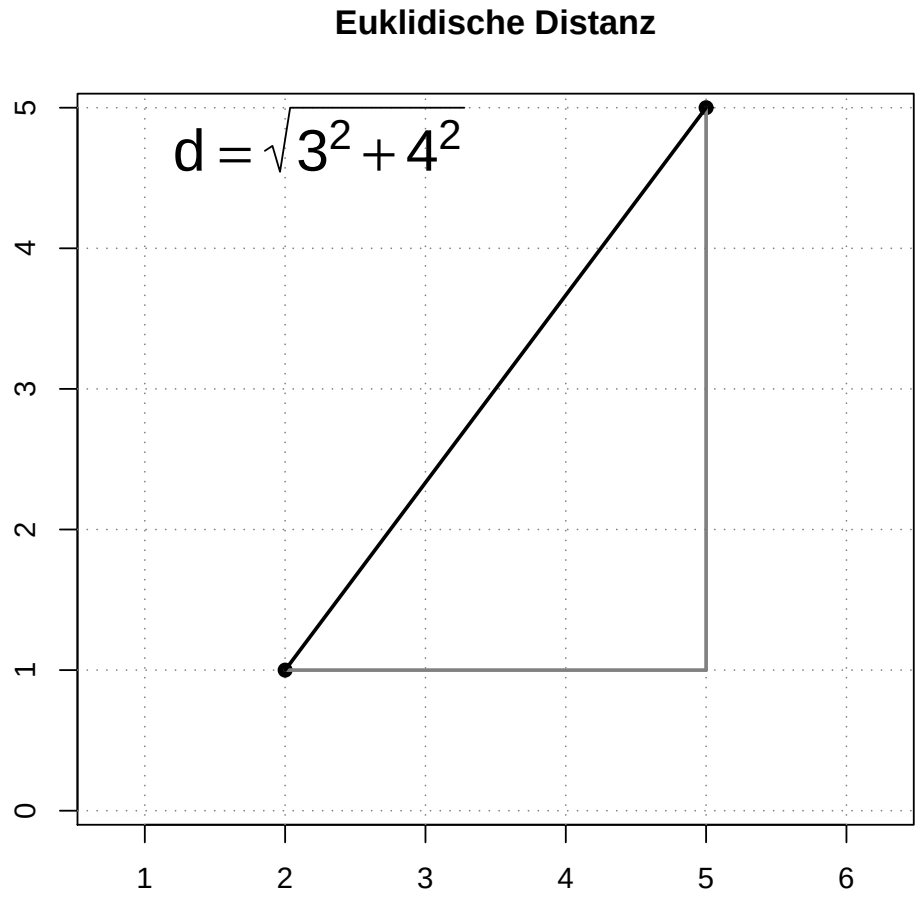
In den vorangegangenen Abbildungen der Hauptkomponenten haben wir immer die Distanzen zwischen Punkten betrachtet: Punkte, die nahe beieinander liegen, sind ähnlich – Punkte, die weit entfernt liegen, sind unähnlich.

Nun kann man sich also fragen, wie man genereller die Distanzen (d.h. Unähnlichkeiten) zwischen zwei Beobachtungen (d.h. Zeilen) der Datenmatrix $X$ messen kann.

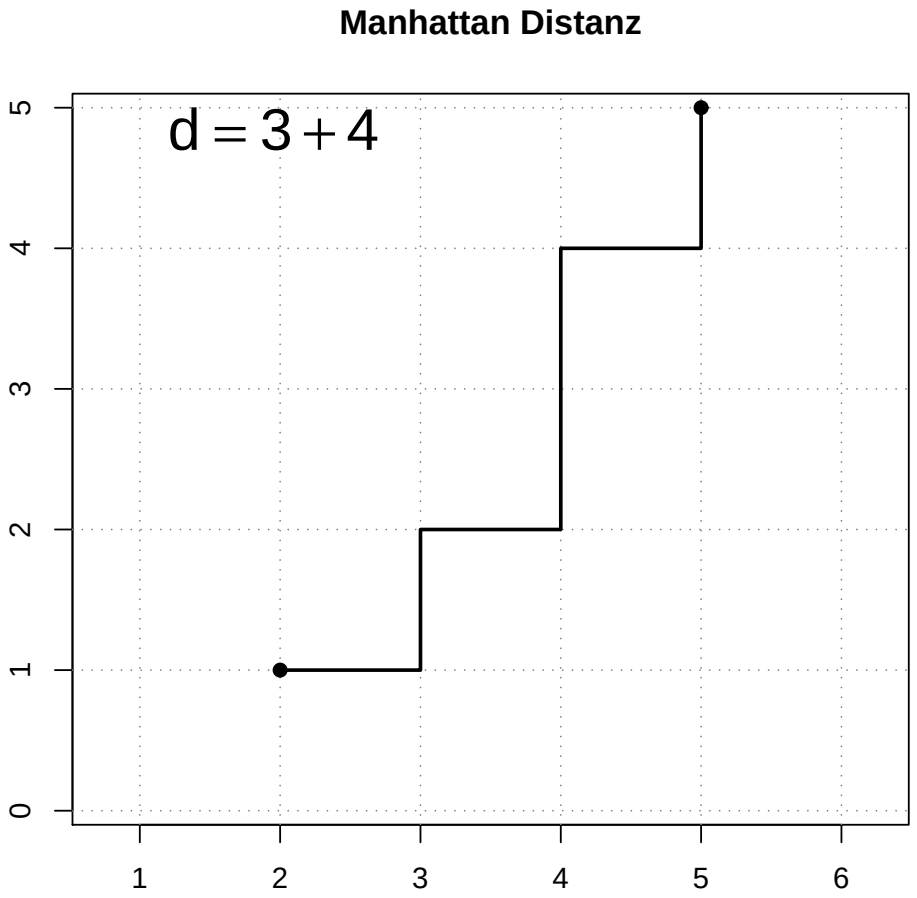
# Distanzen



# Distanzen

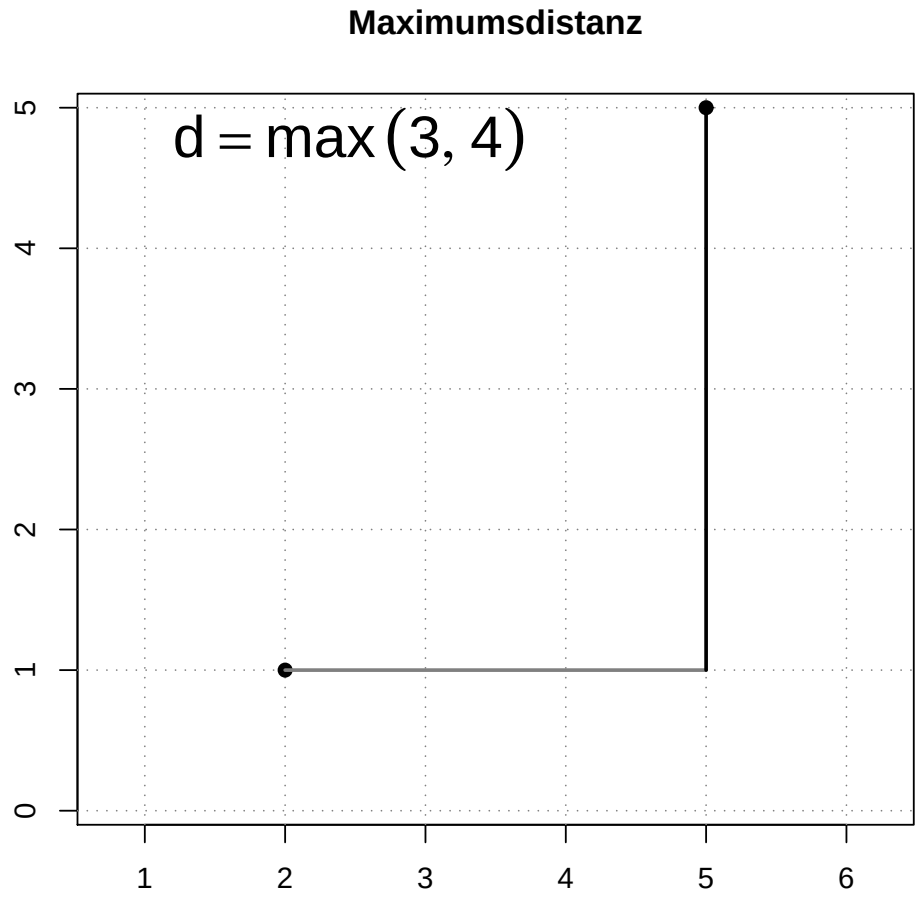


# Distanzen



# Distanzen

# Distanzen



# Distanzen

Seien also $x_1$ und $x_2$ zwei verschiedene Beobachtungen/Zeilen aus $X$, dann sind:

**Manhattan Distanz:**

$$d_1(x_1, x_2) = \sum_{j=1}^{p} |x_{1j} - x_{2j}|$$

**Euklidische Distanz:**

$$d_2(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (x_{1j} - x_{2j})^2}$$

# Distanzen

**Maximumsdistanz:**

$$d_\infty(x_1, x_2) \quad = \quad \max_{j=1,\ldots,p} |x_{1j} - x_{2j}|$$

**Canberra Distanz:**

$$d_{\mathrm{C}}(x_1, x_2) \quad = \quad \sum_{j=1}^{p} \frac{|x_{1j} - x_{2j}|}{|x_{1j} + x_{2j}|}$$

# Distanzen

Zusätzlich gibt es spezielle Distanzen für binäre Merkmale. Am gängigsten ist die binäre Distanz, die jeweils den Anteil von Variablen angibt, die für beide Beobachtungen kein Erfolg ist, unter den Variablen, wo zumindest eine der Beobachtungen ein Erfolg ist.

Man läßt also zuerst alle Spalten weg, wo beide Merkmale $0$ sind. Dann berechnet man den Anteil diskordanter Spalten (wo eine Beobachtung $1$ und die andere $0$ ist) unter den verbleibenden Spalten.

# Multidimensionale Skalierung

Mit jeder der vorher definierten Distanzen kann aus einer gegebenen $n \times p$ Matrix von Beobachtungen $X$ eine $n \times n$ Matrix von paarweisen Distanzen $D$ berechnet werden.

**Frage:** Ist die Umkehrung auch möglich?

**Antwort:**

Ja, ist immer exakt möglich, wenn $p = n - 1$ (unter bestimmten Regularitätsvoraussetzungen an $D$).

# Multidimensionale Skalierung

**Frage:** Ist dasselbe auch möglich für "kleines" $p$ (bspw. $p = 2$)?

**Antwort:**

Ja, wenn so eine Lösung existiert. Sonst können approximative Lösungen gefunden werden.

Solche Verfahren nennt man Multidimensionale Skalierung (MDS), das bekannteste ist die sogenannte klassische MDS. Die resultierende Matrix $X$ der Skalierung nennt man Konfiguration, sie ist bis auf Rotationen (insbesonders also Vorzeichenwechsel) definiert.

# Multidimensionale Skalierung

**Beispiel:** Distanzen zwischen Sprachen

Um Abstände zwischen verschiedenen Sprachen (u.a. Englisch, Deutsch, Dänisch, ...) zu messen, wird eine sehr einfache Distanz verwendet: man zählt einfach, wie viele Wörter für die Zahlen 1 bis 10 mit unterschiedlichen Buchstaben beginnen.

In Englisch und Deutsch fangen one/eins, two/zwei, three/drei, four/vier, eight/acht, ten/zehn mit unterschiedlichen Buchstaben an, hingegen five/fünf, six/sechs, seven/sieben und nine/neun mit den gleichen Buchstaben. Der Abstand beträgt also 6.

