

Kanonische Korrelation

Kanonische Korrelation versucht Assoziationen zwischen Gruppen von Variablen zu finden und zu quantifizieren. Dabei wird die Korrelation einer Linearkombination einer Menge von Variablen mit einer Linearkombination einer anderen Menge von Variablen verglichen. Anwendung ist meist die Vorhersage der Werte der einen Menge durch die Werte der anderen Menge, wie z.B.

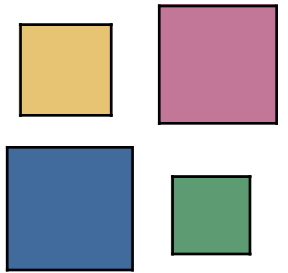
- (1) Vergleich von Lesetests und Rechentests
- (2) Messungen vor und nach einer Behandlung
- (3) Leistung vor und nach dem Studium.

Zwei Mengen von Variablen, \mathbf{X} , \mathbf{Y} .

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_p]$$

$$\mathbf{Y} = [\mathbf{Y}_1 \ \mathbf{Y}_2 \ \dots \ \mathbf{Y}_q]$$

Modelliert werden soll \mathbf{BY} durch \mathbf{AX} d.h., schätze \mathbf{A} und \mathbf{B} um die Korrelation zwischen \mathbf{AY} , \mathbf{BX} zu maximieren.



Notation

Gegeben seien Daten

$$\mathbf{X}_{n \times (p+q)} = \left[\mathbf{X}^{(1)} \mid \mathbf{X}^{(2)} \right]$$

und eine Partitionierung der Varianz-Covarianz Matrix

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right]_{(p+q) \times (p+q)}$$

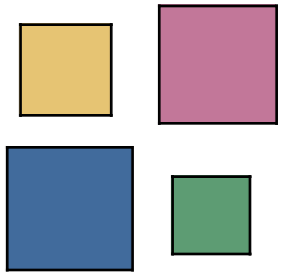
Ziel der Kanonischen Korrelation ist es nun ein paar *wenige* Korrelationen auszuwählen, um die Assoziation zwischen $\mathbf{X}^{(1)}$ und $\mathbf{X}^{(2)}$ zu beschreiben.

Dazu werden die Linearkombinationen

$$\mathbf{U} = \mathbf{a}'\mathbf{X}^{(1)}$$

$$\mathbf{V} = \mathbf{b}'\mathbf{X}^{(2)}$$

betrachtet.



Maximierungsaufgabe

Es gilt

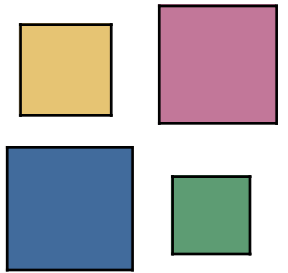
$$\begin{aligned}\text{Var}(\mathbf{U}) &= \mathbf{a}' \text{Cov}(\mathbf{X}^{(1)}) \mathbf{a} = \mathbf{a}' \boldsymbol{\Sigma}_{11} \mathbf{a} \\ \text{Var}(\mathbf{V}) &= \mathbf{b}' \text{Cov}(\mathbf{X}^{(2)}) \mathbf{b} = \mathbf{b}' \boldsymbol{\Sigma}_{22} \mathbf{b} \\ \text{Cov}(\mathbf{U}, \mathbf{V}) &= \mathbf{a}' \text{Cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \mathbf{b} = \mathbf{a}' \boldsymbol{\Sigma}_{12} \mathbf{b}\end{aligned}$$

Gesucht werden also Vektoren \mathbf{a} und \mathbf{b} derart, dass

$$\text{Cor}(\mathbf{U}, \mathbf{V}) = \frac{\mathbf{a}' \boldsymbol{\Sigma}_{12} \mathbf{b}}{\sqrt{\mathbf{a}' \boldsymbol{\Sigma}_{11} \mathbf{a}} \sqrt{\mathbf{b}' \boldsymbol{\Sigma}_{22} \mathbf{b}}}$$

maximiert wird.

