C A U
Christian-Albrechts-Universität zu Kiel

12_korrespondenzanalyse

CA, Seriation und (kurz) weiterführende Verfahren

Christian-Albrechts-Universität zu Kiel

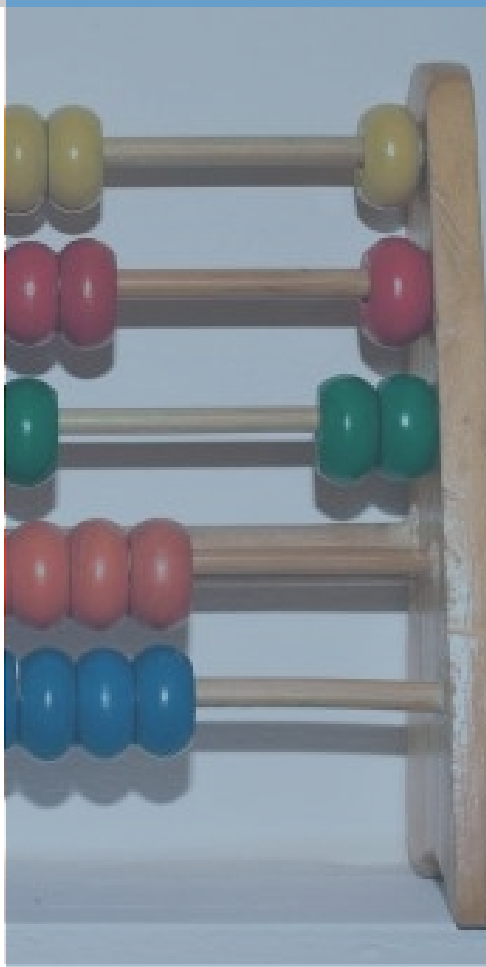
Korrespondenzanalyse: Idee und Grundlagen [1]

Ähnliche Dinge haben ähnliche Merkmale...[2]

Visuelles explorativ/deskriptives Verfahren
Korrespondenzanalyse arbeitet nicht mit Signifikanzen, hat damit keine „Beweiskraft“
Visualisierung von Kontingenztafeln oder Präsenz-Absenz-Matrizen

Idee
Darstellung von Merkmalsträgern (sites) und Merkmalen (species) in einem gemeinsamen Raum (Koordinatensystem)
Daten, die miteinander in Beziehung stehen, werden dabei näher beieinander dargestellt
Ähnlichkeiten werden über Chi-Quadrat-Verfahren berechnet

Voraussetzungen
Eine Datenmatrix mit mindestens nominal skalierten Variablen, daher besonders für archäologische Fragestellungen geeignet

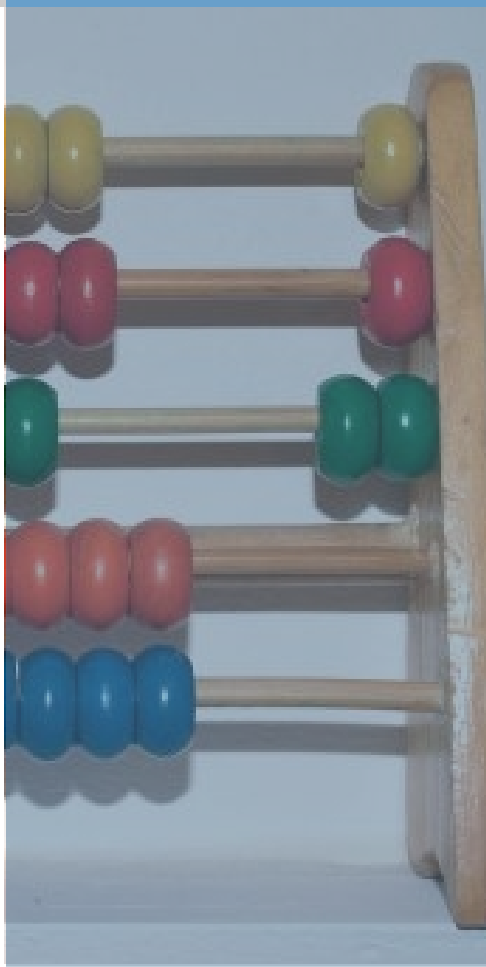


C | A | U
Christian-Albrechts-Universität zu Kiel

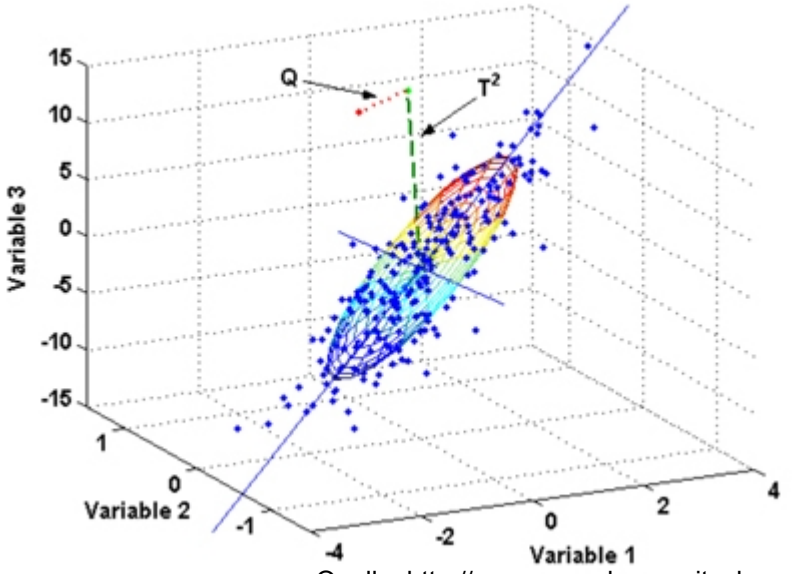
Korrespondenzanalyse: Idee und Grundlagen [2]

Ähnliche Dinge haben ähnliche Merkmale...

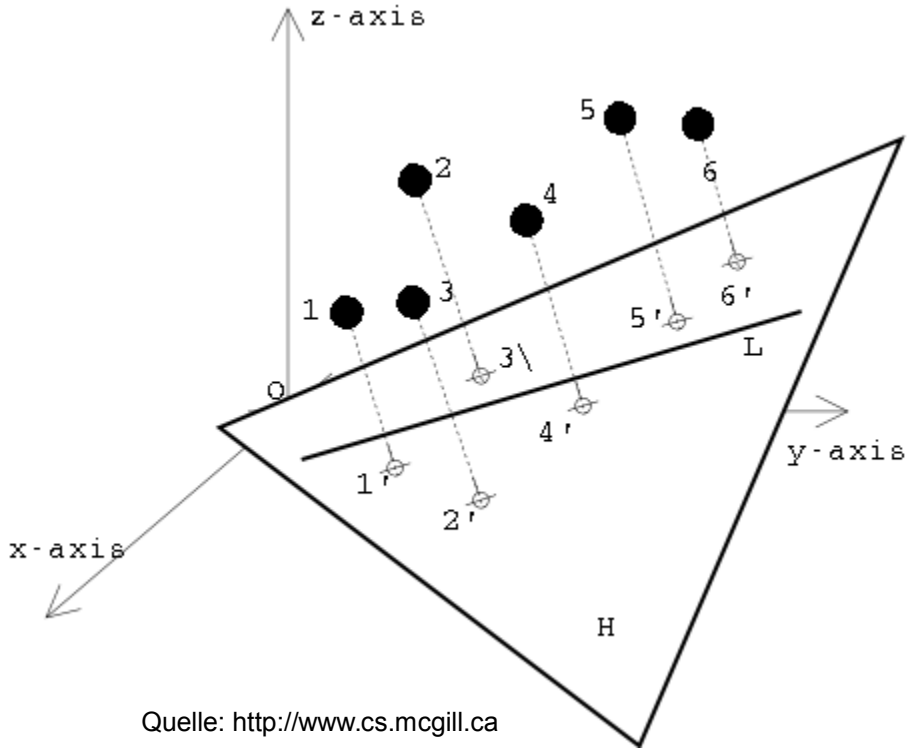
Generelles Vorgehen
Standardisieren der Daten auf ein vergleichbares Maß
„Projektion“ der Daten in einen multidimensionalen Variablenraum
Ermitteln der Vektoren, die stufenweise die meiste Information (Variabilität) in den Daten aufnehmen und senkrecht zueinander orientiert sind
„Projizieren“ der Daten auf diese Vektoren
Darstellung der Lage der Daten auf diesen Vektoren in einem Diagramm



