

Statistik für Digital Humanities

Multivariate Analysis of Variance (MANOVA)

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

01. März 2021

[Letzte Aktualisierung: 01/03/2021, 10:57]

Anknüpfungspunkt t-Test & ANOVA

- ANOVA untersucht Unterschiede einer abhängigen Variable bei mehreren Gruppen
- MANOVA untersucht Unterschiede mehrerer abhängiger Variablen (Outcomes)

- ANOVA: univariat
- MANOVA: multivariat
- Prinzipien von ANOVA übertragbar (Faktoren, unabhängig/abhängig, Post Hoc, Kontraste, Interaktion)
- Folgeanalysen mittels ANOVA oder Diskriminantenanalyse (Siehe Moodle)

Warum nicht mehrere ANOVA durchführen?

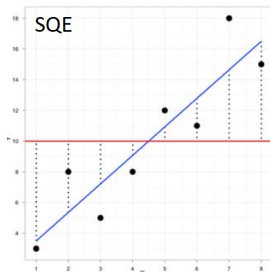
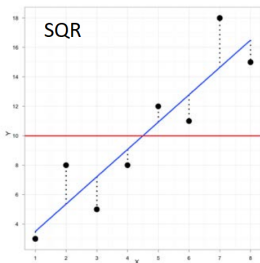
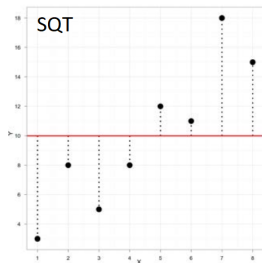
Warum nicht mehrere ANOVA durchführen?

- Problem: familienbezogene / experimentbezogene Fehlerrate $1 - (1 - \alpha)^k$ mit $k = \text{Anzahl der Kombinationen}$ α ist die Typ 1 Fehlerwahrscheinlichkeit → Siehe ANOVA
- MANOVA betrachtet alle abhängigen Variablen, während jede ANOVA nur je eine betrachten würde
 - MANOVA erlaubt Aussagen über Kombinationen der Outcomes
 - exakter Zusammenhang zwischen Korrelation der Outcomes und Aussagekraft der MANOVA strittig
 - → Cole & Maxwell & Arvey & Salas (1994): *How the power of MANOVA can both increase and decrease as a function of the intercorrelations between the dependent variables*
 - Einschätzung der Aussagekraft (Power) bzgl. Interkorrelation generell schwierig, also am besten Vergleichsarbeiten suchen

Berechnung

Wiederholung ANOVA:

- $F = \frac{MQE}{MQR} = \frac{\text{Systematische Variation}}{\text{Unsystematische Variation}}$
- $MQE = \frac{SQE}{k-1}$
- $MQR = \frac{SQR}{n-k}$
- $k = \text{Anzahl der Gruppen}$



Bei MANOVA ersetzen wir *einfach* Zahlen der univariaten ANOVA Analyse mit Matrizen

Berechnung

MANOVA:

- Signifikanzmatrix $HE^{-1} = H * E^{-1} = \frac{H}{E} = \frac{\text{Systematische Variation}}{\text{Unsystematische Variation}}$
Matrixdivision ist Multiplikation der inversen Matrix
- $H = SQKPE$ = Quadratsummenkreuzproduktmatrix erklärt Hypothesenmatrix
- $E = SQKPR$ = Quadratsummenkreuzproduktmatrix der Residuen Errormatrix

Quadratsummenkreuzproduktmatrix

$SQKP^* =$

	Variable 1	Variable 2	...
Variable 1	SQ*	KP*	...
Variable 2	KP*	SQ*	...
...

- SQT, SQR, SQE analog zu vorher
- $KPT = \sum_{i=1}^n (x_{i,var1} - \overline{x_{var1}}) * (x_{i,var2} - \overline{x_{var2}}) * \dots$
- $KPE = \sum_{i=1}^n (\overline{x_{group,var1}} - \overline{x_{var1}}) * (\overline{x_{group,var2}} - \overline{x_{var2}}) * \dots$
- $KPR = \sum_{i=1}^n (x_{i,var1} - \overline{x_{group,var1}}) * (x_{i,var2} - \overline{x_{group,var2}}) * \dots$