Das erste Paar kanonischer Variablen ist U_1 und V_1 , welche obige Korrelation maximieren. Allgemein maximiert das i -te Paar kanonischer Variablen U_i und V_i die obige Korrelation und ist mit allen $i - 1$ vorausgehenden kanonischen Variablen unkorreliert.



Lösung

Die Koeffizienten Vektoren \mathbf{a} und \mathbf{b} aus $\mathbf{U} = \mathbf{a}'\mathbf{X}^{(1)}$ und $\mathbf{V} = \mathbf{b}'\mathbf{X}^{(2)}$ ergeben sich aus

$$U_1 = \underbrace{\mathbf{e}'_1 \Sigma_{11}^{-1/2}}_{\mathbf{a}'_1} \mathbf{X}^{(1)}$$

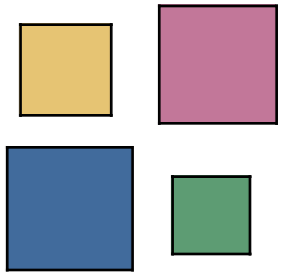
$$V_1 = \underbrace{\mathbf{f}'_1 \Sigma_{22}^{-1/2}}_{\mathbf{b}'_1} \mathbf{X}^{(2)}$$

und haben maximale Korrelation ρ_1^* .

Die \mathbf{e}_i und die \mathbf{f}_i sind die Eigenvektoren von:

$$\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2} \quad \text{bzw.} \quad \Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$$

Die dazugehörigen Eigenwerte sind die $\rho_1^{*2} \geq \rho_2^{*2} \geq \dots \geq \rho_m^{*2}$ mit $m = \min\{p, q\}$.



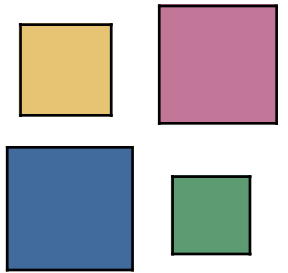
Eigenschaften

Es gilt:

$$\begin{aligned} \text{Var}(U_k) &= \text{Var}(V_k) = 1 \\ \text{Cov}(U_k, U_l) &= \text{Cor}(U_k, U_l) = 0 \quad k \neq l \\ \text{Cov}(V_k, V_l) &= \text{Cor}(V_k, V_l) = 0 \quad k \neq l \\ \text{Cov}(U_k, V_l) &= \text{Cor}(U_k, V_l) = 0 \quad k \neq l \end{aligned}$$

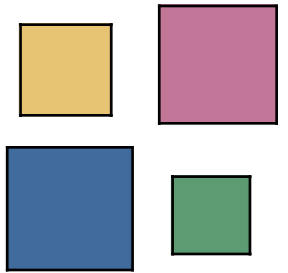
Für die Korrelationsmatrix $R_{2m \times 2m}$ der kanonischen Variablen (\mathbf{U}, \mathbf{V}) ergibt sich somit:

$$R = \left(\begin{array}{cc|cc} R_{11} & & R_{12} & \\ \hline R_{21} & & R_{22} & \end{array} \right) = \left(\begin{array}{cccccccc} 1 & 0 & \cdots & 0 & \rho_1^* & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \rho_2^* & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & \rho_m^* \\ \rho_1^* & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \rho_2^* & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & \rho_m^* & 0 & 0 & \cdots & 1 \end{array} \right)$$



Bemerkungen

- Die Kanonische Korrelation eliminiert also die Korrelation zwischen den Variablen in einer Gruppe, so dass nur noch Korrelationen zwischen den beiden Gruppen übrig bleiben.
- Beim Übergang von Zufallsvariablen zur tatsächlichen Stichprobe wird Σ zu \mathbf{S} , ρ zu r , sowie \mathbf{a} , \mathbf{b} , \mathbf{U} , \mathbf{V} zu den entsprechenden Schätzern $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$.
- Interpretiert werden meist die Koeffizienten \mathbf{a} und \mathbf{b} , und weniger \mathbf{U} und \mathbf{V} .
- Da Korrelationen betrachtet werden, ist es im allgemeinen nicht nötig die Variablen vorher zu standardisieren.