Grundlegende statistische Verfahren für archäologische Datenanalyse in R



Quelle: http://www.aapspharmscitech.org



Quelle: http://www.cs.mcgill.ca

C | A | U
Christian-Albrechts-Universität zu Kiel

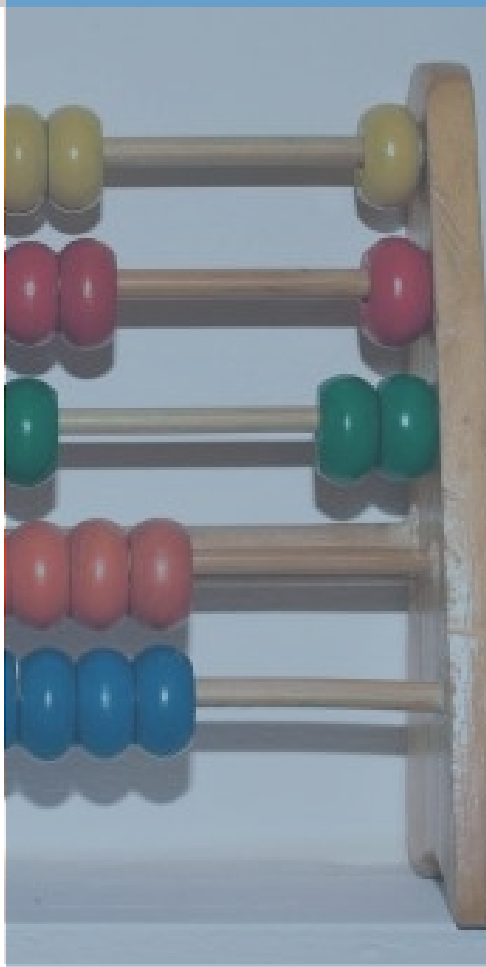
Korrespondenzanalyse: Geschichte

Allgemein

Entwicklung im Umfeld der Biologie und Psychologie
Algebrarische Grundlagen 1940er Jahre (Hartley/Guttman)
Erste explizite Verwendung 1960er Jahre durch Benzéncri in linguistischen Untersuchungen
Weiterentwicklung in verschiedenen Forschergruppen → führte zu verschiedenen Versionen und Benennungen des Verfahrens
1984 Greenacre grundlegende Monographie zum Verfahren

In der Archäologie

Erste Seriation: Sir William Flinders-Petrie 1899
Erste größere Versuche mit seriierenden Verfahren in Deutschland Goldmann 1979 mit „reciprocal averaging“
Weite Anwendung des Verfahrens zur chronologischen Sortierung der rheinländischen Linearbandkeramik
Weiterführung durch Institute Köln und Kiel (Zimmermann, Müller)



C A U
Christian-Albrechts-Universität zu Kiel

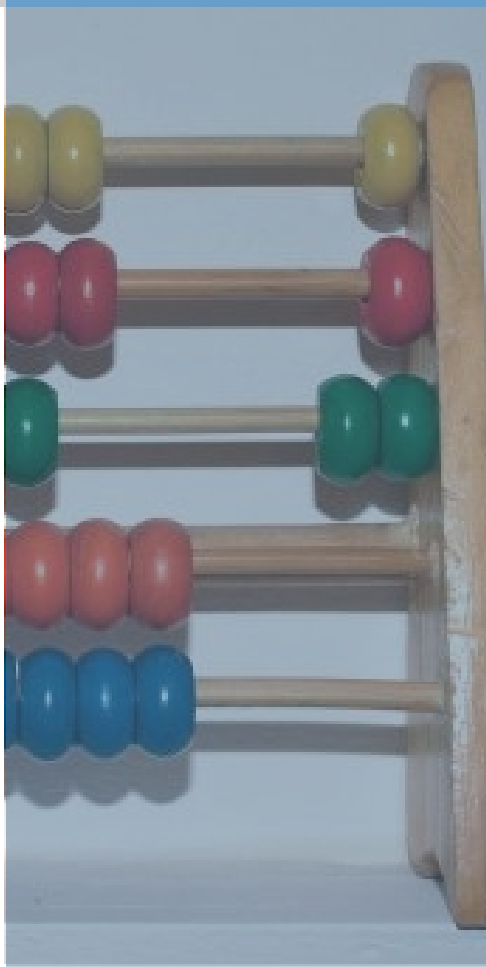
Korrespondenzanalyse: Vorgehen

Vorbereitung: Kontingenztafel, falls nötig

Präsenz-Absenz-Matrix
Notiert das Vorhandensein oder Nichtvorhandensein eines Merkmals für einem Merkmalsträger, in der Archäologie die meistverwendete Ausgangsbasis

	Topf	Tasse	Fibel	
Grab 1	1	1	0	2
Grab 2	0	1	1	2
Grab 3	1	1	1	3
Grab 4	1	0	1	2
Summe	3	3	3	9

Voraussetzung: Summe je Spalte mind. 2, Summe je Zeile mind. 2



Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

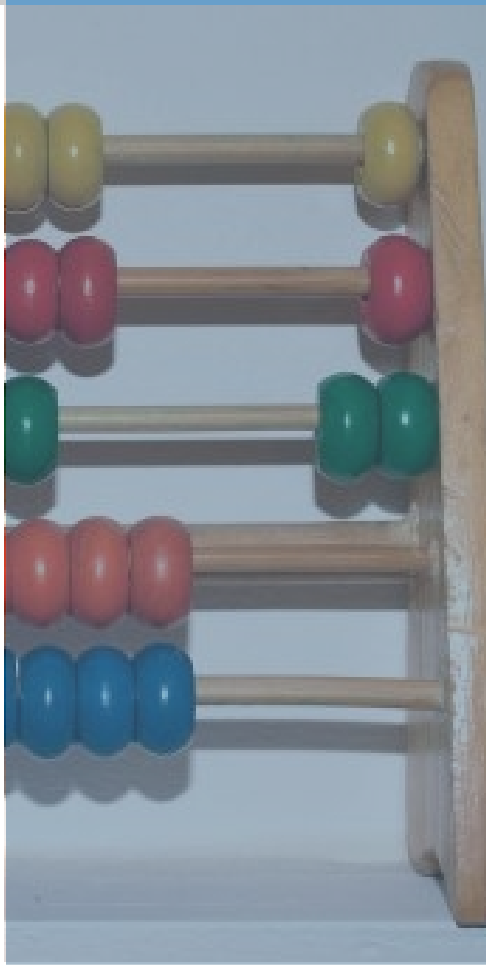
Vorbereitung: Kontingenztafel, falls nötig

Kontingenztafel

Notiert die Anzahl eines Merkmals für einem Merkmalsträger oder eine Gruppe von Merkmalsträgern.

	Topf	Tasse	Fibel
Untergrab	20	23	40
Bodengrab	23	10	6
Obergrab	2	56	4

Außerdem noch möglich: Burt-Matrix, wer will, kann Einzelheiten dazu nach der Sitzung bei mir erfragen...



Grundlegende statistische Verfahren für archäologische Datenanalyse in R

Christian-Albrechts-Universität zu Kiel

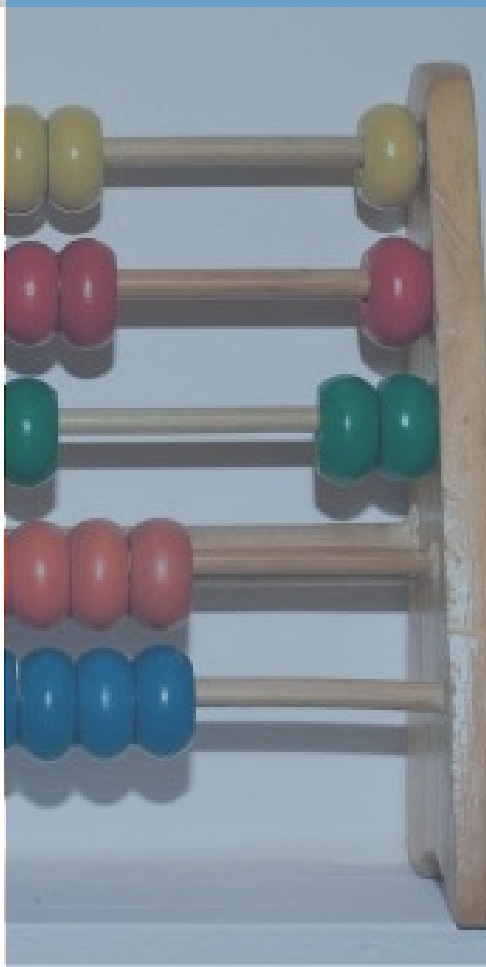
Korrespondenzanalyse: Vorgehen (anhand einer Präsenz-Absenz-Matrix)

Vorbereitung: Spalten- und Reihenprofile

Wieviel Prozent der Gesamtzahl in einer Spalte verteilen sich auf die einzelnen Zellen?
Wieviel Prozent der Gesamtzahl in einer Zeile verteilen sich auf die einzelnen Zellen?