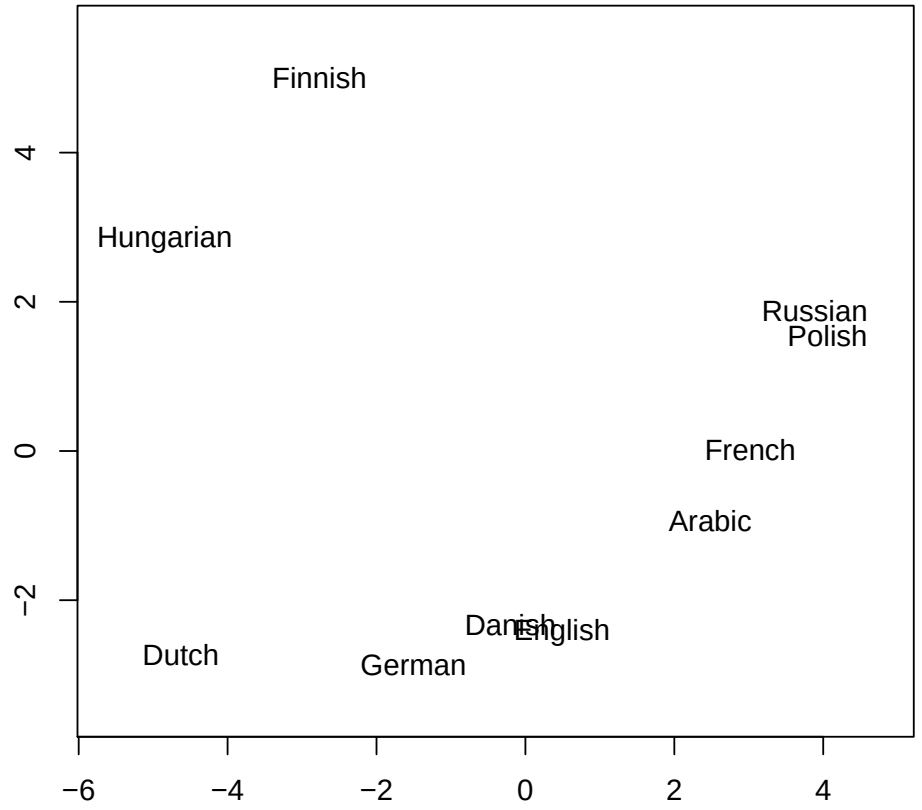
# Multidimensionale Skalierung

Für die ersten vier betrachteten Sprachen ergibt sich daraus eine Distanzmatrix

| | Englisch | Dänisch | Holländisch | Deutsch |
|---|---|---|---|---|
| Englisch | 0 | 2 | 7 | 6 |
| Dänisch | 2 | 0 | 6 | 5 |
| Holländisch | 7 | 6 | 0 | 5 |
| Deutsch | 6 | 5 | 5 | 0 |

Die zugehörigen klassische MDS ergibt folgende Konfiguration.

# Multidimensionale Skalierung



# Tutorium

Multidimensionale Skalierung in R ($MDS.pdf$)

# Übung

**Aufgabe 14:**

Der Datensatz `autodist` gibt die Distanzen zwischen den neun österreichischen Landeshauptstädten in Kilometern an (nach dem Shell-Online Autoatlas). Versuchen Sie aus diesen Distanzen eine Karte der Landeshauptstädte zu rekonstruieren. (Hinweis: Konfigurationen können gespiegelt werden.)

Der Datensatz `oebbdist` gibt dieselben Distanzen gemäß der OeBB-Bahnverbindungen an. Konstruieren Sie auch hier eine Karte und vergleichen Sie die Konfiguration mit der auf den Autodistanzen basierenden.

# Clusteranalyse

Eine Klassifikation von Objekten (Beobachtungen) ist eine Einteilung dieser Objekte in Gruppen (Cluster), so daß der Abstand der Objekte innerhalb einer Gruppe möglichst klein, aber zwischen den Gruppen möglichst groß ist.

Man möchte also

- Homogenität/Ähnlichkeit innerhalb der Cluster,
- Heterogenität/Unähnlichkeit zwischen den Clustern

erzielen.

# Clusteranalyse

Die Klassifikationen, die wir hier betrachten, sind entweder Partitionen oder Hierarchien von Partitionen.

Eine **Partition** ist eine vollständige und alternative Zerlegung der $n$ Objekte in $k$ Cluster, d.h. jede Beobachtung ist genau einem Cluster zugehörig.

Die Vereinigung aller Cluster ergibt also die Gesamtheit aller Objekte, während alle paarweisen Durchschnitte der Cluster leer sind.

# Clusteranalyse

**Formal:** Eine Partition $\mathcal{C}$ ist eine Menge von $k$ Clustern

$$\mathcal{C} \quad = \quad \{C_1, \ldots, C_k\}$$

wobei jeder Cluster $C_j$ eine Menge von Objekten ist, so daß

$$C_1 \cup \ldots \cup C_k \quad = \quad \{x_1, \ldots, x_n\}$$
$$C_i \cap C_j \quad = \quad \emptyset$$

Eine **Hierarchie** von Partitionen ist eine Folge von Partitionen, so daß $\mathcal{C}_j$ und $\mathcal{C}_{j+1}$ sich nur dadurch unterscheiden, daß mindestens ein Cluster aus $\mathcal{C}_j$ nochmals partitioniert wurde.

# Clusteranalyse

Um nun die Heterogenität zwischen den Clustern beurteilen zu können, benötigen wir geeignete Distanzen $D(\cdot, \cdot)$ zwischen Clustern. Diese werden üblicherweise basierend auf den paarweisen Distanzen zwischen den Objekten aus den Clustern $d(\cdot, \cdot)$ berechnet.

Dadurch kann diesen Heterogenitätsmaßen jedes beliebige Distanzmaß von Objekten zugrunde gelegt werden. Einige solcher Distanzmaße haben wir bereits kennengelernt.

# Clusteranalyse

Die Distanz von zwei Clustern wird durch die Distanz der beiden ähnlichsten Objekte der Cluster definiert.

$$D_{\mathrm{s}}(C_1, C_2) \quad = \quad \min_{x \in C_1, y \in C_2} d(x, y)$$

Cluster Verfahren, die dieses Heterogenitätsmaß verwenden, heißen single linkage Verfahren oder Verfahren der nächsten Nachbarn.

**Problem:** Die Heterogenität wird tendenziell unterschätzt.

# Clusteranalyse

Die Distanz von zwei Clustern wird durch die Distanz der beiden unähnlichsten Objekte der Cluster definiert.

$$D_{\mathrm{c}}(C_1, C_2) \quad = \quad \max_{x \in C_1, y \in C_2} d(x, y)$$

Cluster Verfahren, die dieses Heterogenitätsmaß verwenden, heißen complete linkage Verfahren oder Verfahren der weitesten Nachbarn.

**Problem:** Die Heterogenität wird tendenziell überschätzt.

# Clusteranalyse

Die Distanz von zwei Clustern wird durch die mittlere Distanz der Objekte der Cluster definiert.

$$D_{\mathrm{a}}(C_1, C_2) \quad = \quad \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y)$$

Cluster Verfahren, die dieses Heterogenitätsmaß verwenden, heißen average linkage Verfahren.

# Clusteranalyse

Eine weitere Methode um Distanzen zwischen Clustern zu definieren, ist die Methode von **Ward**.

Die Idee ist dabei, daß eine Art Varianzanalyse durchgeführt wird, die die Fehlerquadratsumme in zwei Clustern mit der Fehlerquadratsumme des resultierenden vereinigten Clusters vergleicht.

# Clusteranalyse

Analog können Maße für die Homogenität innerhalb eines Clusters definiert werden:

- maximale Distanz,
- minimale Distanz oder
- durchschnittliche Distanz

von jeweils zwei Objekten innerhalb desselben Clusters.

# Hierarchisches Clustern

Clusterverfahren, die Hierarchien von Partitionen erzeugen, nennt man hierarchische Clusterverfahren. Diese lassen sich unterteilen in divisive und agglomerative Verfahren.

**Divisive** Verfahren beginnen mit einem einzigen Cluster, der alle Objekte enthält, partitionieren diesen Cluster, und wiederholen dies rekursiv für jeden Cluster der entstandenen Partition. In der Regel wird in jedem Schritt genau ein Cluster in zwei neue Cluster zerlegt.

# Hierarchisches Clustern

**Agglomerative** Verfahren gehen genau umgekehrt vor: Sie fangen also mit einer Partition an, in der jeder Cluster nur genau ein Objekt enthält, und legen dann rekursiv Cluster zusammen.

Die hier betrachteten Verfahren sind von der Form:

1. Starte mit $n$ Clustern (einer für jedes Objekt). Die Distanzen zwischen den Clustern sind einfach die Distanzen zwischen den Objekten.
2. Lege die beiden ähnlichsten Cluster zusammen.
3. Berechne die Distanzen des neuen zu allen übrigen Clustern.
4. Wiederhole 2. und 3. bis es nur noch einen Cluster (mit allen Objekten) gibt.

# Hierarchisches Clustern

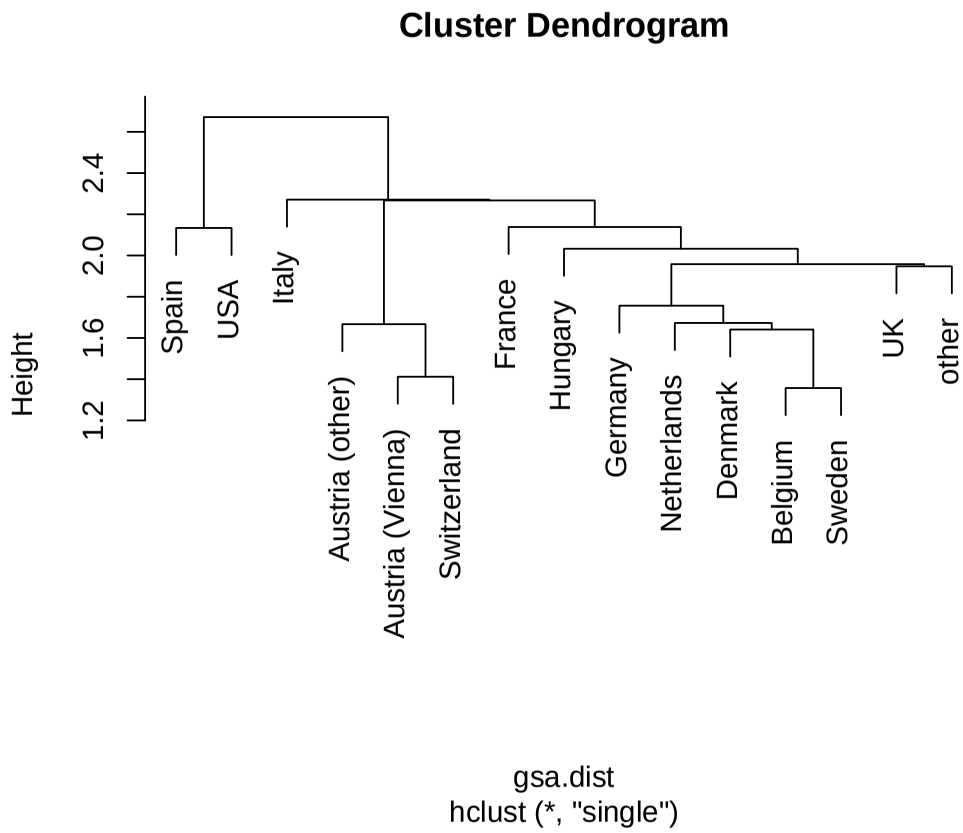
In Schritt 1. bei der Berechnung der **Distanzen zwischen den Objekten** kann ein beliebiges Distanzmaß verwendet werden. Bei metrischen Merkmalen werden in aller Regel euklidische Distanzen verwendet.

In Schritt 3. bei der Berechnung der **Distanzen zwischen den Clustern** kann eines der oben definierten Distanzmaße verwendet werden. Einige Eigenschaften der verschiedenen Verfahren lassen sich festhalten.