Inferenz

Falls $\Sigma_{12} = 0$, dann haben $\mathbf{a}'\mathbf{X}^{(1)}$ und $\mathbf{b}'\mathbf{X}^{(2)}$ die Kovarianz $\mathbf{a}'\Sigma\mathbf{b} = 0$ für jegliche Vektoren \mathbf{a} und \mathbf{b} .

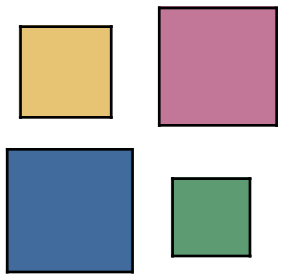
Demnach müssen auch alle kanonischen Korrelationen gleich Null sein, und es ist sinnlos eine kanonische Korrelations Analyse durchzuführen.

Zu testen ist also

$$H_0 : \Sigma_{12} = 0 \quad \text{vs.} \quad H_1 : \Sigma_{12} \neq 0$$

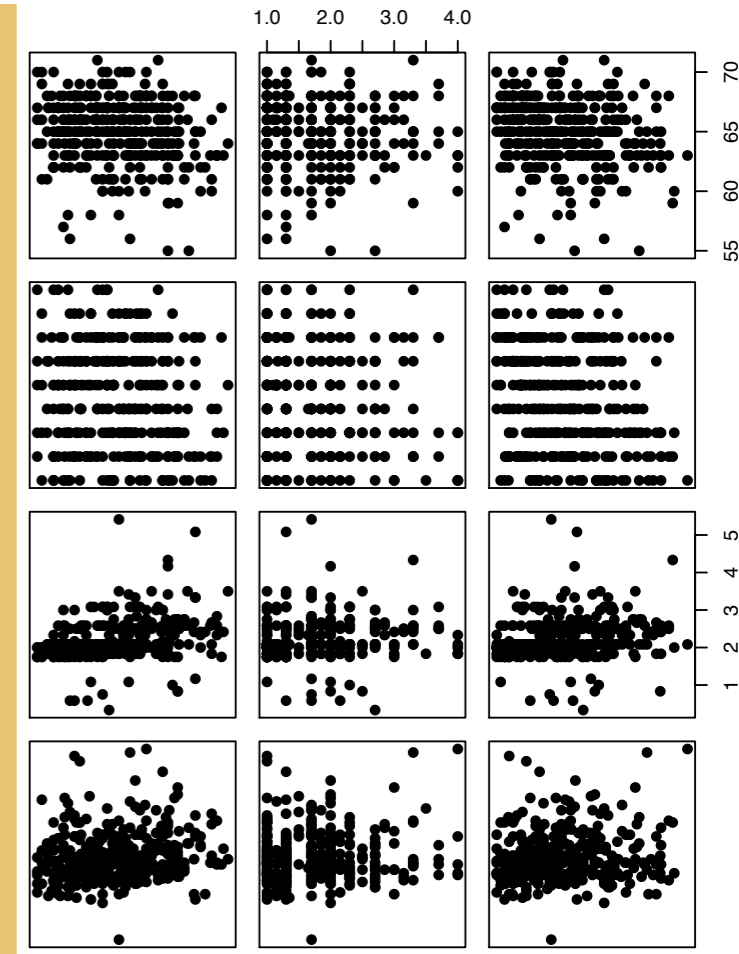
Als Likelihood Ratio Test ergibt sich

$$-2 \ln \Lambda = n \ln \left(\frac{|S_{11}| |S_{22}|}{|S|} \right) = -n \prod_{i=1}^m (1 - \rho_i^{*2}) > \chi_{pq}^2(\alpha)$$