	Topf	Tasse	Fibel	
Grab 1	1	1	0	2
Grab 2	0	1	1	2
Grab 3	1	1	1	3
Grab 4	1	0	1	2
Summe	3	3	3	9

Berechnung: Teilen jeder Zelle durch die Gesamtzahl der jeweiligen Spalte/Zeile.



Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen (anhand einer Präsenz-Absenz-Matrix)

Vorbereitung: Spalten- und Reihenprofile

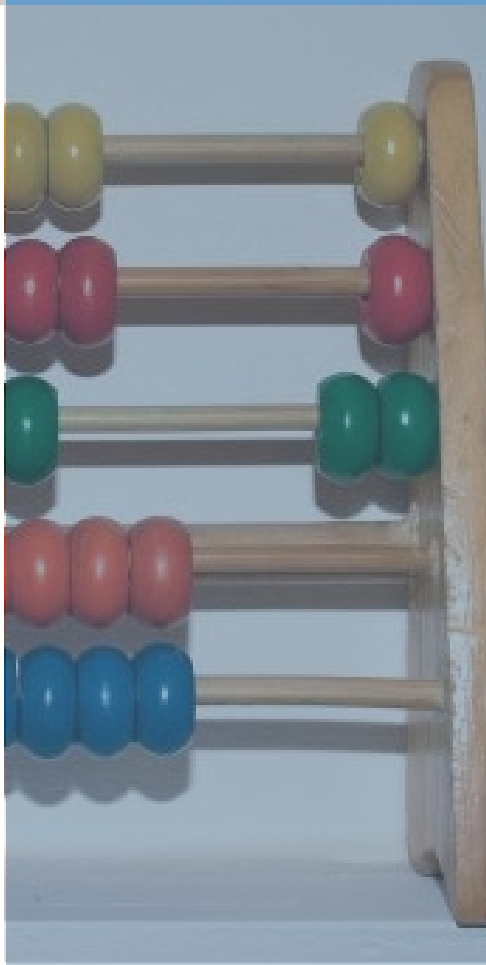
	Topf	Tasse	Fibel	Mittelwert
Grab 1	0,33	0,33	0	0,22
Grab 2	0	0,33	0,33	0,22
Grab 3	0,33	0,33	0,33	0,33
Grab 4	0,33	0	0,33	0,22
Summe	1	1	1	1

Masse

Spaltenprofil

Zeilenprofil

	Topf	Tasse	Fibel	Summe
Grab 1	0,5	0,5	0	1
Grab 2	0	0,5	0,5	1
Grab 3	0,33	0,33	0,33	1
Grab 4	0,5	0	0,5	1
Mittelwert	0,33	0,33	0,33	1



C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

Vorbereitung: Streuung der Daten

Berechnung der Chi-Quadrat-Abweichung

Erwartungswert	Topf	Tasse	Fibel	
Grab 1	0,67	0,67	0,67	2
Grab 2	0,67	0,67	0,67	2
Grab 3	1	1	1	3
Grab 4	0,67	0,67	0,67	2
	3	3	3	9

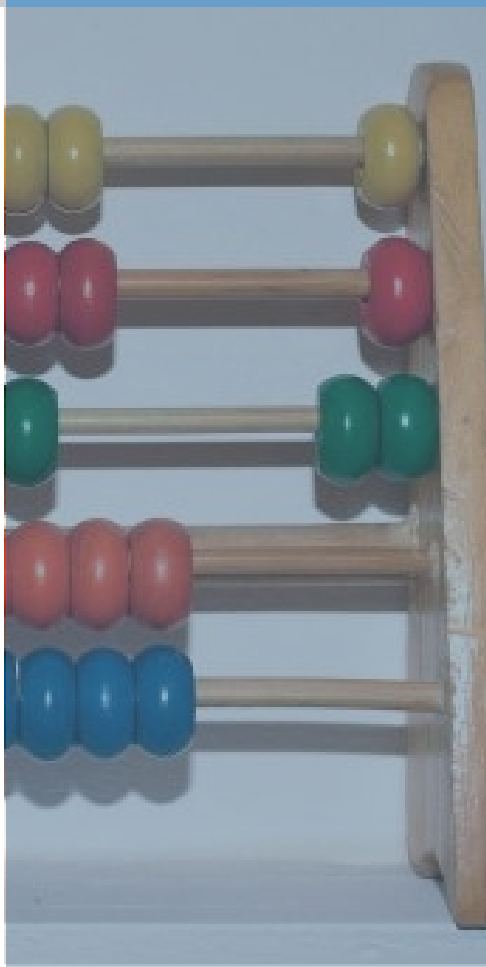
$$\chi^2 = \sum_{i=1}^{k} \frac{(B_i - E_i)^2}{E_i}$$

Chi-Quadrat-Abweichung	Topf	Tasse	Fibel	Summe
Grab 1	0,17	0,17	0,67	1
Grab 2	0,67	0,17	0,17	1
Grab 3	0	0	0	0
Grab 4	0,17	0,67	0,17	1
Summe	1	1	1	3

Totale Inertia:

$$\frac{\chi^2}{n}$$

$$Beispiel: \frac{3}{9} = 0.\overline{3}$$



Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

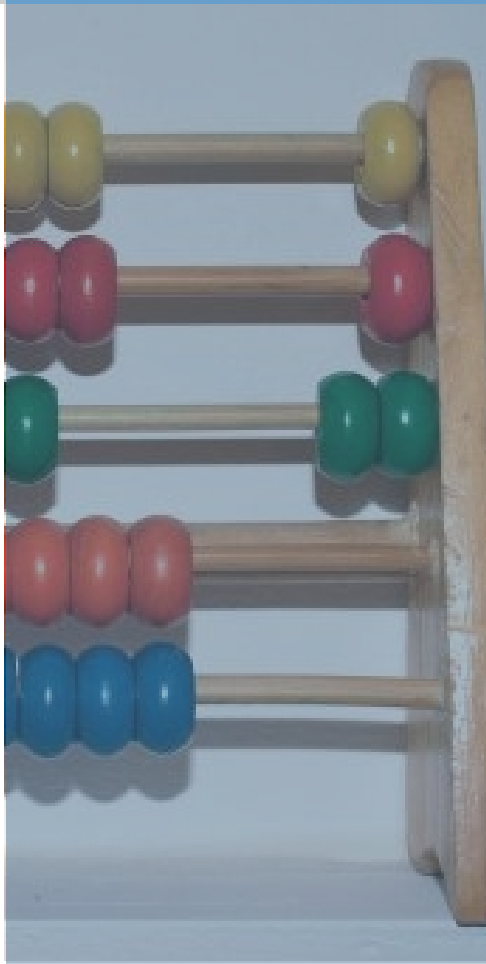
Standardisierung der Daten

Erster Schritt: Erstellung einer Tabelle mit den Relativen Häufigkeiten (Korrespondenztabelle)

	Topf	Tasse	Fibel	
Grab 1	1	1	0	2
Grab 2	0	1	1	2
Grab 3	1	1	1	3
Grab 4	1	0	1	2
Summe	3	3	3	9

Division der einzelnen Zellen durch die Gesamtzahl der Beobachtungen

	Topf	Tasse	Fibel	Summe
Grab 1	0,11	0,11	0	0,22
Grab 2	0	0,11	0,11	0,22
Grab 3	0,11	0,11	0,11	0,33
Grab 4	0,11	0	0,11	0,22
Summe	0,33	0,33	0,33	1



C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

Standardisierung der Daten

Zweiter Schritt: Zentrierung

Erstens: Berechnen der Erwartungswerte für die Korrespondenztabelle

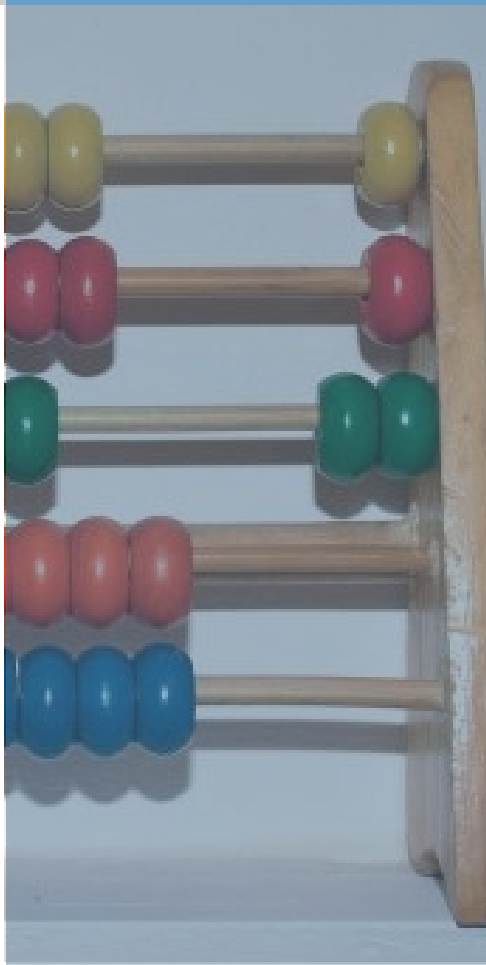
Erwartungswert	Topf	Tasse	Fibel	
Grab 1	0,07	0,07	0,07	0,22
Grab 2	0,07	0,07	0,07	0,22
Grab 3	0,11	0,11	0,11	0,33
Grab 4	0,07	0,07	0,07	0,22
	0,33	0,33	0,33	1