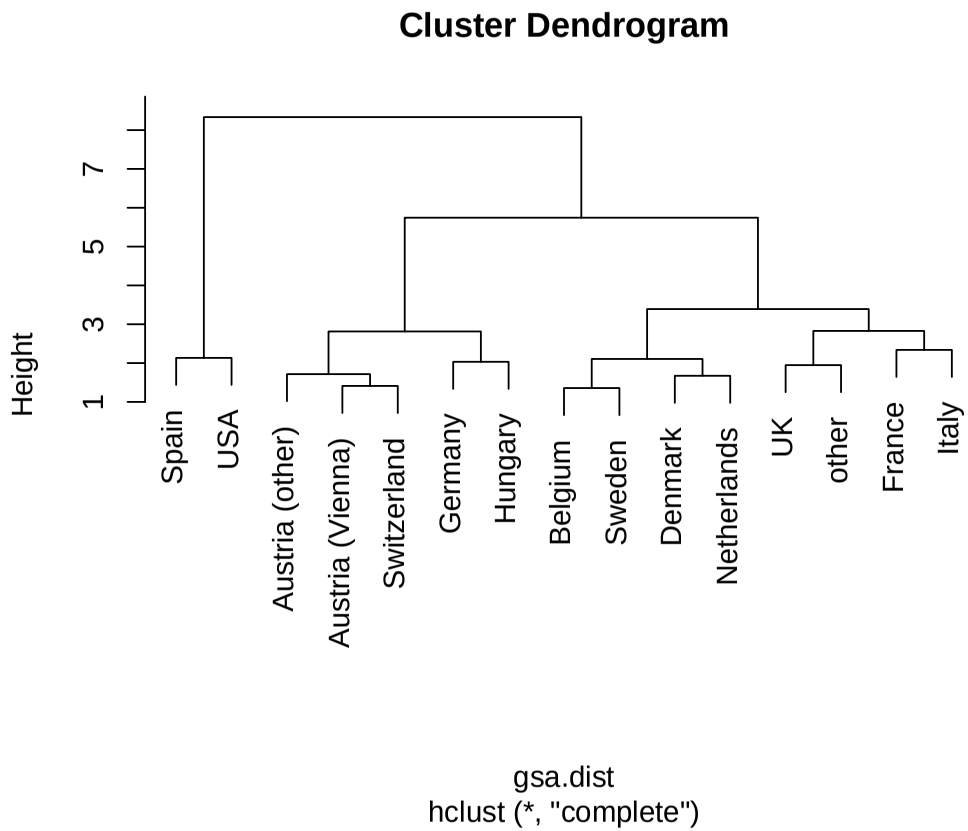
# Hierarchisches Clustern

- **Single linkage** verwendet eine 'Freunde von Freunden'-Strategie, um die Cluster zu konstruieren, da ein einziges Objekt zwei ansonsten weit auseinander liegende Cluster verbinden kann. Dies führt oft zu 'Verkettungen' von Clustern.
- **Complete linkage** versucht sehr homogene Cluster zu finden, manchmal 'zu' homogene.
- **Average linkage** ist ein Kompromiß zwischen single und complete linkage.
- Die **Ward**-Methode versucht kompakte sphärische Cluster zu finden.

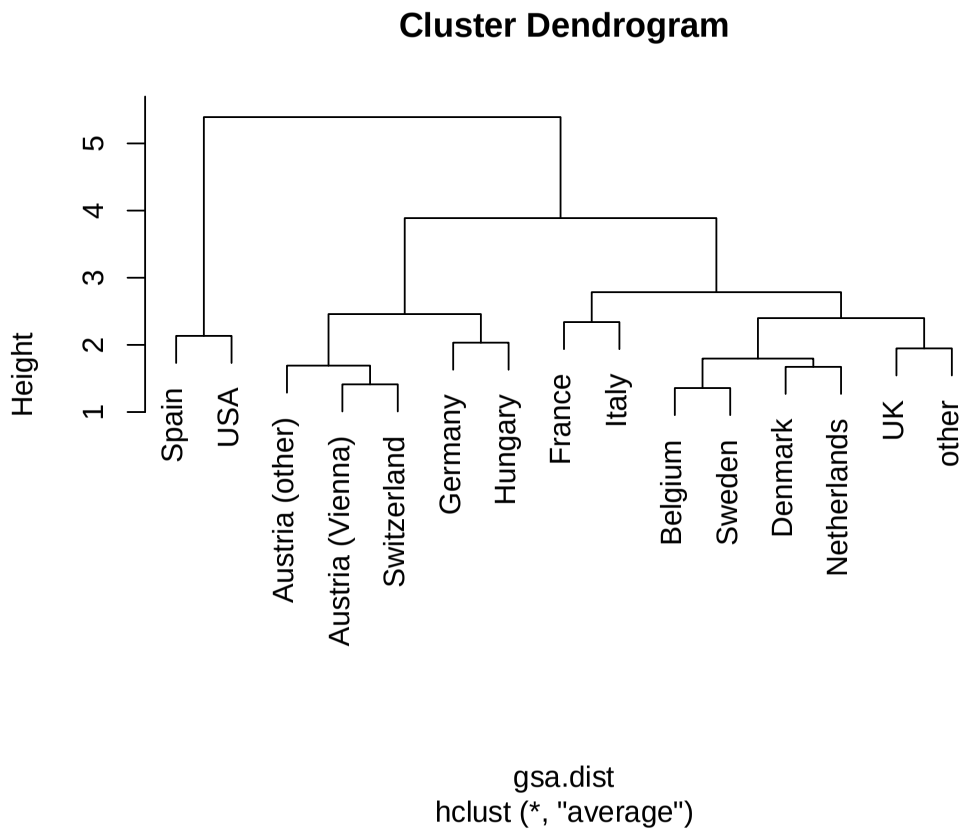
# Hierarchisches Clustern



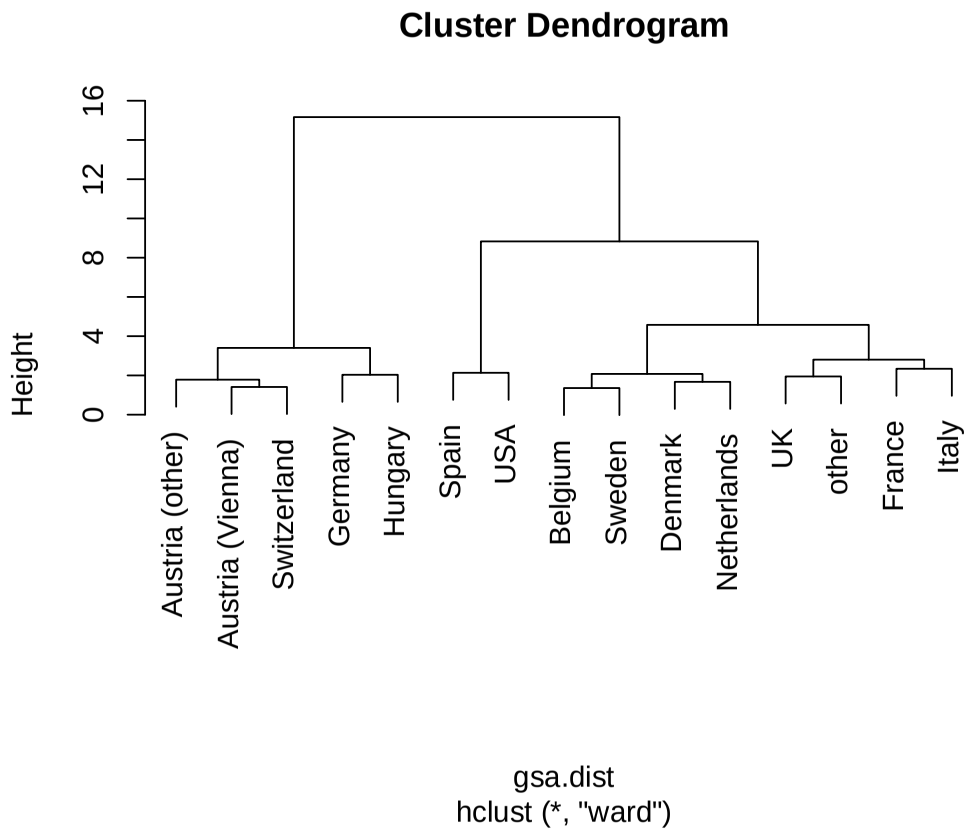
# Hierarchisches Clustern



# Hierarchisches Clustern



# Hierarchisches Clustern



# Hierarchisches Clustern

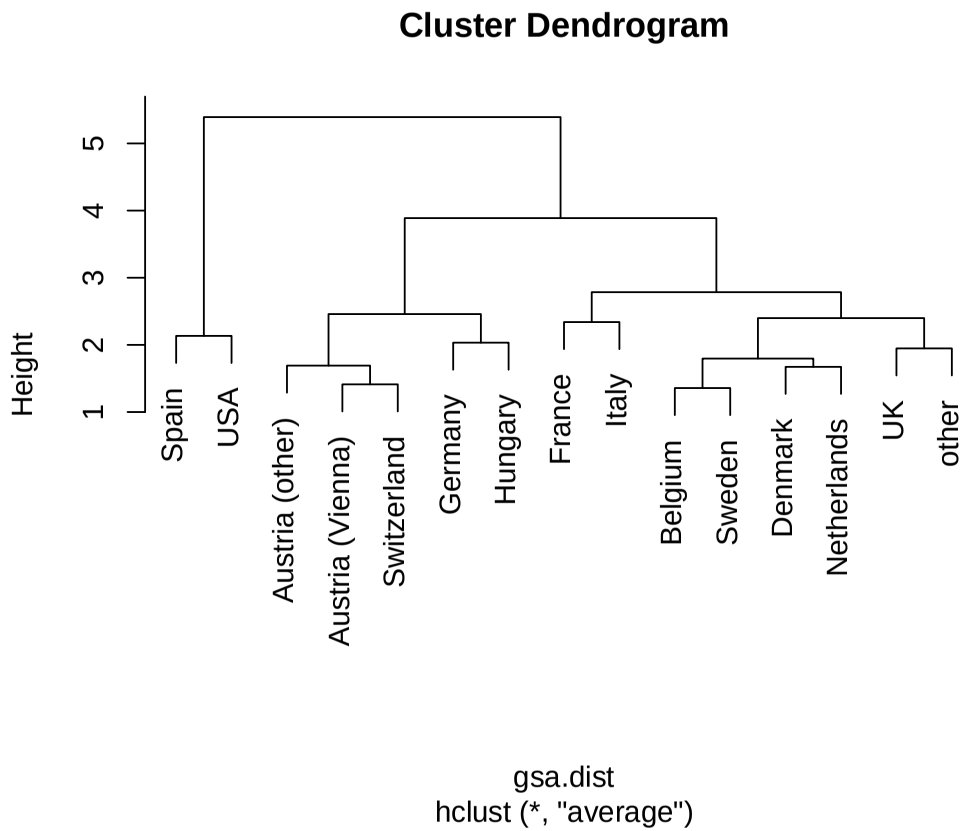
Das Dendrogramm einer Hierarchie von Partitionen visualisiert als Baum in welcher Reihenfolge die Cluster zusammengelegt werden. Die Höhe gibt dabei an, ab welcher Distanz zwei Cluster zusammengefaßt werden.

Je höher also der Schritt zur nächsten Zusammenlegung ist, desto unähnlicher sind die Cluster die zusammengelegt werden.