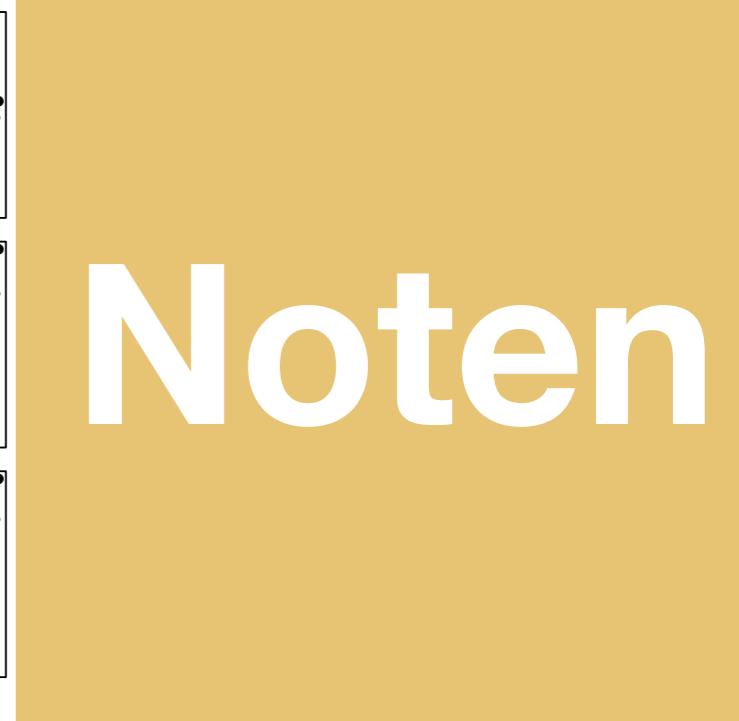
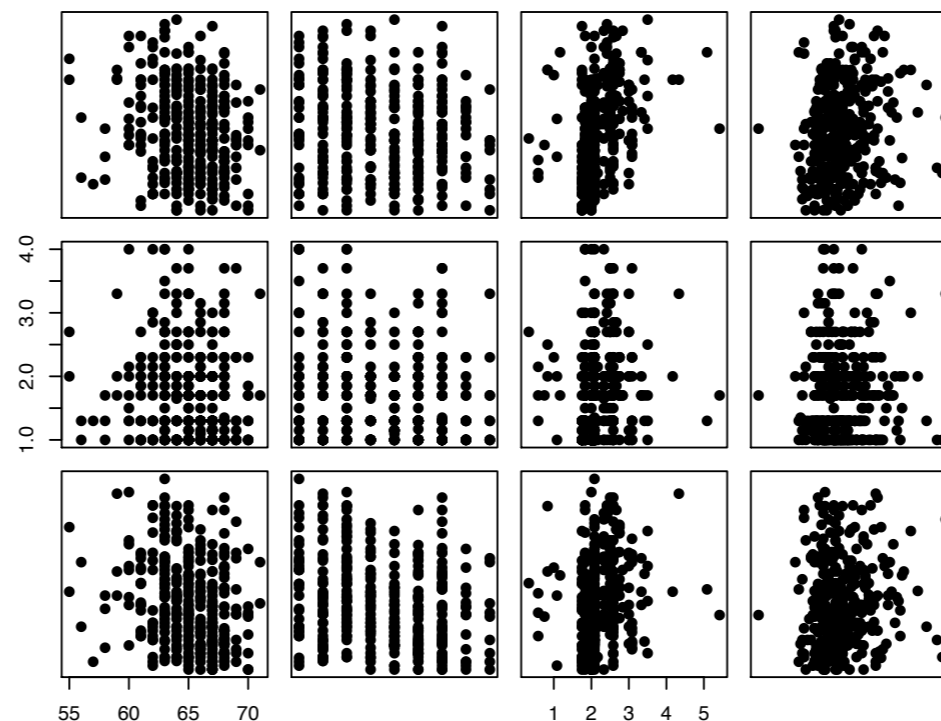