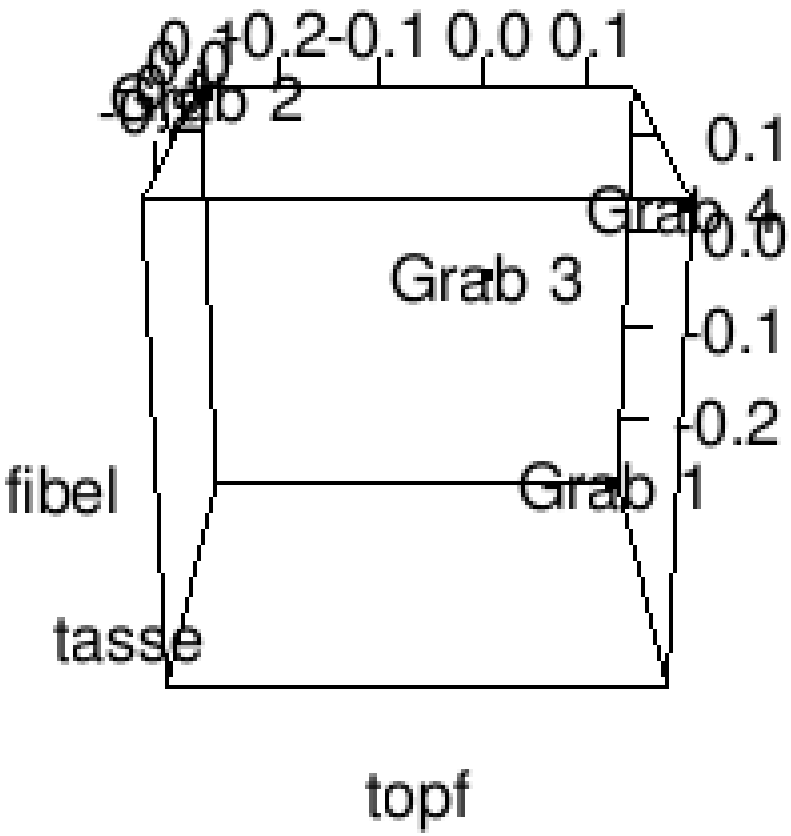
$$z_{ij} = \frac{p_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij}}}$$

$$Beide\ Schritte\ auf\ einmal: \quad z_{ij} = \frac{n_{ij}}{\sqrt{n_{i.} n_{.j}}} - \frac{\sqrt{n_{i.} n_{.j}}}{n}$$

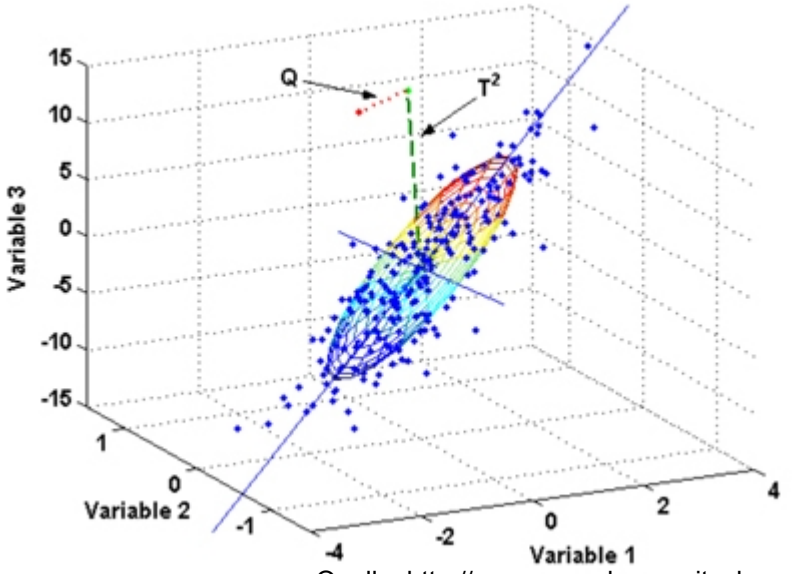
Zweitens: Verschiebung
Der Zentren der Spalten und Zeilen auf 0
(später Koordinaten-Ursprung)

Zentriert zum Erwartungswert	Topf	Tasse	Fibel
Grab 1	0,14	0,14	-0,27
Grab 2	-0,27	0,14	0,14
Grab 3	0	0	0
Grab 4	0,14	-0,27	0,14

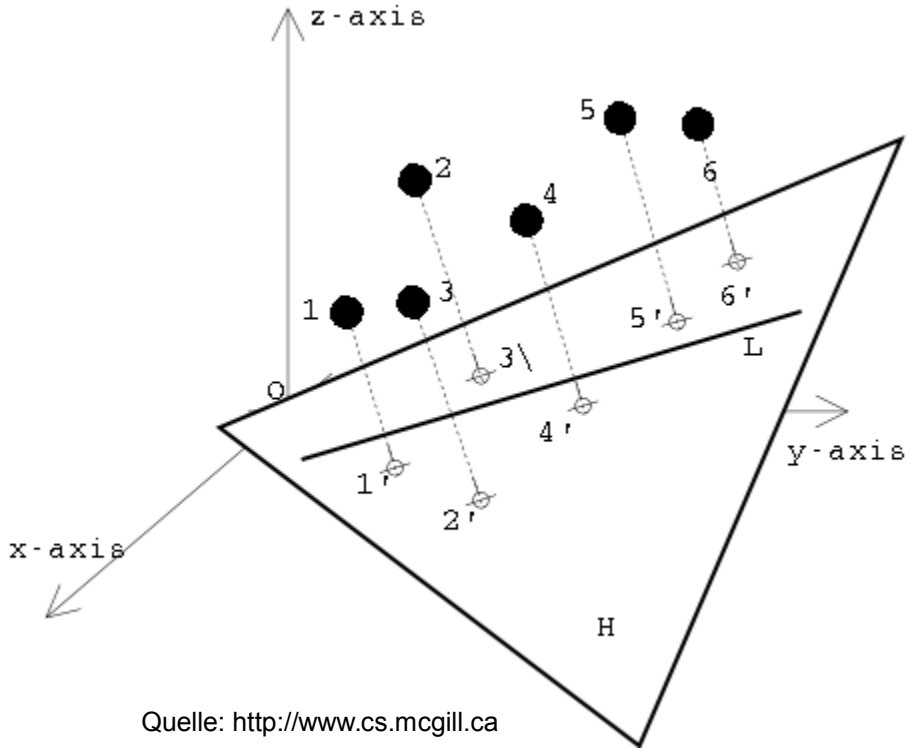



-0.2
-0.1
0.0
0.1
Grab 2
0.1
Grab 4
0.0
Grab 3
0.1
0.2
fibel
Grab 1
tasse
topf

Grundlegende statistische Verfahren für archäologische Datenanalyse in R



Quelle: http://www.aapspharmscitech.org



Quelle: http://www.cs.mcgill.ca

Christian-Albrechts-Universität zu Kiel

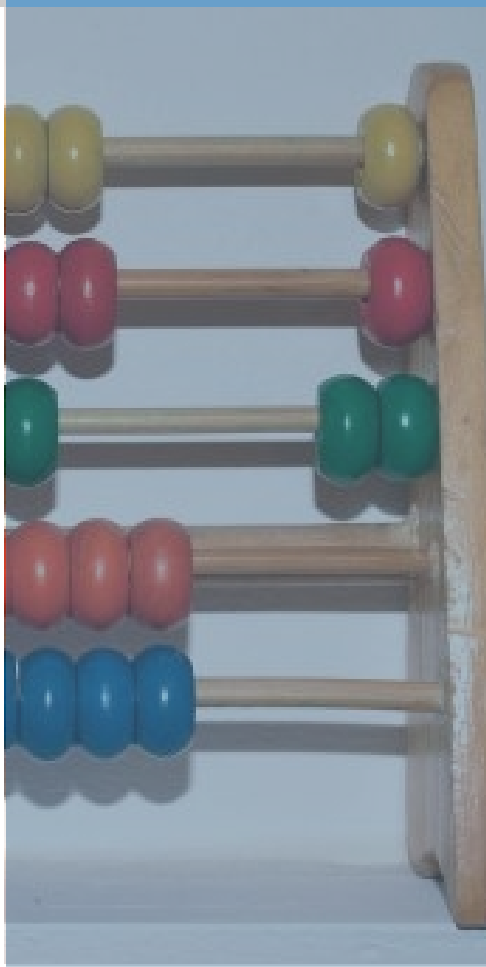
Korrespondenzanalyse: Vorgehen

Extraktion der Dimensionen

SVD
Singulärwertzerlegung, Verfahren zur Dimensionsreduzierung bei minimalem Informationsverlust

$$Z = U * S * V'$$

Z : Matrix mit den standardisierten Daten
U : Matrix für die Zeilenelemente
V : Matrix für die Spaltenelemente
S : Diagonalmatrix mit den Singulärwerten



Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

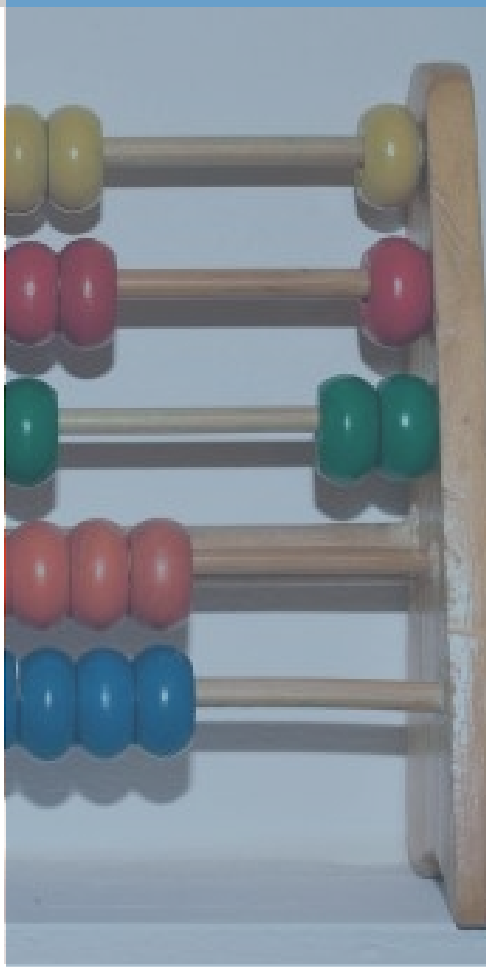
Extraktion der Dimensionen

In R

```
> graeber.zent<-read.csv2("graeber_zent.csv",row.names=1)
> graeber.svd<-svd(graeber.zent)
> graeber.svd
$d
[1] 4.082483e-01 4.082483e-01 8.002750e-18

$u
           [,1]          [,2]       [,3]
[1,] -0.4082483 -7.071068e-01 -0.5773503
[2,]  0.8164966  2.202828e-16 -0.5773503
[3,]  0.0000000  0.000000e+00  0.0000000
[4,] -0.4082483  7.071068e-01 -0.5773503

$v
           [,1]          [,2]       [,3]
[1,] -0.8164966 -1.433065e-16 -0.5773503
[2,]  0.4082483 -7.071068e-01 -0.5773503
[3,]  0.4082483  7.071068e-01 -0.5773503
```



C | A | U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

SVD und Inertia

Die Singulärwerte (Eigenwerte) geben die Inertia wieder

Die Eigenwerte

```
> graeber.svd$d
[1] 4.082483e-01 4.082483e-01 8.002750e-18
```

Die Quadrierten Eigenwerte sind die Inertia der einzelnen Dimensionen

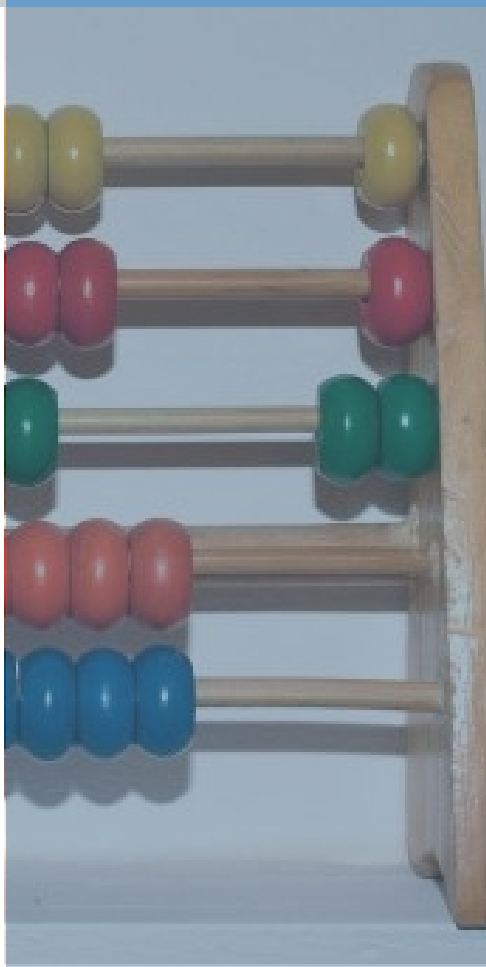
```
> graeber.svd$d^2
[1] 1.666667e-01 1.666667e-01 6.404401e-35
```

Die Summe der Quadrierten Eigenwerte is gleich der Gesamtinertia

```
> sum(graeber.svd$d^2)
[1] 0.3333333
```

Teilt man die Inertia der einzelnen Dimensionen durch die Gesamtinertia, so erhält man den (Eigenwert-)Anteil der Dimensionen

```
> graeber.svd$d^2/sum(graeber.svd$d^2)
[1] 5.000000e-01 5.000000e-01 1.921320e-34
```



C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

Normalisierung der Koordinaten

Skalierung der Koordinaten so, dass

Die Dimensionen nach ihrem Anteil an der Gesamtinertia gewichtet werden

Die Zeilen/Spalten nach ihrem Anteil an der Masse gewichtet werden

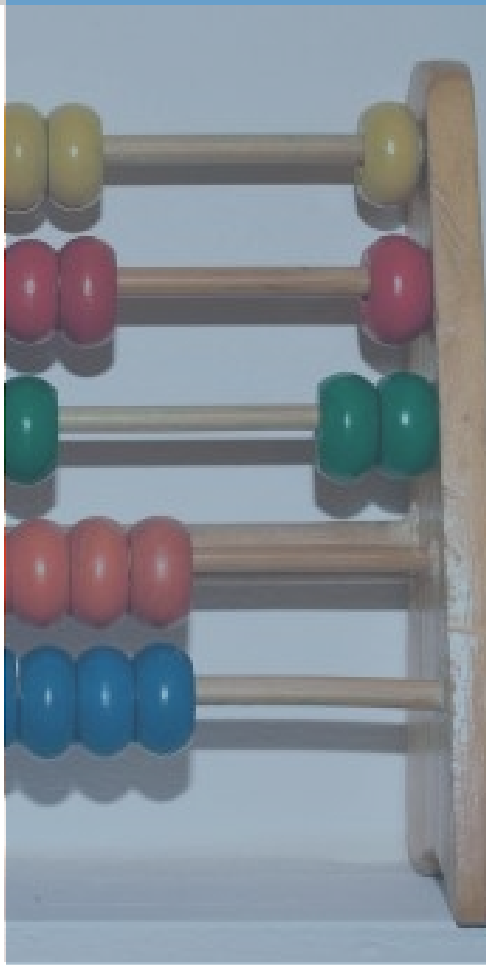
$$\textit{Zeilenpunkte}: \quad r_{ik} = \frac{u_{ik} * \sqrt{s_k}}{\sqrt{p_i}} \qquad \textit{Spaltenpunkte}: \quad c_{ik} = \frac{v_{jk} * \sqrt{s_k}}{\sqrt{p_j}}$$

Mit:

$u, v \rightarrow$ Matrizen der Zeilen/Spalten aus der SVD

$s_k \rightarrow$ Diagonalmatrix

$p_i, p_j \rightarrow$ Massen der Zeilen/Spalten aus der relativen Häufigkeit



C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

Normalisierung der Koordinaten

Beispiel Spalten

Matrize V:

	Dim 1	Dim 2	Dim 3
Topf	-0,82	-1,43E-016	-0,58
Tasse	0,41	-0,71	-0,58
Fibel	0,41	0,71	-0,58