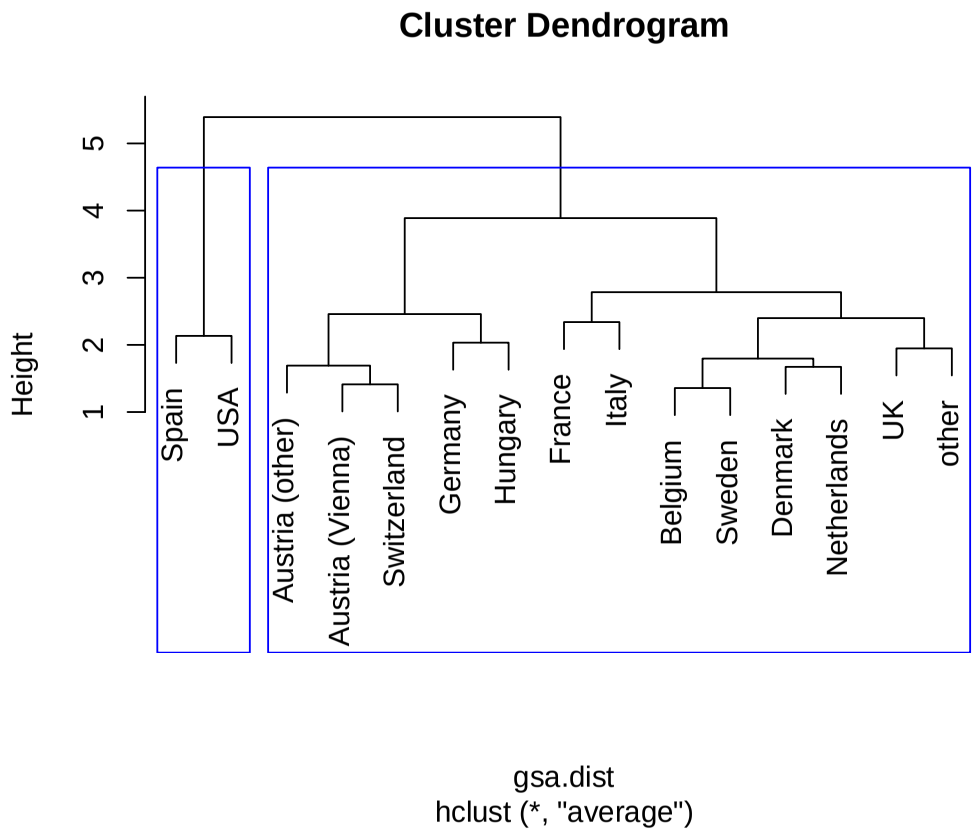
Typischerweise sind die zu überschreitenden Höhen zunächst klein und werden dann immer größer. Man hört in der Regel dann auf, Cluster zusammenzufassen, wenn die Distanzen "zu groß" werden.

Ein "Zerschneiden" des Baumes auf einer bestimmten Höhe ergibt immer eine Partition der Daten.

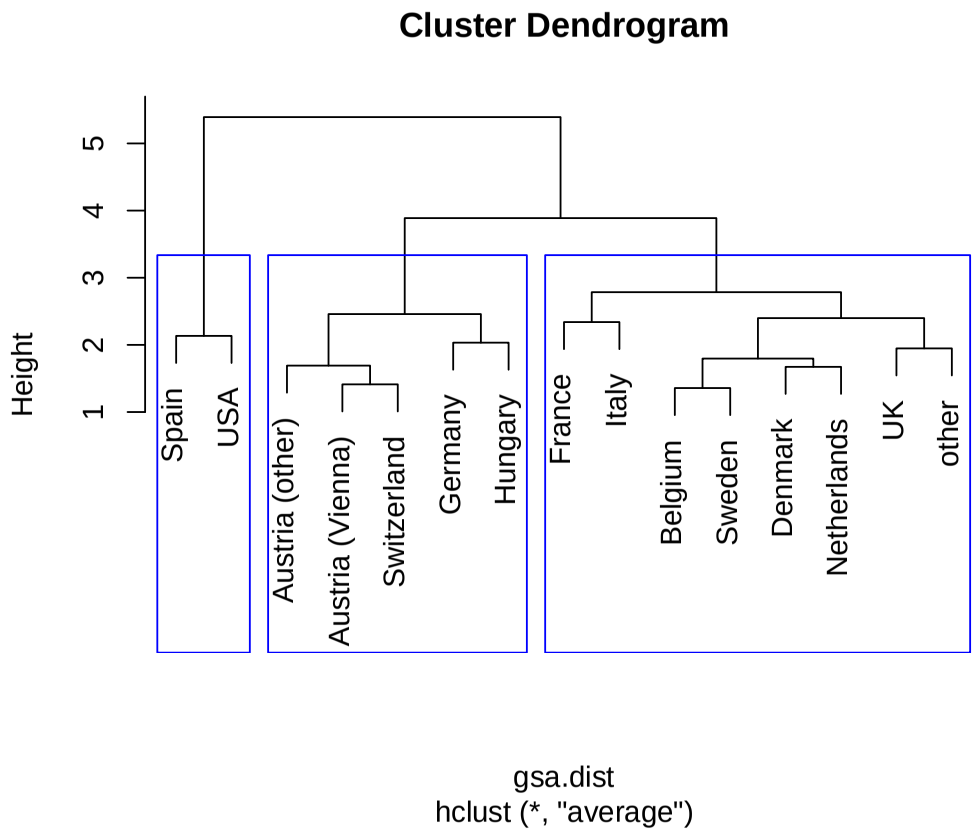
# Hierarchisches Clustern



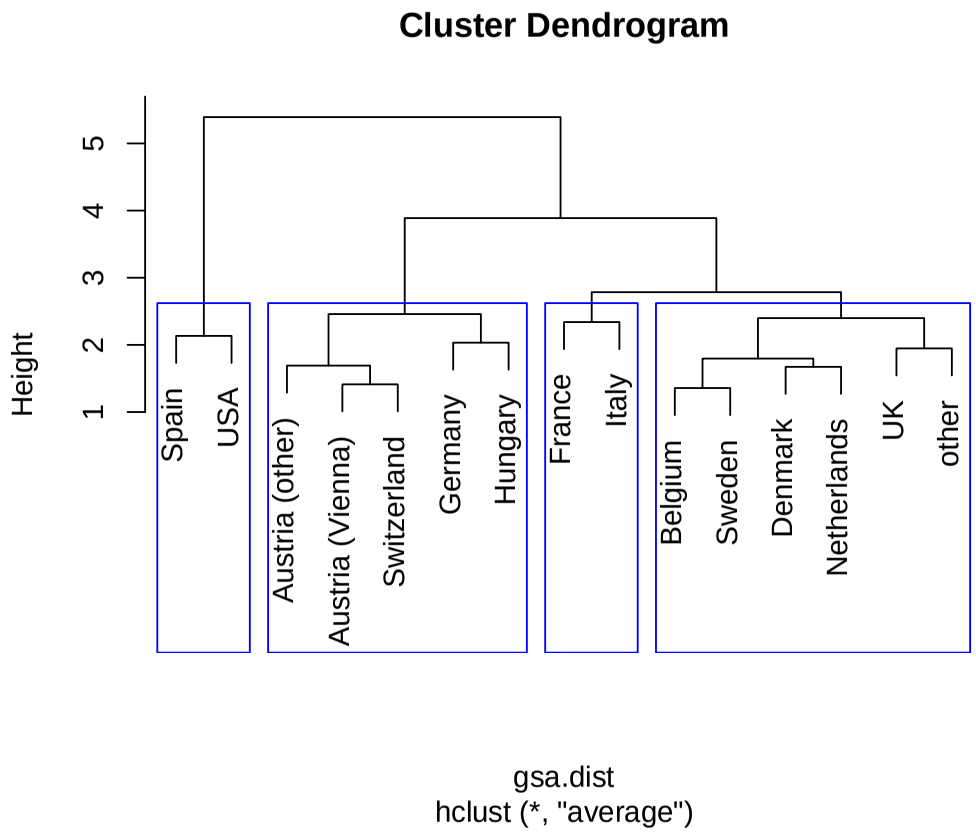
# Hierarchisches Clustern



# Hierarchisches Clustern



# Hierarchisches Clustern



# Tutorium

Hierarchisches Clustern in R (*HClust.pdf*)

# Übung

**Aufgabe 15:**

Clustern Sie die `Ice` Daten (nach Skalierung) hierarchisch.

- Welche Distanzmethode halten Sie für die geeignetste?
- Wie viele Cluster würden Sie wählen?
- Welche der Eigenschaften unterscheidet sich besonders stark zwischen den Clustern?

# Übung

**Aufgabe 16:**

Clustern Sie die `SwissBank` Daten (ohne Skalierung) hierarchisch.

- Welche Distanzmethode halten Sie für die geeignetste?
- Wie viele Cluster würden Sie wählen?
- Welche der Eigenschaften unterscheidet sich besonders stark zwischen den Clustern?

# Übung

**Aufgabe 17:**

Aggregieren Sie die `GSA` Daten nach dem Zielbundesland `province` (anstatt nach dem Herkunftsland) für die Motivationsvariablen 01, 02, 06, 07, 08, 09 (anstatt der Sommeraktivitäten).

- Visualisieren Sie die Daten mit Hilfe von Sternen.
- Führen Sie eine Hauptkomponentenanalyse durch.
- Clustern Sie die Daten hierarchisch.

Wie und in welchen Eigenschaften bzgl. der Motivationen unterscheiden sich die Zielbundesländer?

# $k$-Means

Neben Algorithmen, die Hierarchien von Partitionen berechnen, gibt es auch Algorithmen, die direkt nur Partitionen für eine bestimmte Zahl $k$ von Clustern eine Partition berechnet.

**Frage:** Welche aller möglichen Partitionen von $n$ Objekten in $k$ Cluster, soll aber gewählt werden?

**Antwort:** Benutze eine Zielfunktion und wähle die Partition, die die Zielfunktion optimiert.

# $k$-Means

**Frage:** Was ist eine geeignete Zielfunktion?

**Antwort:** Eine mögliche Zielfunktion, die gerade bei Verwendung euklidischer Distanzen intuitiv ist, ist die Fehlerquadratsumme $SS$. Dies ist die Summe der quadratischen euklidischen Distanzen der Beobachtungen von ihrem Cluster-Mittelwert.

Mittelwert in Cluster $j$:

$$\bar{x}_j \quad = \quad \frac{1}{|C_j|} \sum_{x \in C_j} x$$

# $k$-Means

Fehlerquadratsumme in Cluster $j$ und gesamte Fehlerquadratsumme:

$$\begin{aligned} WSS(C_j) &= \sum_{x \in C_j} d_2(x, \bar{x}_j)^2 \\ SS(\mathcal{C}) &= \sum_{j=1}^{k} WSS(C_j) \end{aligned}$$

Damit hat man also die Partition $\mathcal{C}$ durch $k$ Mittelwerte $\bar{x}_1, \ldots, \bar{x}_k$ repräsentiert. Diese nennt man auch Prototypen. Die Partition $\mathcal{C}$ (oder äquivalent: die zugehörigen $k$ Mittelwerte), die

## $k$-Means

die Fehlerquadratsumme $SS(\mathcal{C})$ minimieren, nennt man $k$-Means-Partition.