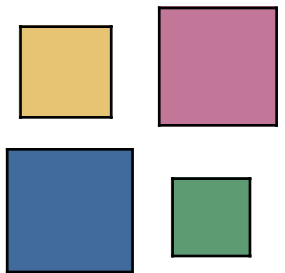
Beispiel

- Mathestudenten
 - Dauern:
 - Geburtsjahr
 - Anfang
 - VorDauer
 - HauptDauer
 - Noten
 - VDnote
 - Dnote
 - HDnote



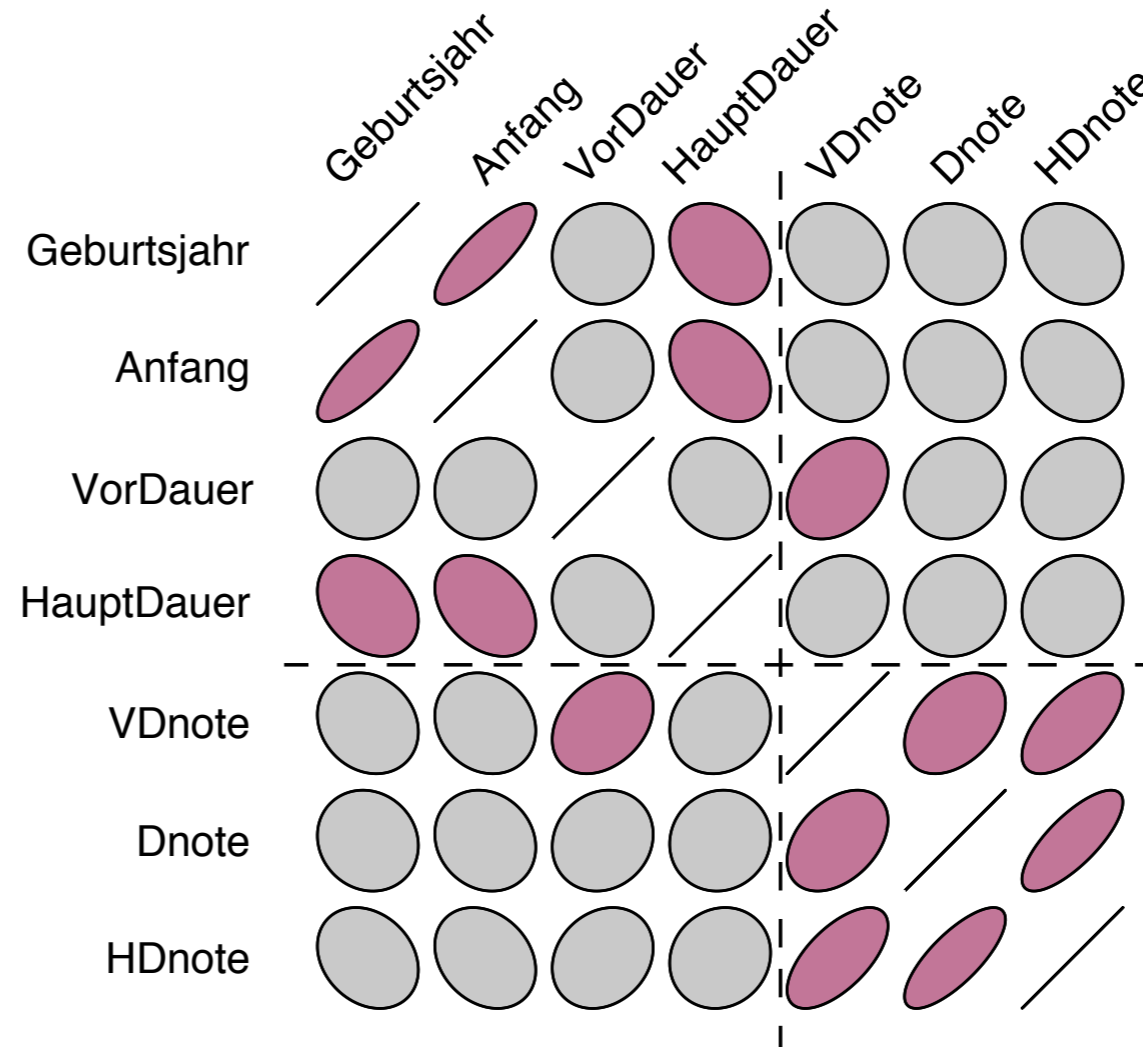
- Frage:
Wie sind Noten mit Zeiten assoziiert?