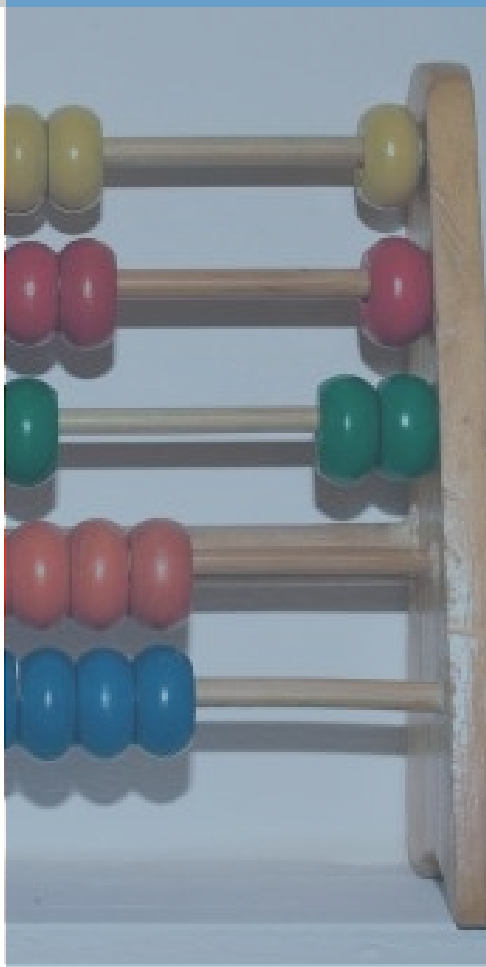
Eigenwerte s_k: 0,408 0,408 8,00E-018
Wurzel von s_k: 0,64 0,64 2.828913E-09

$$Spaltenpunkte: \quad c_{ik} = \frac{v_{jk} * \sqrt{s_k}}{\sqrt{p_j}}$$

Matrizenmultiplikation von v_{jk} mit der Wurzel von s_k …

Massen der Spalten p_j: 0.3333333 0.3333333 0.3333333
Wurzel von p_j: 0.5773503 0.5773503 0.5773503

1. Dim. Mit 1. Wert von s_k, 2. Dim. Mit 2. Wert von p_j…



Christian-Albrechts-Universität zu Kiel

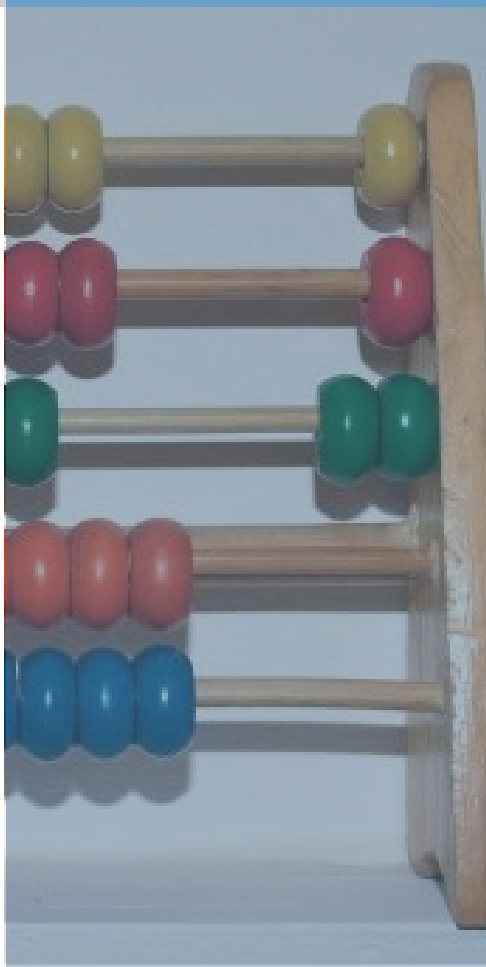
Korrespondenzanalyse: Vorgehen

Normalisierung der Koordinaten

In R

```
> graeber.svd$u%*%sqrt(diag(graeber.svd$d)) /
sqrt(rowSums(graeber.rel))
           [,1]           [,2]           [,3]
[1,] -0.553341 -9.584147e-01 -3.464697e-09
[2,]  1.106682  2.985719e-16 -3.464697e-09
[3,]  0.000000  0.000000e+00  0.000000e+00
[4,] -0.553341  9.584147e-01 -3.464697e-09

> graeber.svd$v%*%sqrt(diag(graeber.svd$d)) /
sqrt(colSums(graeber.rel))
           [,1]           [,2]           [,3]
[1,] -0.903602 -1.585947e-16 -2.828913e-09
[2,]  0.451801 -7.825423e-01 -2.828913e-09
[3,]  0.451801  7.825423e-01 -2.828913e-09
> sites<-graeber.svd$u%*%sqrt(diag(graeber.svd$d)) /
sqrt(rowSums(graeber.rel))
> species<-graeber.svd$v%*%sqrt(diag(graeber.svd$d)) /
sqrt(colSums(graeber.rel))
```



C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Vorgehen

Darstellen Koordinaten

In R

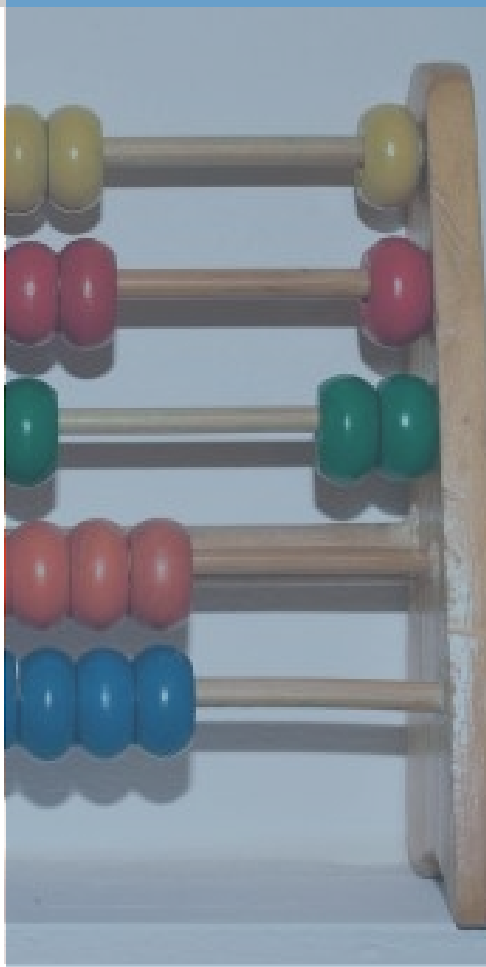
```
> plot(c(-1:2),c(-1:2),type="n")
> text(sites[,1],sites[,2],rownames(sites))
> text(species[,1],species[,2],rownames(species),col="red")
```

Etwas schneller und einfacher als das Verfahren bis hierher:

```
> library(vegan)
This is vegan 1.15-1
> graeber.cca = cca(graeber)
> plot(graeber.cca,scaling=3)
```

Scaling=3: standardmäßig normalisiert R nur die Species (Typen)

Optionen: scaling = 1 : Normalisierung der Sites
scaling = 2 : Normalisierung der Species
scaling = 3 : Symmetrische Normalisierung von Sites und Species
scaling = 0 : Keine Normalisierung



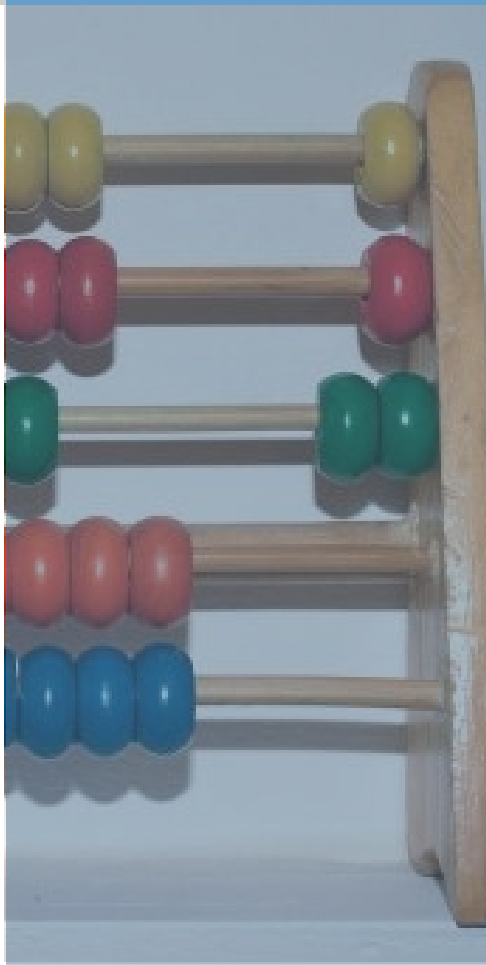
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Aufgabe

Revolutionieren Sie die Chronologie der Trichterbecherkultur

Gegeben sind die Auszählungen verschiedener Dekorationsmuster für verschiedene archäologische Kulturen. Führen Sie eine Korrespondenzanalyse durch (benutzen Sie den kurzen Weg) und stellen Sie das Ergebnis graphisch dar.