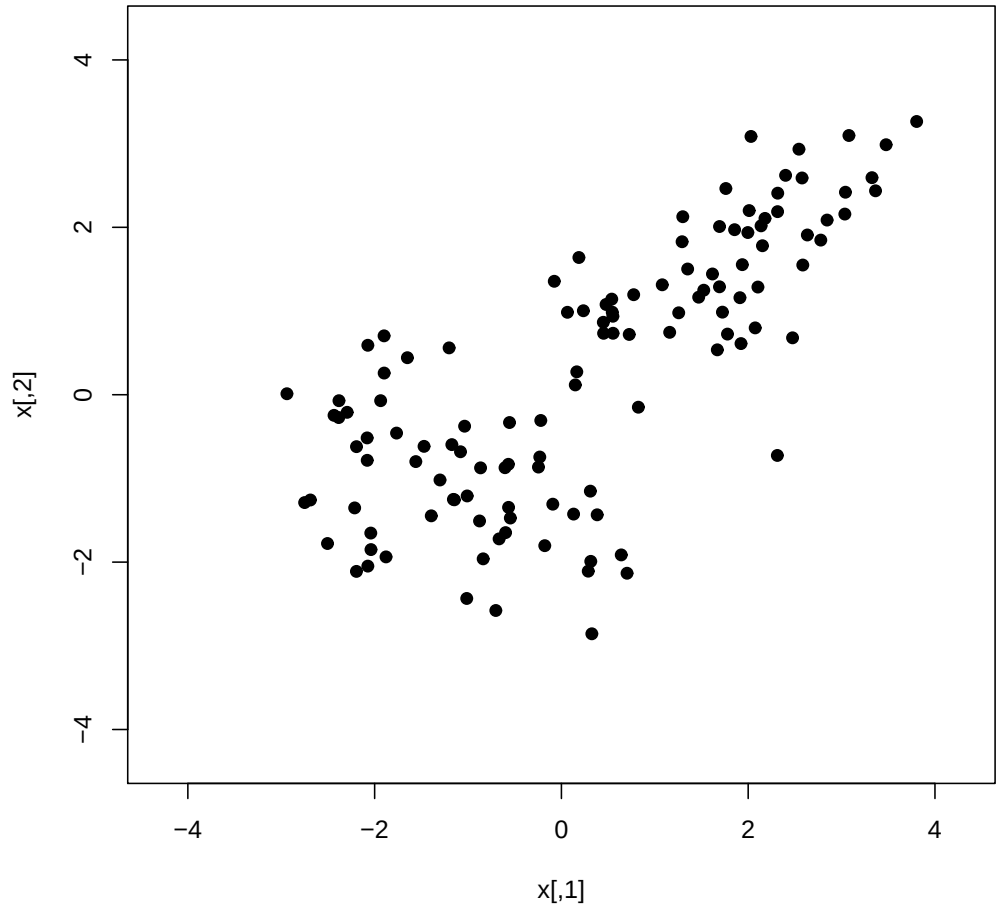
# $k$-Means

Es gibt verschiedene Algorithmen, die eine approximative Lösung für das $k$-Means-Problem berechnen. Der bekannteste ist:

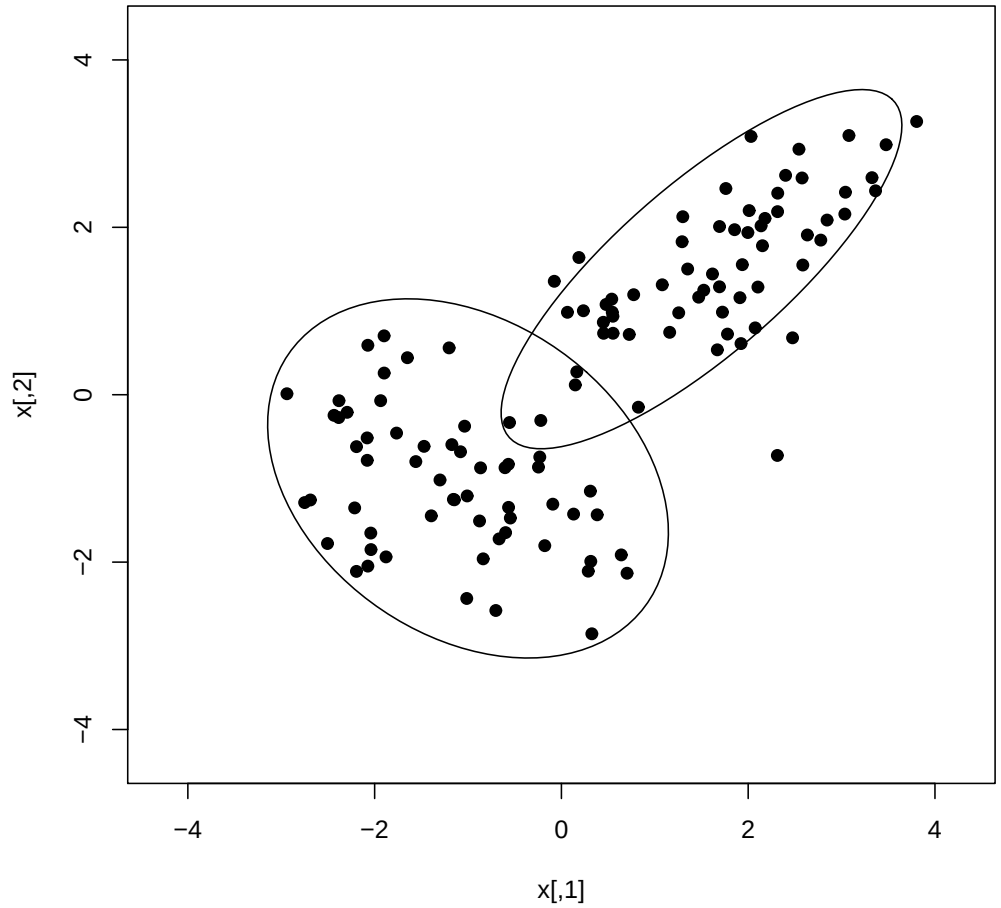
1. Beginne mit $k$ (zufälligen) Mittelwerten $\bar{x}_j$.
2. Ordne jeden Punkt $x_i$ dem Cluster $j$ zu, zu dessem Mittelwerte $\bar{x}_j$ er am nächsten liegt.
3. Berechne die neuen Mittelwerte $\bar{x}_j$ als Mittelwerte der Cluster $j$.
4. Wiederhole 2. und 3. bis sich die Cluster nicht mehr ändern.