Überprüfung per $SQT = SQR + SQE$

Überprüfung per $KPT = KPR + KPE$

Überprüfung per $SQKPT = SQKPR + SQKPE$

Beispiel

Zeichenlänge des Dokumententitels und Dokumentes pro Autor

Autor 1		Autor 2	
Titel	Dokument	Titel	Dokument
35	250	20	400
35	280	30	170
50	400	40	300

$$\overline{\text{Titel}} = 35 \quad \overline{\text{Dokument}} = 300$$

$$\overline{\text{Autor1, Titel}} = 40 \quad \overline{\text{Autor2, Titel}} = 30$$

$$\overline{\text{Autor1, Dokument}} = 310 \quad \overline{\text{Autor2, Dokument}} = 290$$

Beispiel

$$\bar{T} = 35 \quad \bar{D} = 300 \quad \overline{A1, T} = 40 \quad \overline{A2, T} = 30 \quad \overline{A1, D} = 310 \quad \overline{A2, D} = 290$$

A1					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A1, T}$	$D - \overline{A1, D}$
35	250	0	-50	-5	-60
35	280	0	-20	-5	-30
50	400	15	100	10	90

A2					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A2, T}$	$D - \overline{A2, D}$
20	400	-15	100	-10	110
30	170	-5	-130	0	-120
40	300	5	0	10	10

$$SQT_T = 0 + 0 + 15^2 + (-15)^2 + (-5)^2 + 5^2 = 500$$

$$SQT_D = (-50)^2 + (-20)^2 + 100^2 + 100^2 + (-130)^2 + 0 = 39800$$

$$KPT = 0 + 0 + (15 * 100) + (-15 * 100) + (-5 * -130) + 0 = 650$$

$$SQKPT =$$

	T	D
T	500	650
D	650	39800

Beispiel

$$\bar{T} = 35 \quad \bar{D} = 300 \quad \overline{A1, T} = 40 \quad \overline{A2, T} = 30 \quad \overline{A1, D} = 310 \quad \overline{A2, D} = 290$$

A1					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A1, T}$	$D - \overline{A1, D}$
35	250	0	-50	-5	-60
35	280	0	-20	-5	-30
50	400	15	100	10	90

A2					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A2, T}$	$D - \overline{A2, D}$
20	400	-15	100	-10	110
30	170	-5	-130	0	-120
40	300	5	0	10	10

$$SQE_T = 3 * (40 - 35)^2 + 3 * (30 - 35)^2 = 150$$

$$SQE_D = 3 * (310 - 300)^2 + 3 * (290 - 300)^2 = 600$$

$$KPE = 3 * (40 - 35) * (310 - 300) + 3 * (30 - 35) * (290 - 300) = 300$$

	T	D
SQKPE =	T 150	300
	D 300	600

Beispiel

$$\bar{T} = 35 \quad \bar{D} = 300 \quad \overline{A1, T} = 40 \quad \overline{A2, T} = 30 \quad \overline{A1, D} = 310 \quad \overline{A2, D} = 290$$

A1					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A1, T}$	$D - \overline{A1, D}$
35	250	0	-50	5	-60
35	280	0	-20	5	-30
50	400	15	100	20	90

A2					
T	D	$T - \bar{T}$	$D - \bar{D}$	$T - \overline{A2, T}$	$D - \overline{A2, D}$
20	400	-15	100	-20	110
30	170	-5	-130	-10	-120
40	300	5	0	0	10

$$SQR_T = (35 - 40)^2 + (35 - 40)^2 + (50 - 40)^2 + (20 - 30)^2 + 0 + (40 - 30)^2 = 350$$

$$SQR_D = (250 - 310)^2 + (280 - 310)^2 + (400 - 310)^2 + (400 - 290)^2 + (170 - 290)^2 + (300 - 290)^2 = 39200$$

$$KPR = (35 - 40) * (250 - 310) + (35 - 40) * (280 - 310) + (50 - 40) * (400 - 310) + (20 - 30) * (400 - 290) + 0 + (40 - 30) * (300 - 290) = 350$$

$$SQKPR = \begin{array}{|c|c|c|} \hline & T & D \\ \hline T & 350 & 350 \\ \hline D & 350 & 39200 \\ \hline \end{array}$$

Beispiel

$$E = SQKPR =$$

	T	D
T	350	350
D	350	39200

$$H = SQKPE =$$

	T	D
T	150	300
D	300	600

$$HE^{-1} = H * E^{-1} // \text{Lösung via } R$$

```
E <- matrix( c(350,350,
               350,39200
             ), nrow=2, byrow=TRUE)
H <- matrix( c( 150,300,
               300,600
             ), nrow=2, byrow=TRUE)
H %*% solve(E)
```

$$HE^{-1} =$$

0.440	-0.012
-0.880	0.023

wir können keine Matrix gegen einen Signifikanzwert vergleichen
deshalb ...