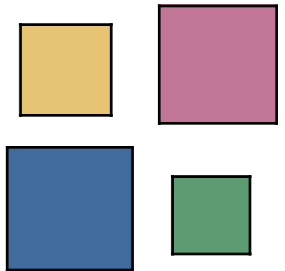


Beispiel cont.

- Korrelationen der Rohdaten



	<i>Gbj</i>	<i>Anf</i>	<i>VD</i>	<i>HD</i>	<i>VDn</i>	<i>Dn</i>	<i>HDn</i>
<i>Geburtsjahr</i>							
<i>Anfang</i>	0.816						
<i>VorDauer</i>	0.044	0.06					
<i>HauptDauer</i>	-0.294	-0.36	-0.141				
<i>VDnote</i>	-0.168	-0.15	0.319	0.137			
<i>Dnote</i>	-0.099	-0.13	0.122	0.066	0.40		
<i>HDnote</i>	-0.237	-0.24	0.200	0.101	0.64	0.763	



Beispiel cont.

- R-Code:

```
options(digits=2)

Dauern<-MSS[,1:4]
Noten <-MSS[,5:7]

plot(MSS, pch=16)

round(cor(MSS), 3)

plotcorr(cor(MSS))

MSSC <- cancel(Dauern, Noten)

U <- as.matrix(Dauern)%*%MSSC$xcoef
V <- as.matrix(Noten)%*%MSSC$ycoef

UV <- cbind(U[,1:3],V)

round(cor(UV), 3)

plot(as.data.frame(UV))

plotcorr(cor(as.data.frame(UV)))
```