Datei: trbchron.csv



Christian-Albrechts-Universität zu Kiel

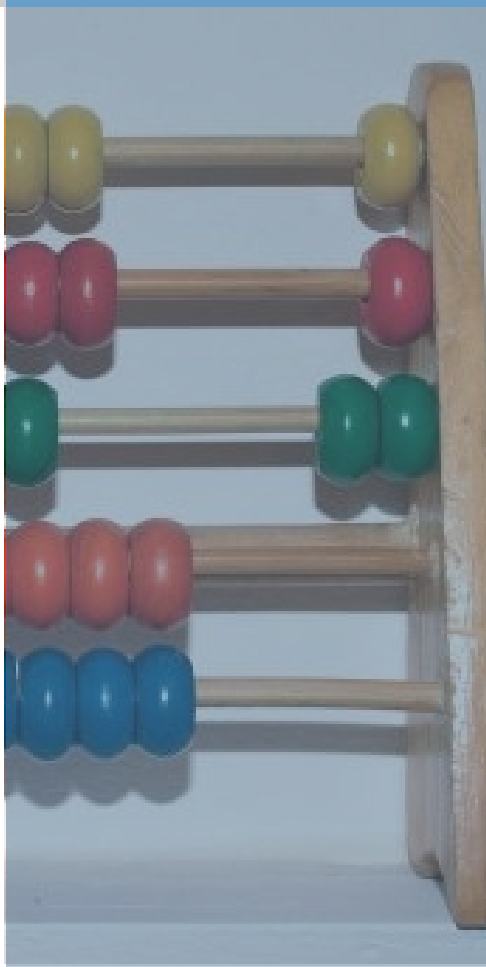
Korrespondenzanalyse: Lösung

Revolutionieren Sie die Chronologie der Trichterbecherkultur

Gegeben sind die Auszählungen verschiedener Dekorationsmuster für verschiedene archäologische Kulturen. Führen Sie eine Korrespondenzanalyse durch (benutzen Sie den kurzen Weg) und stellen Sie das Ergebnis graphisch dar.

Datei: trbchron.csv

```
> trb<-read.csv2("trbchron.csv",row.names=1)
> trb.cca<-cca(trb)
> plot(trb.cca)
```



Christian-Albrechts-Universität zu Kiel

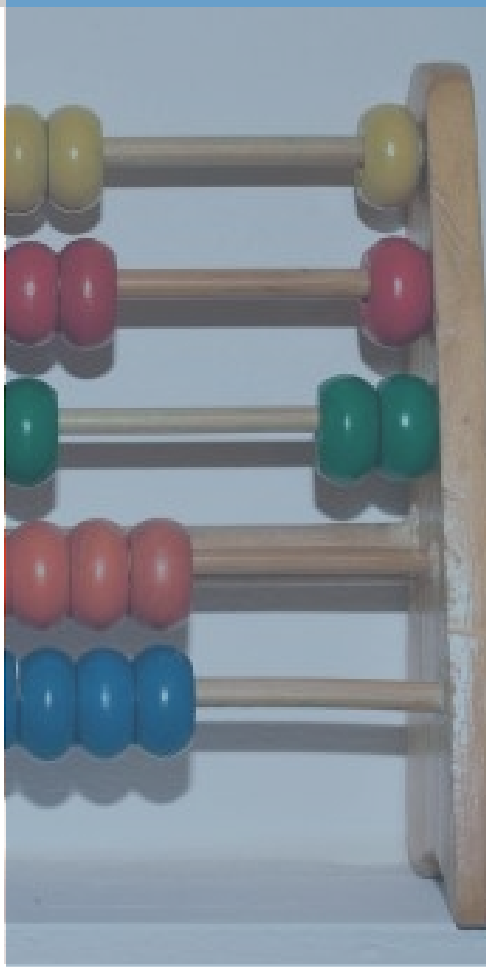
Korrespondenzanalyse: Interpretation

Interpretation

Dimensionen und Inertia
Im Beispiel bilden 1. und 2. Dimension die Daten fast vollständig ab (je ca. 50%)

Mittelpunkt des Koordinatensystems
Gibt den absoluten „Durchschnittsfall“ an. Im Beispiel weist Grab 3 keine Besonderheiten hinsichtlich der Typen auf (Da alle vertreten sind).

Abstand der Punkte zueinander
Das Beispiel ist symmetrisch konstruiert (Es gibt keine Gräber, die sich ähnlicher sind als andere).
Wenn ähnlichere sites oder species existieren, so werden sie näher beieinander abgebildet.
Die Interpretation des Abstandes von sites zu species ist gefährlich:
Es gibt keine mathematische Begründung dafür (sie werden im gleichen Raum/Koordinatensystem abgebildet, stammen aber aus unterschiedlichen Räumen)

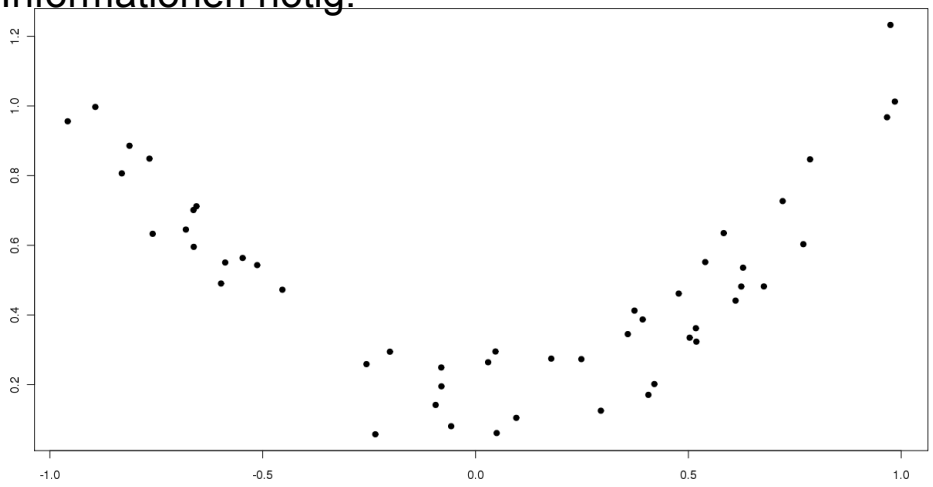


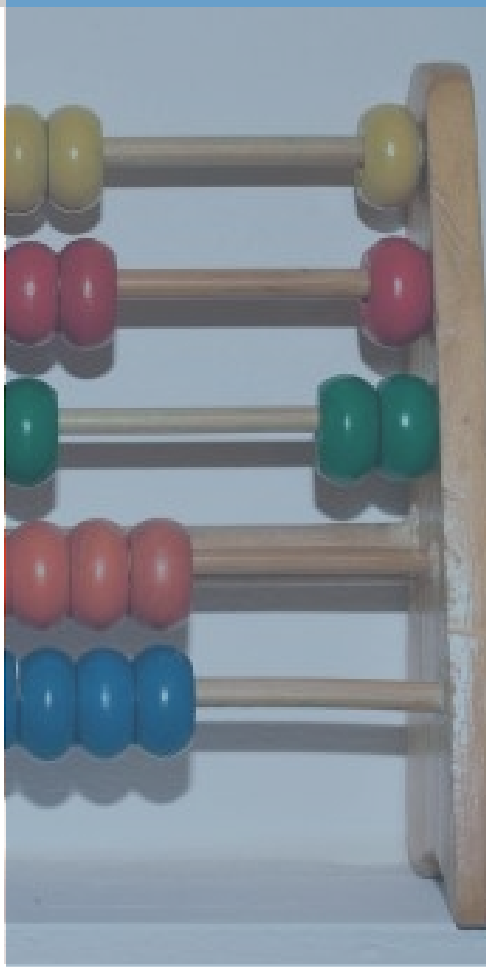
C | A | U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Interpretation

Guttman-Effekt (horseshoe, Parabel)

Wird in der Archäologie häufig als Nachweis für eine Zeitliche Orientierung angesehen.
Der Guttman-Effekt tritt ein, wenn ein Prozess die Daten auf mehreren Ebenen beeinflußt.
Der größte beeinflussende Faktor, gegeben eine längere Laufzeit, ist meist die Zeit, aber:
Das muß nicht immer der Fall sein.
Abprüfen gegen andere Informationen nötig.