Diskriminante Funktionsvariate

$$HE^{-1} = \begin{array}{|c|c|} \hline 0.440 & -0.012 \\ \hline -0.880 & 0.023 \\ \hline \end{array}$$

wir können keine Matrix gegen einen Signifikanzwert vergleichen
deshalb ...

- Eigenvektoren als Diskriminante Funktionsvariate
 - Lineare Funktionen, die auf Basis des Outcomes die Prädiktoren (Gruppen) vorhersagt / diskriminiert
 - Jede Variate ist ein multiples Regressionsmodell mit den Outcomes als Prädiktoren und Elementen der Eigenvektoren als Regressionskoeffizienten
 - mehrere möglich
- Eigenwerte λ (aus den Eigenvektoren)
- Berechnung komplex und unnötig, also per R-Skript

```
A <- H %*% solve(E)
ev <- eigen(A)
ev$values
[1] 0.432432 -8.673617e-19 <---Eigenwerte der beiden Variaten
```

Eigenwerte entsprechen konzeptionell F-Werten bei ANOVA, müssen aber noch transformiert werden ...

Transformation der Eigenwerte

- Pillai-Bartlett Trace
- Hotelling's T^2
- Wilk's Lambda
- Roy's Largest Root

Pillai(-Bartlett) Trace

- $V = \sum_{i=1}^s \frac{\lambda_i}{1+\lambda_i}$ mit s =Anzahl der Variaten
- entspricht $\frac{\text{Erklärte Variation}}{\text{Totale Variation}} = \frac{SQE}{SQT} = R^2$

Beispiel:

Eigenwerte : 0.432 und 0

$$\frac{0.432}{1.432} + 0 = 0.302$$

Hotelling's T^2

- $T^2 = \sum_{i=1}^s \lambda_i$ mit $s = \text{Anzahl der Variaten}$
- entspricht $\frac{\text{Erklärte Variation}}{\text{Unerklärte Variation}} = \frac{SQE}{SQR} = F$

Beispiel:

Eigenwerte : 0.432 und 0

$$0.432 + 0 = 0.432$$

Wilk's Lambda

- $\Delta = \prod_{i=1}^s \frac{1}{1+\lambda_i}$ mit s =Anzahl der Variaten
- entspricht $\frac{\text{Unerklärte Variation}}{\text{Totale Variation}} = \frac{SQR}{SQT}$
- Kleine Werte zeigen höhere Signifikanz

Beispiel:

Eigenwerte : 0.432 und 0

$$\frac{1}{1.432} * \frac{1}{1} = 0.698$$

Roy's Largest Root

- $\Theta = \max(\lambda)$
- manchmal auch $\Theta = \frac{\max(\lambda)}{1+\max(\lambda)}$, aber nicht in R
- entspricht $\frac{\text{Erklärte Variation}}{\text{Unerklärte Variation}} = \frac{SQE}{SQR} = F$ der ersten (einflussreichsten) Variate
- Oft am aussagekräftigsten, da es den maximalen Effekt beschreibt

Beispiel:

Eigenwerte : 0.432 und 0

→ 0.432

Aussagekraft

- Bei kleinen Stichproben wenig Unterschied
- Wenn die erste Variate sehr viel größer ist \rightarrow Roy $>$ Hotelling $>$ Wilk $>$ Pillai
- Wenn Effekte sich eher gleichmäßig verteilen \rightarrow Roy $<$ Hotelling $<$ Wilk $<$ Pillai
- Olson (1974): *Comparative robustness of six tests in multivariate analysis of variance*
- Olson (1976): *On choosing a test statistic in multivariate analysis of variance*
- Olson (1979): *Practical considerations in choosing a MANOVA test statistic*
- Generell weniger als 10 Outcomegruppen empfehlenswert
- Stevens (1980): *Power of the multivariate analysis of variance*

Grundannahmen

Annahmen von ANOVA plus

- Unabhängige Beobachtungen
- Randomisierte Stichproben
- mindestens intervallskaliert Daten
- Multivariate Normalverteilung
 - Outcomes in Gruppen normalverteilt
 - → Multivariater Shapiro Test
- Homogenität der Varianz-Kovarianz Matrix
 - → Homogene Korrelationen sämtlicher Paare von Outcomegruppen und homogene Varianzen der Outcomegruppen
 - Box's Test nicht signifikant → Gut
 - aber Box's Test gilt als unzuverlässig
 - bei gleichen Gruppengrößen (und v.a. 2 Gruppen) kann Hotelling und Pillai eher robust angesehen werden
 - bei unterschiedlichen Gruppengrößen könnte man zufällige Einträge in den größeren Gruppen löschen