- Ergebnis von cancel

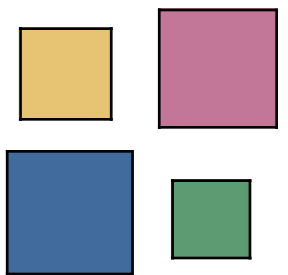
```
$cor
[1] 0.403 0.208 0.077

$xcoef
      [,1]      [,2]      [,3]      [,4]
Geburtsjahr -0.0076 -0.0069 -0.0321 -0.0031
Anfang      -0.0031 -0.0157  0.0402  0.0057
VorDauer     0.0815 -0.0440 -0.0073 -0.0360
HauptDauer   0.0131 -0.0222 -0.0024  0.0402

$ycoef
      [,1]      [,2]      [,3]
VDnote  0.059 -0.086  0.014
Dnote  -0.017 -0.055 -0.103
HDnote  0.047  0.164  0.052

$xcenter
Geburtsjahr      Anfang      VorDauer      HauptDauer
      65.0          85.4          2.2          4.0

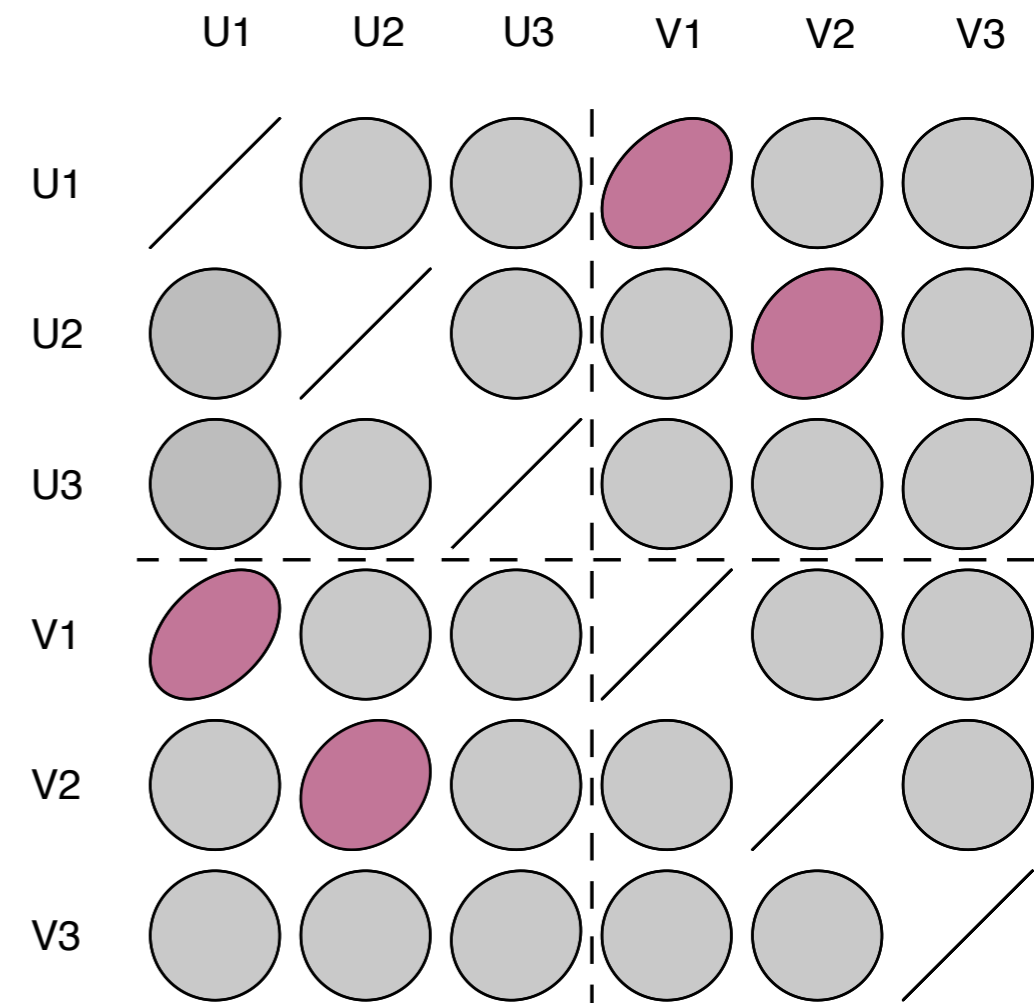
$ycenter
VDnote  Dnote  HDnote
      2.2    1.7    1.9
```

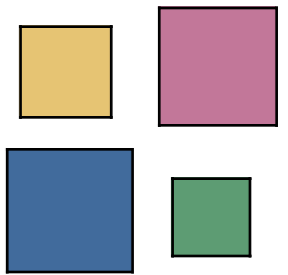


Beispiel cont.

- Kanonische Korrelationen

	U_1	U_2	U_3	V_1	V_2	V_3
U_1	1.00	0.00	0.000	0.40	0.00	0.000
U_2	0.00	1.00	0.000	0.00	0.21	0.000
U_3	0.00	0.00	1.000	0.00	0.00	0.077
V_1	0.40	0.00	0.000	1.00	0.00	0.000
V_2	0.00	0.21	0.000	0.00	1.00	0.000
V_3	0.00	0.00	0.077	0.00	0.00	1.000





Beispiel cont.

- Kanonische Variablen
- Korrelation ist eine hochgradig lineare, auf Normalverteilt. basierende Größe
- Nicht-lineare Assoziationen von nicht normalverteilten Daten werden somit meist nicht entdeckt!