**Problem:** Dieser Algorithmus findet nur ein lokales Minimum von $SS(\mathcal{C})$ und nicht das globale Minimum.

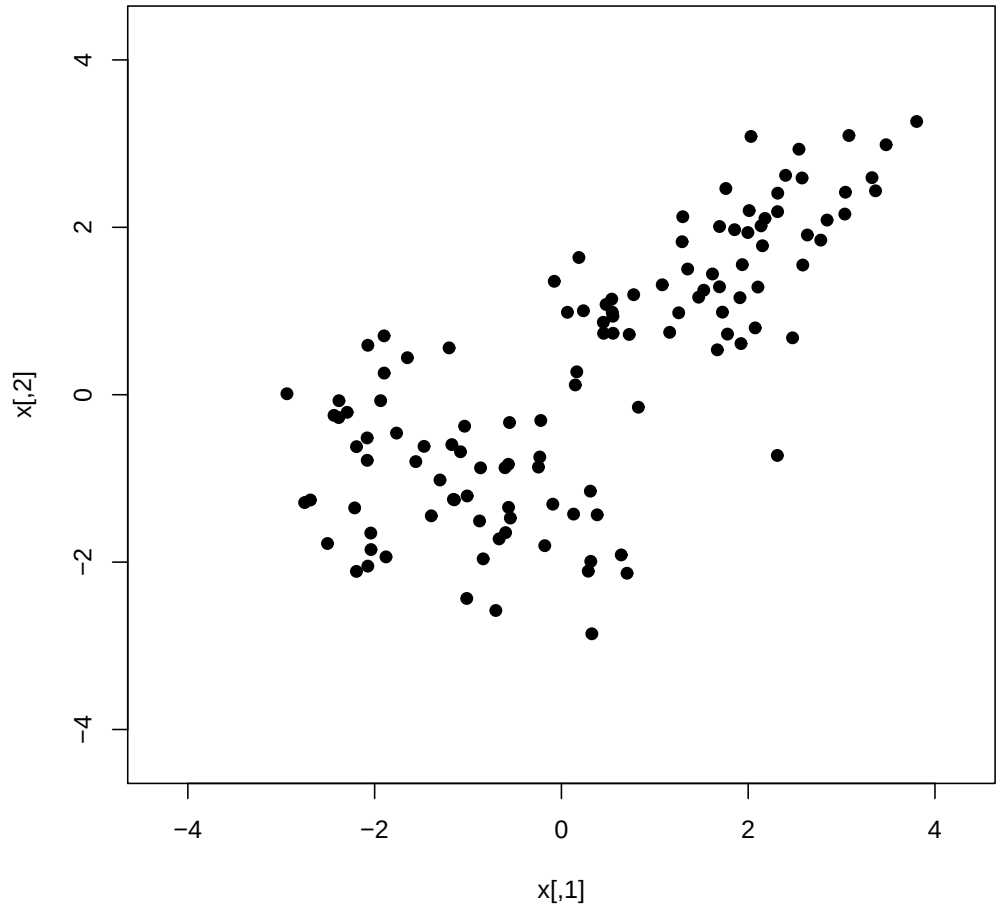
# $k$-Means



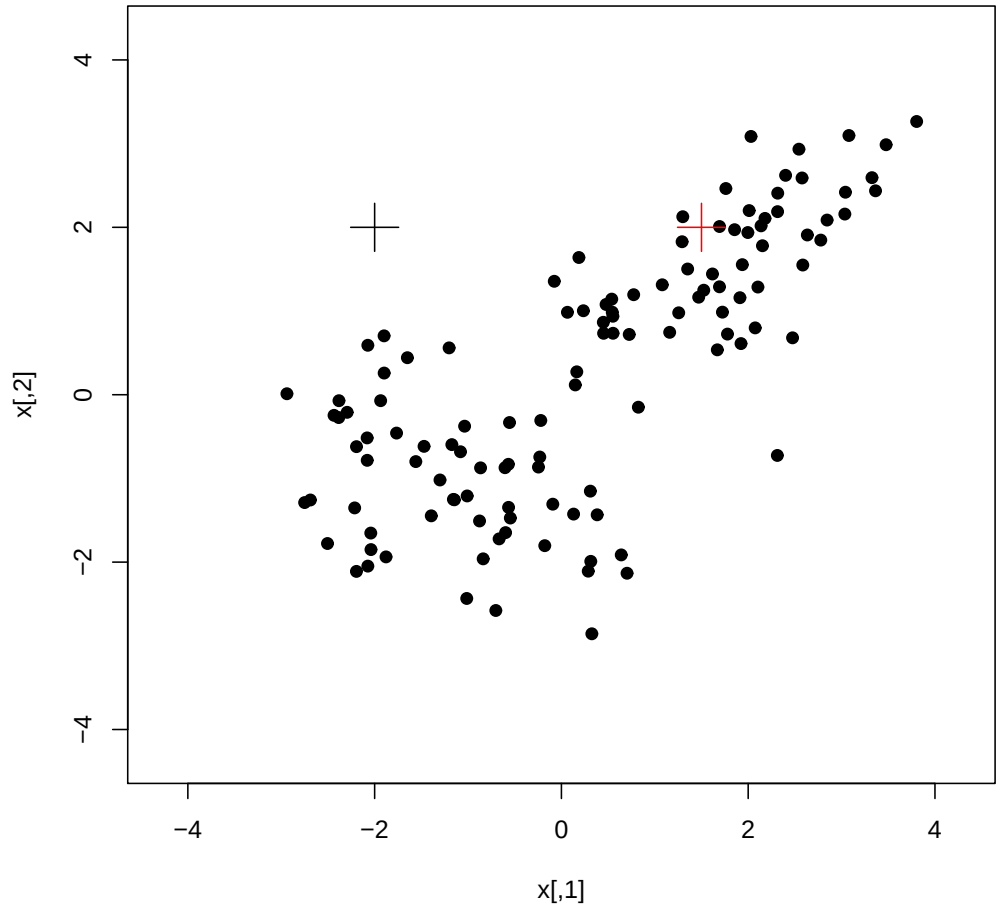
# $k$-Means



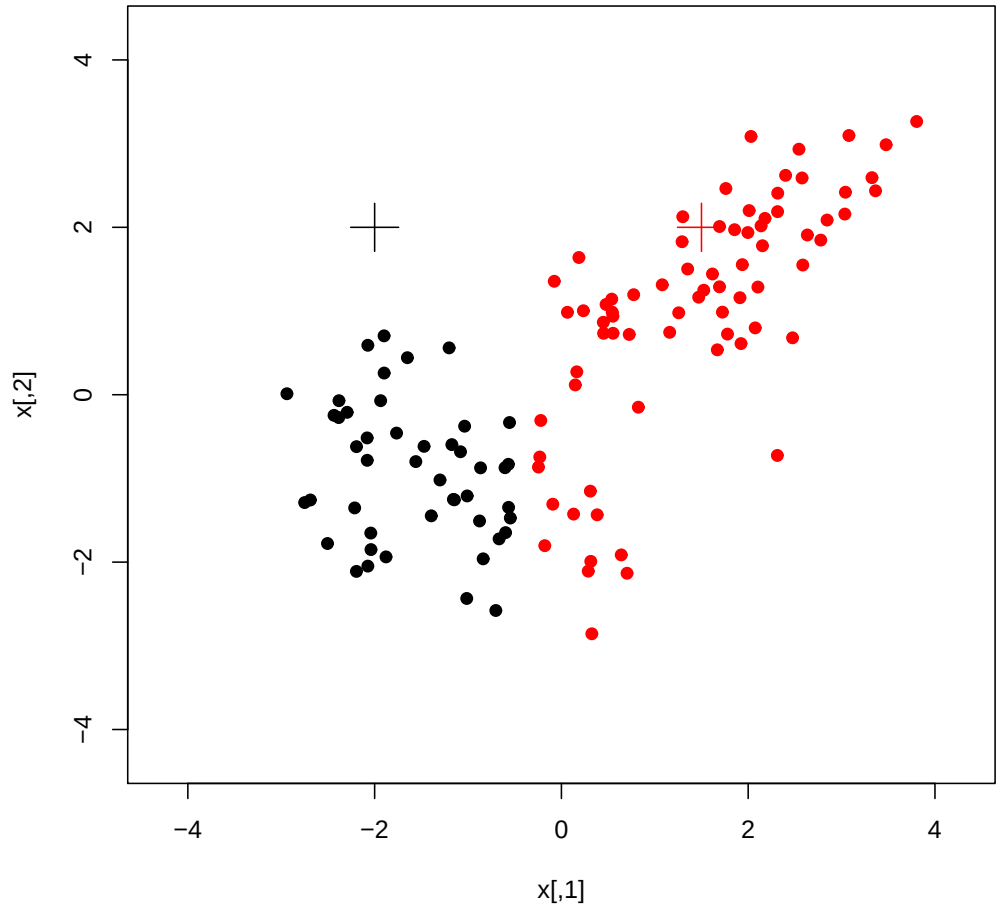
# $k$-Means



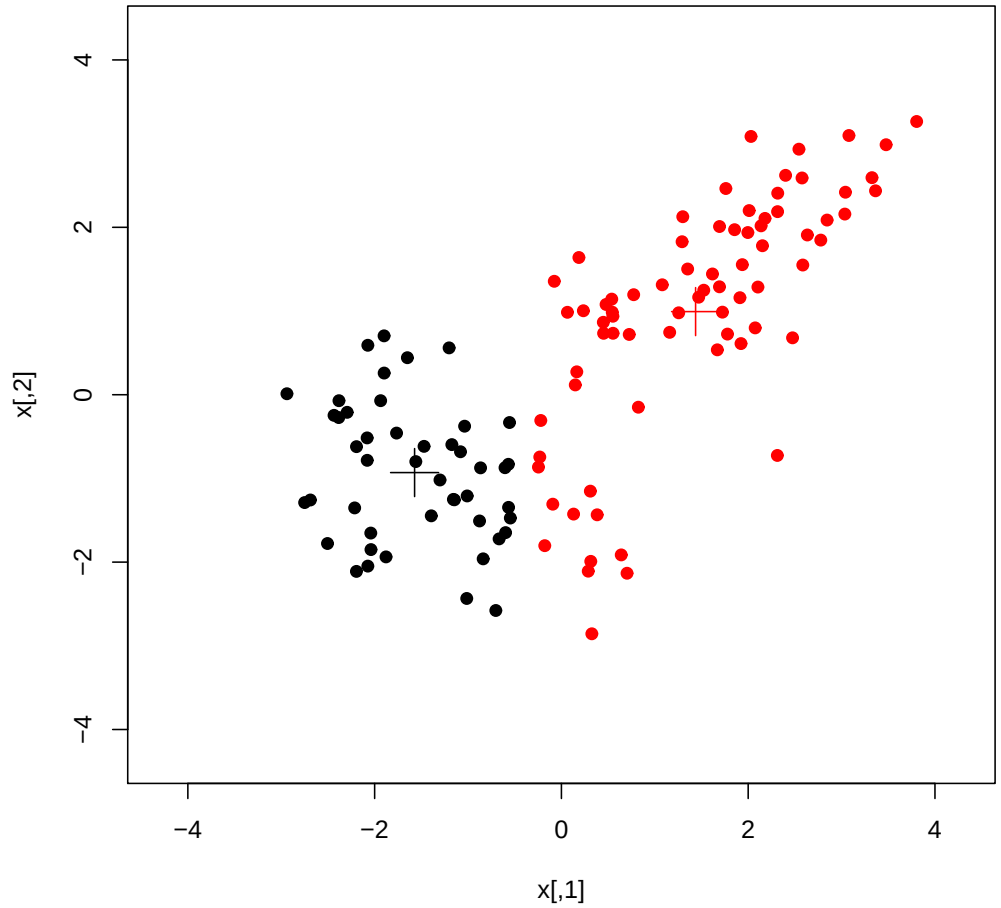
# $k$-Means



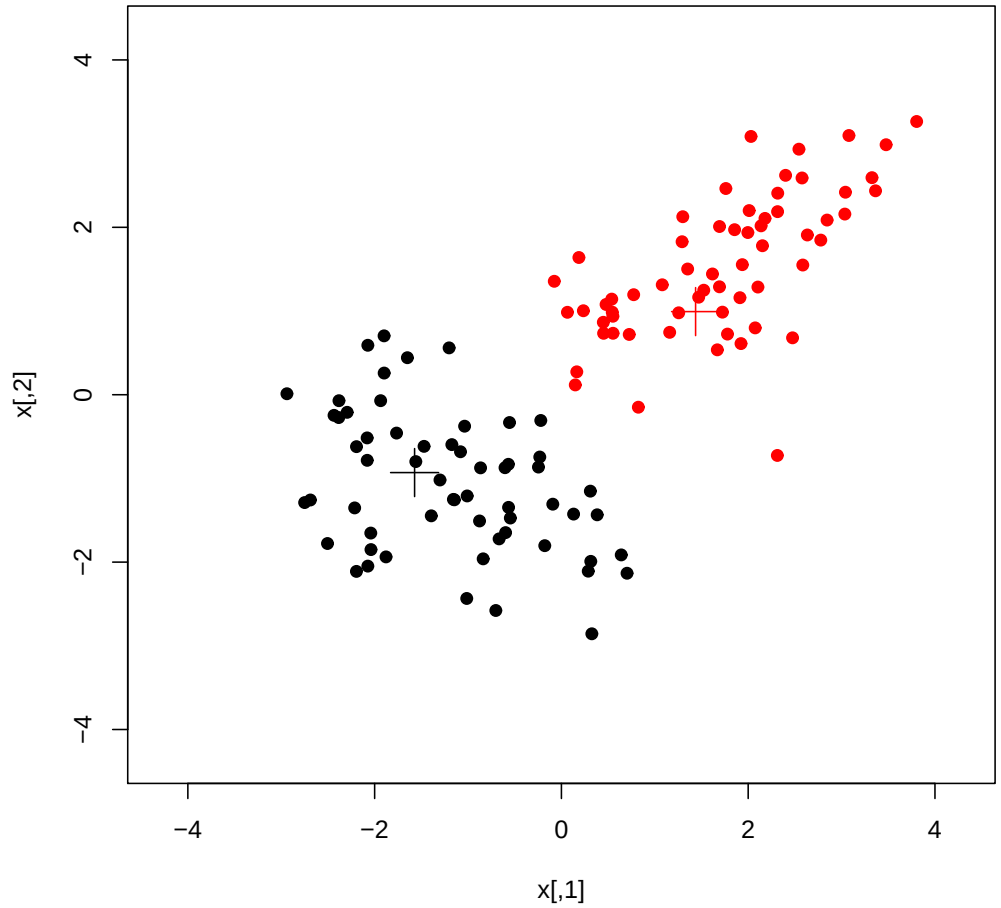
# $k$-Means



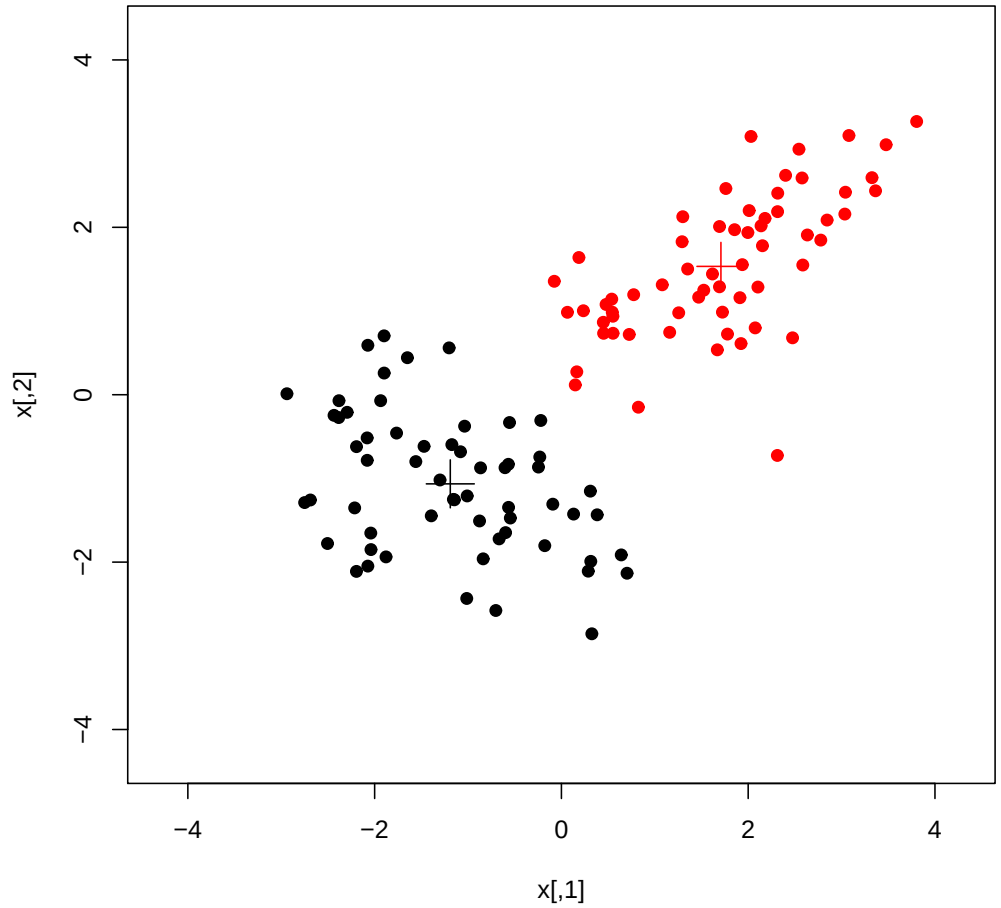
# $k$-Means



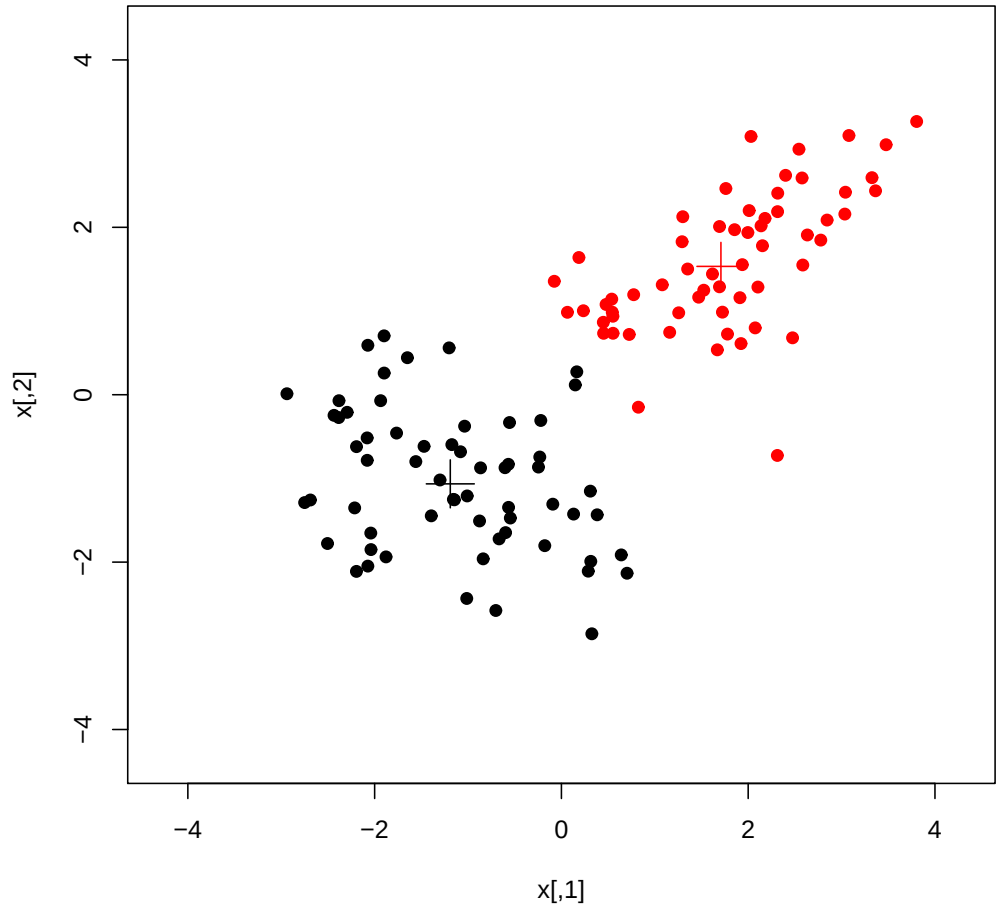
# $k$-Means



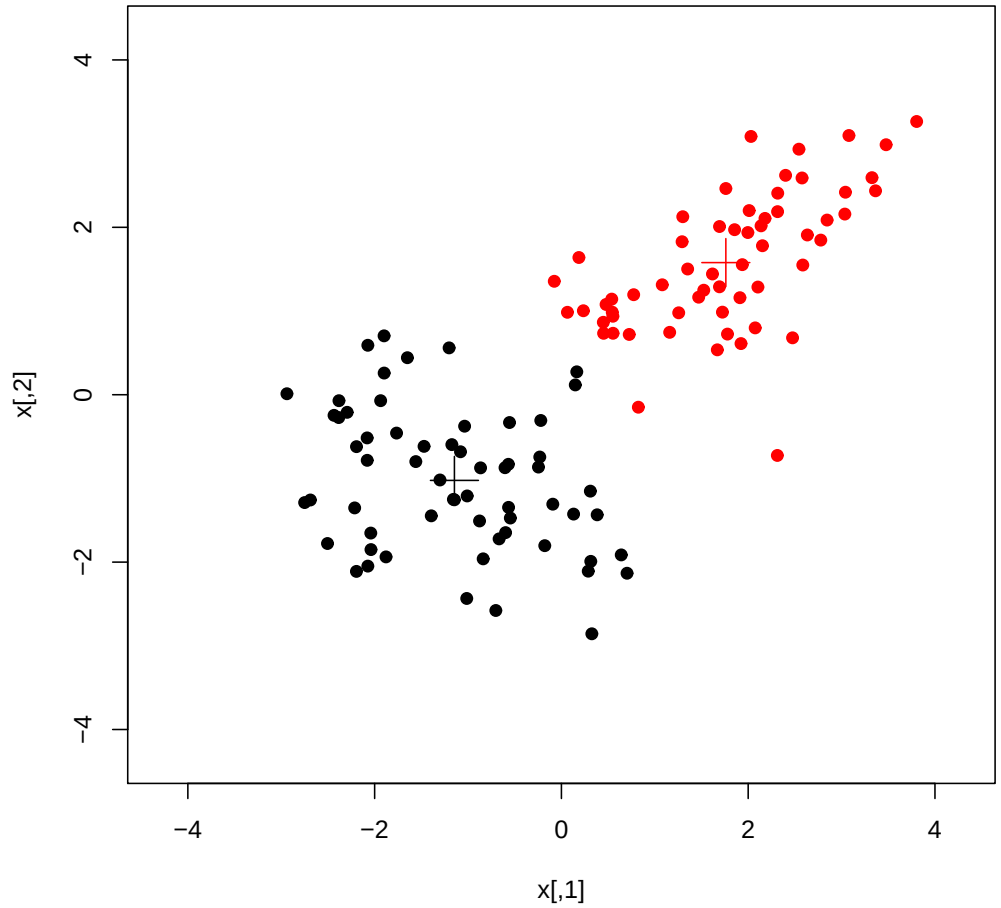
# $k$-Means



# $k$-Means



# $k$-Means

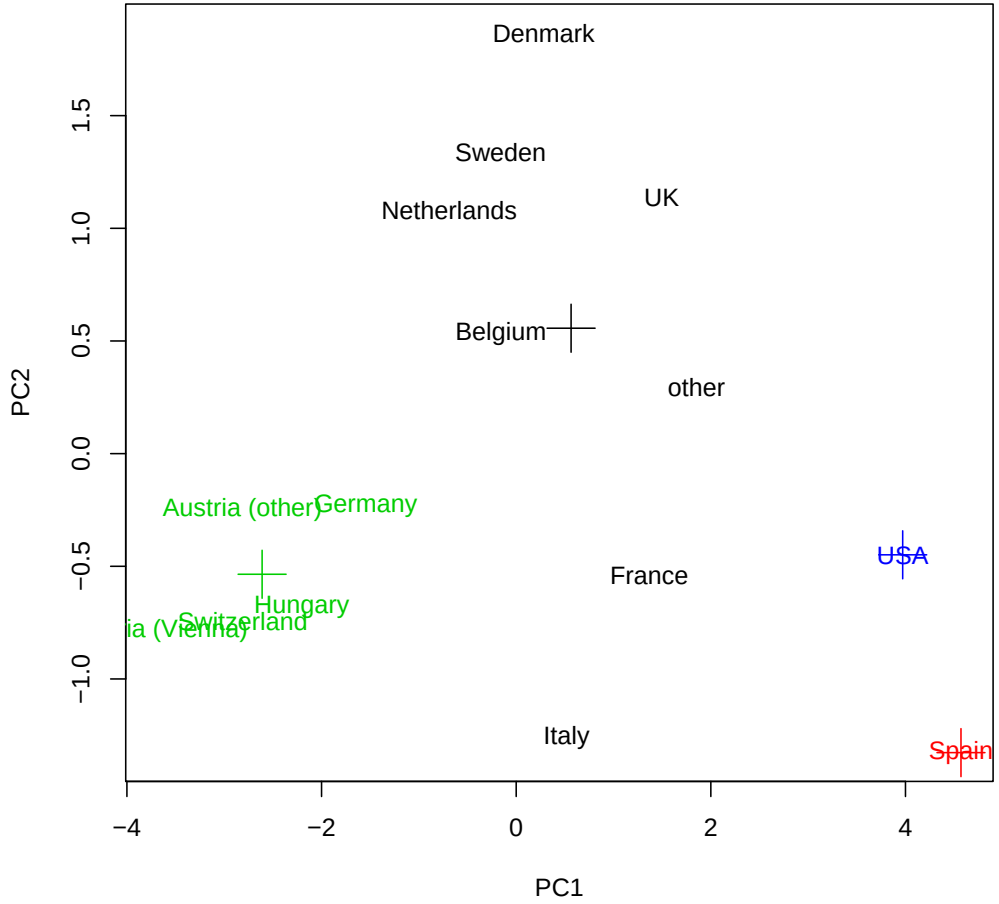


# $k$-Means

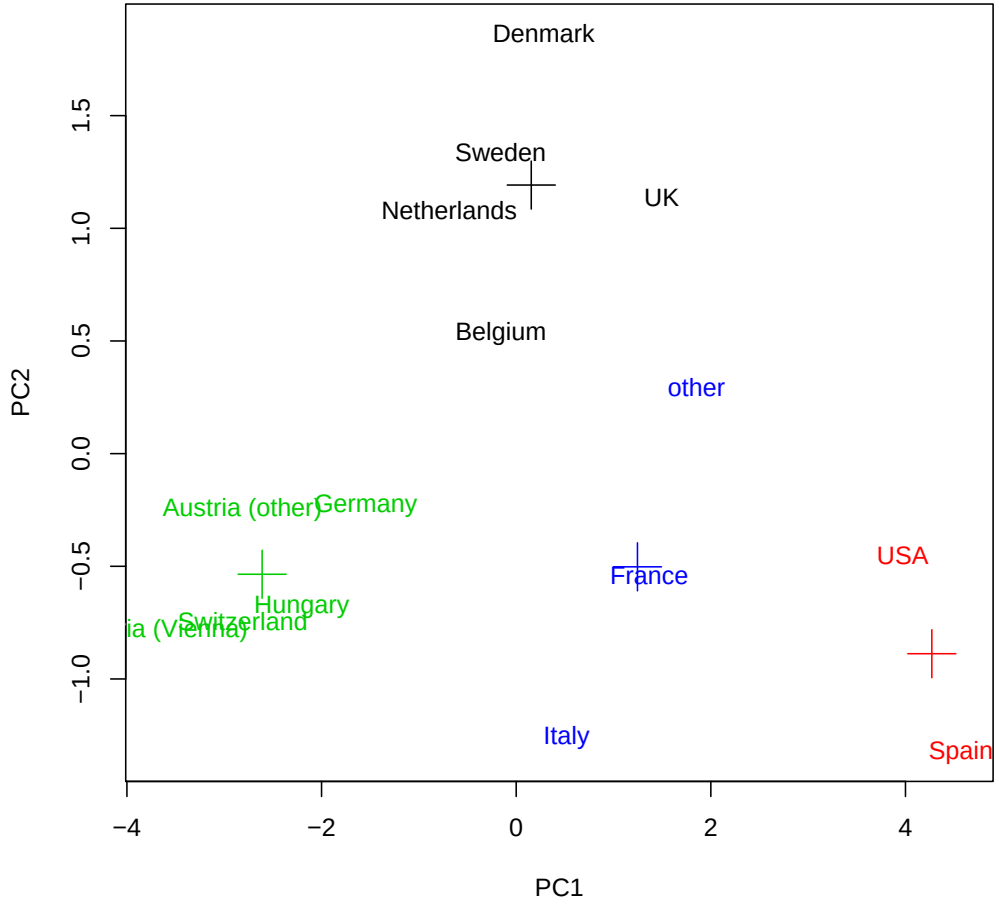
**Beispiel:** Zwei verschiedene Läufe von $k$-Means mit unterschiedlichen Startwerten auf den skalierten `GSA` Daten.

**Bemerkung:** Die Visualisierung der Partition verwendet die ersten beiden Hauptkomponenten. Die Distanzen wurden im $\mathbb{R}^8$ berechnet, werden aber auch in der Projektion in den $\mathbb{R}^2$ noch gut widergespiegelt.

# $k$-Means


Denmark
Sweden
Netherlands
UK
Belgium
other
Austria (other)
Germany
France
USA
Hungary
Switzerland
ia (Vienna)
Italy
Spain
PC2
PC1
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

# $k$-Means


Denmark
Sweden
Netherlands
UK
Belgium
other
Austria (other)
Germany
Hungary
Switzerland
ia (Vienna)
France
USA
Italy
Spain
PC2
PC1
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

# $k$-Means

Es sind also aufgrund unterschiedlicher Startwerte, unterschiedliche Partitionen gewählt worden. Mindestens eine der beiden Partitionen hat also nur ein lokales Minimum gefunden.