C A U
Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Anwendung

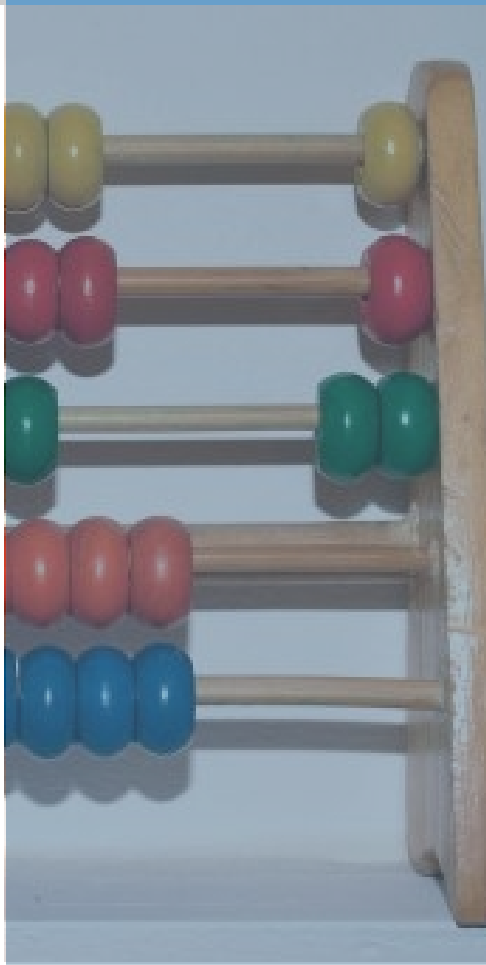
Seriation

Ordnung von Fundorten/Fundtypen nach Ähnlichkeit
Verschiedene Verfahren in Anwendung, Identifikation einer möglichen zeitlichen Abfolge
Wenn eine Dimension als zeitlich beeinflußt identifiziert werden kann, kann diese als Ordnungskriterium für Seriation verwendet werden (z.B. Winbasp)

```
> vegemite(graeber,cca(graeber))

      GGGG
      rrrr
      aaaa
      bbbb

      4132
Topf  111.
Fibel 1.11
Tasse .111
 sites species
     4       3
```

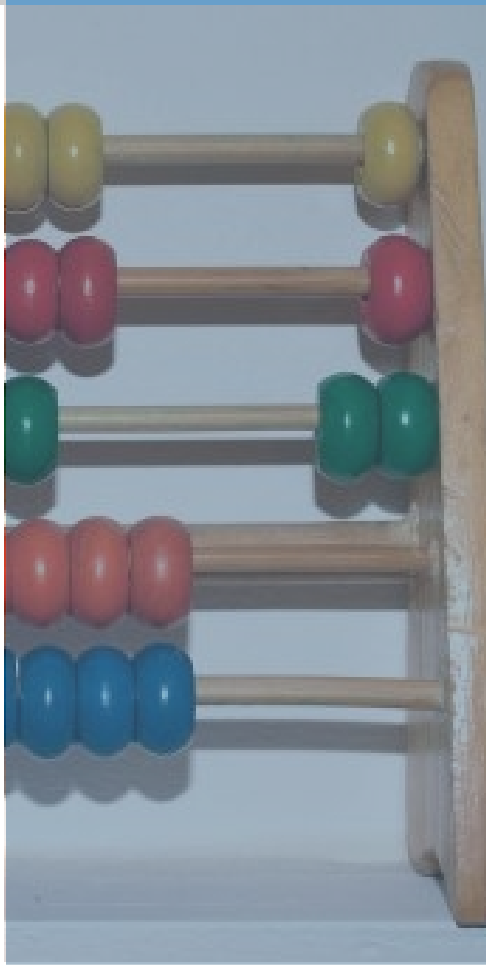


Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Seriation: Aufgabe

Revolutionieren Sie die Chronologie der Trichterbecherkultur

Gegeben sind die Auszählungen verschiedener Dekorationsmuster für verschiedene archäologische Kulturen (geladen in der letzten Aufgabe). Führen Sie eine Seriation anhand der Korrespondenzanalyse durch.



Christian-Albrechts-Universität zu Kiel

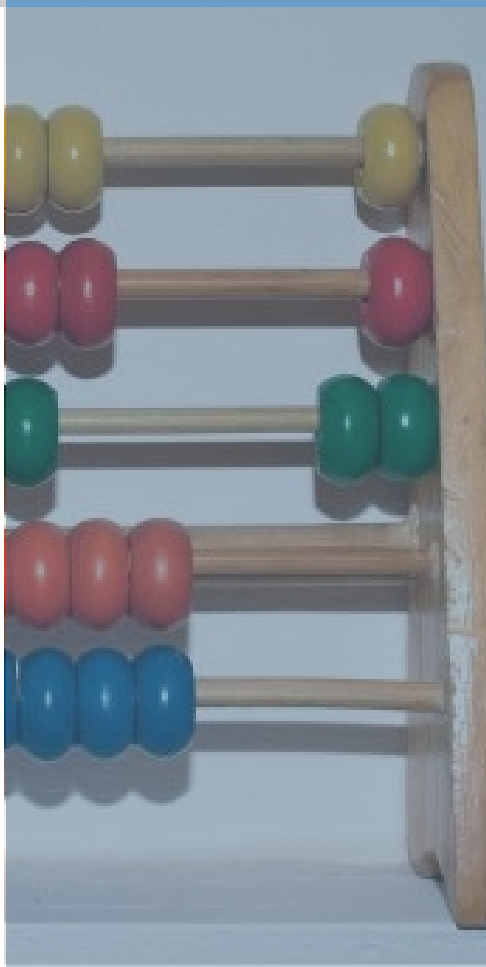
Korrespondenzanalyse: Seriation: Lösung

Revolutionieren Sie die Chronologie der Trichterbecherkultur

Gegeben sind die Auszählungen verschiedener Dekorationsmuster für verschiedene archäologische Kulturen (geladen in der letzten Aufgabe). Führen Sie eine Seriation anhand der Korrespondenzanalyse durch.

```
> vegemite(trb,trb.cca)

                      K
                      l
                     Fi
                     un
                     ct
                     he
                     sb
                    Oba
                    xekE
                    irkG
                    egeK
Stacheldrahtmotiv  9132
Leiterbandmotiv    6234
Schurverzierung    3165
 sites species
     4       3
```



Christian-Albrechts-Universität zu Kiel

Korrespondenzanalyse: Varianten

Wem Korrespondenzanalyse noch nicht reicht...

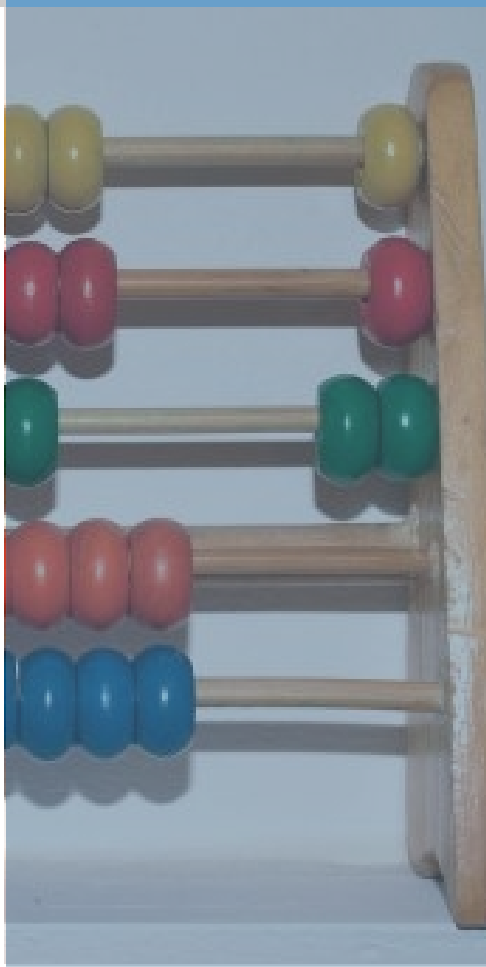
Kanonische Korrespondenzanalyse
Es wird ein Set an „Umweltvariablen“ für jeden Datensatz mitgegeben. An diesen werden die Dimensionen ausgerichtet (nicht an der optimalen Wiedergabe der Information).

Ziel ist es, herauszufinden, wieviel Prozent der Information durch die Umweltvariablen erklärt wird.

Partielle (kanonische) Korrespondenzanalyse
Es wird ein Set an „Umweltvariablen“ für jeden Datensatz mitgegeben. Es wird eine kanonische Korrespondenzanalyse mit den Umweltvariablen durchgeführt. Als Ergebnis werden jedoch nur die Werte einbezogen, die sich **nicht** durch die Umweltvariablen erklären lassen.

Ziel ist es, Einflüsse aus den Daten herauszufiltern

Alles in R mit `cca` möglich...



Das war die letzte Sitzung

Bitte lesen Sie Shennan 13 bzw.
Backhaus et al. 2006 (11. Auflage) oder
Greenacre 1984 oder
Blasius 2001 (wenn Sie es sich dreckig geben wollen)