Hier ist

$$\begin{aligned} SS(\mathcal{C}_1) &= 30.723 \\ SS(\mathcal{C}_2) &= 24.565 \end{aligned}$$

und deshalb ist $\mathcal{C}_2$ vorzuziehen.

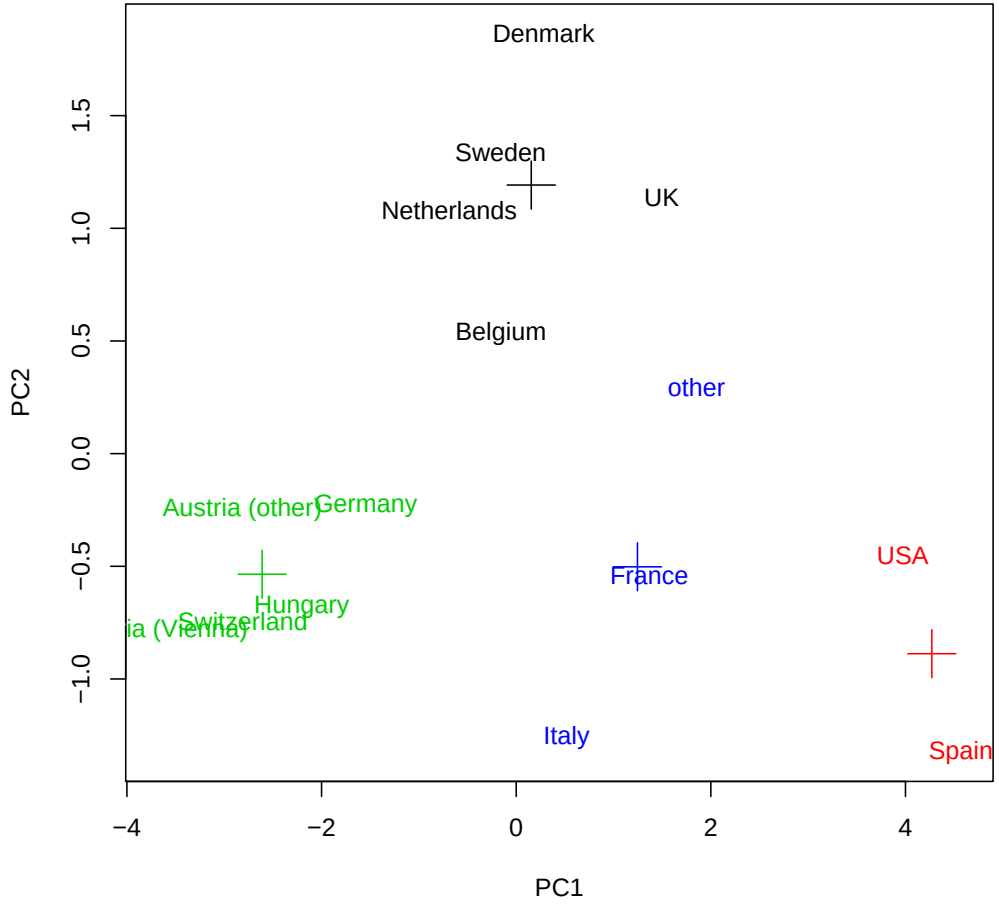
In der Praxis: Berechne 20 Partitionen mit unterschiedlichen Startwerten. Behalte nur die mit der kleinsten Fehlerquadratsumme.

# $k$-Means

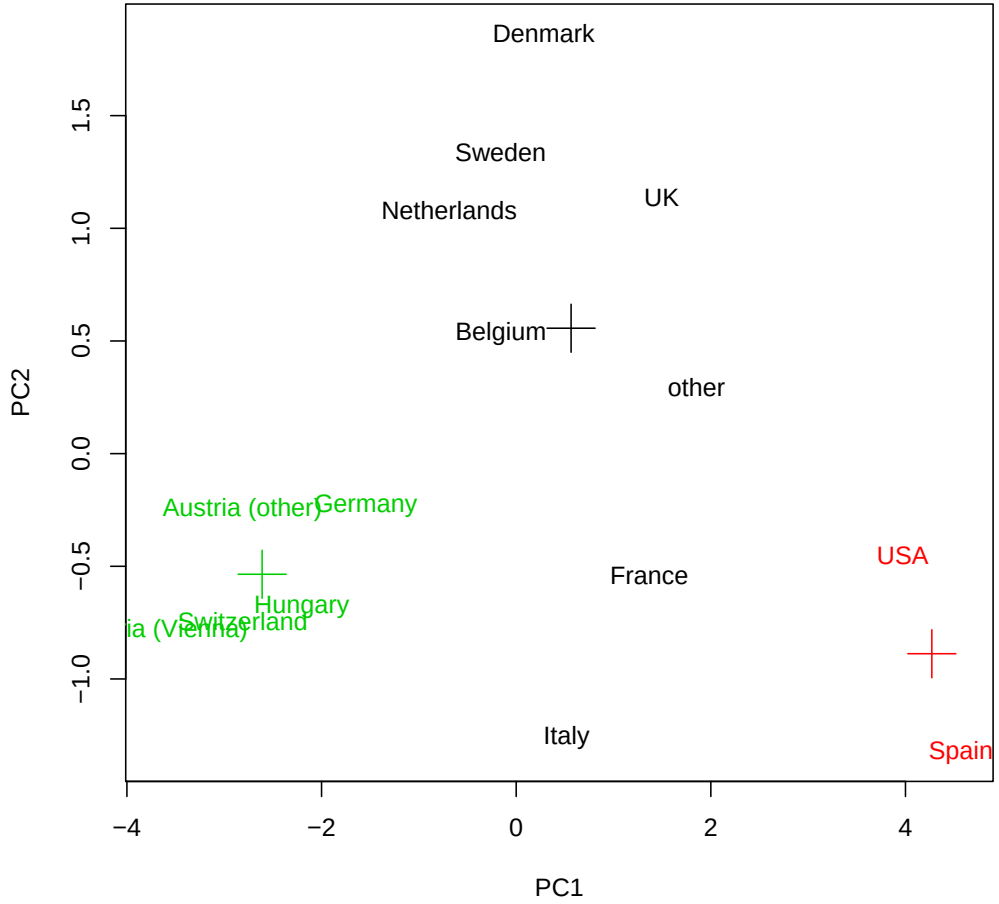
**Frage:** Welche Anzahl $k$ von Clustern soll verwendet werden?

**Antwort:** Berechne für $k = 2, 3, \ldots$ jeweils eine geeignete $k$-Means-Partition und visualisiere die zugehörigen Fehlerquadratsummen. Man hört in der Regel dann auf, Cluster hinzuzufügen, wenn die Verbesserung der Fehlerquadratsumme zu gering wird.

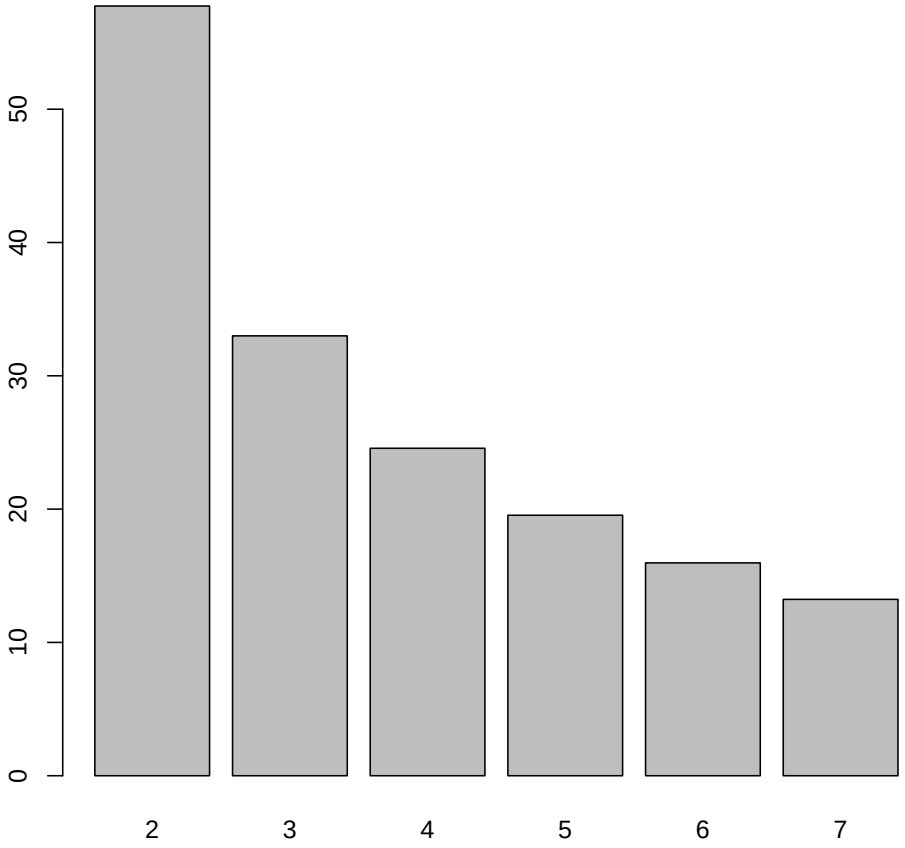
# $k$-Means


Denmark
Sweden
Netherlands
UK
Belgium
other
Austria (other)
Germany
Hungary
Switzerland
ia (Vienna)
France
USA
Italy
Spain
PC2
PC1
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

# $k$-Means


Denmark
Sweden
Netherlands
UK
Belgium
other
Austria (other)
Germany
Hungary
Switzerland
ia (Vienna)
France
USA
Italy
Spain
PC2
PC1
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

# $k$-Means



# $k$-Means

**Alternativ:** Es gibt verschiedene Güte-Indizes, die versuchen einen geeigneten Trade-off zwischen der Anzahl der Cluster $k$, der Heterogenität zwischen den Clustern und der Homogenität innerhalb der Cluster festzulegen.

Die Anzahl der Cluster wird dann durch Optimierung des jeweils verwendeten Güte-Index durchgeführt.

# Tutorium

$k$-Means Clustern in R ($kmeans.pdf$)

# Übung

**Aufgabe 18:**

Clustern Sie die `Ice` Daten (nach Skalierung) mit Hilfe von $k$-Means.

- Wie viele Cluster würden Sie wählen?
- Durch welche Eigenschaften sind die Prototypen charakterisiert?

**Aufgabe 19:**

Clustern Sie die `SwissBank` Daten (ohne Skalierung) mit Hilfe von $k$-Means.

- Wie viele Cluster würden Sie wählen?
- Durch welche Eigenschaften sind die Prototypen charakterisiert?

# Übung

**Aufgabe 20:**

Aggregieren Sie die `GSA` Daten nach dem Zielbundesland `province` (anstatt nach dem Herkunftsland) für die Motivationsvariablen 01, 02, 06, 07, 08, 09 (anstatt der Sommeraktivitäten).

Clustern Sie die resultierenden Daten mit Hilfe von $k$-Means.

- Wie viele Cluster würden Sie wählen?
- Durch welche Eigenschaften sind die Prototypen charakterisiert?

# Verwandte Methoden

**Hauptkomponentenanalyse:**

- Independent component analysis (ICA),
- Faktoranalyse.

**Multidimensionale Skalierung:** nicht-metrische Erweiterungen

- Sammon-Mapping,
- Kruskals MDS,
- Self-organizing maps (SOMs).

# Verwandte Methoden

## Clustern

- divisive hierarchische Algorithmen (DIANA),
- optimale Partitionierung:
  partitioning around medoids (PAM),
  fuzzy clustering (FANNY),
  Neural Gas,
  Convex Clustering,
  Learning Vector Quantization.