Robustheit

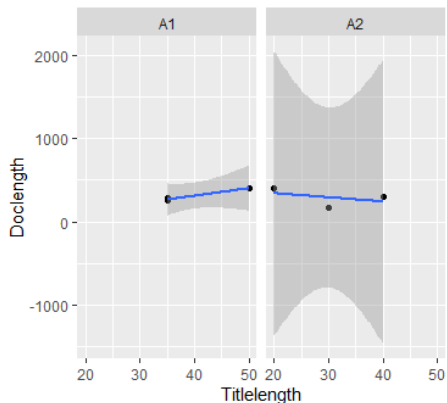
- Relativ robust gegenüber Verletzung der Multivariaten Normalverteilung
- Roy nicht robust gegenüber heterogenen Varianz-Kovarianz Matrizen
- Bei gleichen Gruppengrößen Pillai am robustesten
- ... sonst Pillai anfällig für heterogene Varianz-Kovarianz Matrizen und Verletzung der Multivariaten Normalverteilung

Daumenregel:

Achte auf homogene Varianz-Kovarianz Matrizen und multivariate Normalverteilung und verwende Pillai (oder Roy wenn fast nur 1 Variate Einfluss hat)

MANOVA in R

```
library(ggplot2)
Group<-gl(2, 3, labels = c("A1", "A2"))
Titlelength<-c(35,35,50,20,30,40)
Doclength<-c(250,280,400,400,170,300)
df<-data.frame(Group, Titlelength, Doclength)
scatter <- ggplot(df, aes(Titlelength, Doclength))
scatter + geom_point() + geom_smooth(method = "lm") + facet_wrap(~Group, ncol = 2)
```



MANOVA in R

```
library(pastecs)
Group<-gl(2, 3, labels = c("A1", "A2"))
Titlelength<-c(35,35,50,20,30,40)
Doclength<-c(250,280,400,400,170,300)
df<-data.frame(Group, Titlelength, Doclength)
by(df[,2:3],df$Group, cov)
```

df\$Group: A1

	Titlelength	Doclength	
Titlelength	75	675	//Diagonal Varianzen
Doclength	675	6300	//Nichtdiagonal Kovarianzen
-----//der Outcomes nach Gruppen			

df\$Group: A2

	Titlelength	Doclength
Titlelength	100	-500
Doclength	-500	13300

Die Werte sind hier stark unterschiedlich (aber immerhin sind die Gruppennzahlen gleich)

MANOVA in R

```
library(mvnormtest)
Group<-gl(2, 3, labels = c("A1", "A2"))
Titlelength<-c(35,35,50,20,30,40)
Doclength<-c(250,280,400,400,170,300)
df<-data.frame(Group, Titlelength, Doclength)
a1t<-t(df[1:3, 2:3])
a2t<-t(df[4:6, 2:3])
mshapiro.test(a1t)
mshapiro.test(a2t)
```

Shapiro-Wilk normality test

data: Z

W = 0.8, p-value <2e-16

Shapiro-Wilk normality test

data: Z

W = 0.8, p-value <2e-16 //Deutliche Abweichung von multivariater Normalverteilung -

MANOVA in R

```
...df Siehe vorher
outcome<-cbind(df$titlelength, df$doclength)
model<-manova(outcome ~ Group, data = df)
summary(model, intercept = TRUE)
summary(model, intercept = TRUE, test = "Wilks")
summary(model, intercept = TRUE, test = "Hotelling")
summary(model, intercept = TRUE, test = "Roy")
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
(Intercept)	1	0.96955	47.768	2	3	0.005312	**
Group	1	0.30189	0.649	2	3	0.583296	//Default = Pillai
Residuals	4						
Group	1	0.69811	0.649	2	3	0.583296	//Wilks
Group	1	0.432	0.649	2	3	0.583296	//Hotelling-Lawley
Group	1	0.432	0.649	2	3	0.583296	//Roy



Und es kamen tatsächlich dieselben Werte heraus

Zusammenfassung

- MANOVA ist ein multi-multiples Regressionsmodell mit vertauschtem Outcome und Prädiktoren zur Untersuchung der Outcomes auf signifikante Unterschiede
- Berechnung analog zu ANOVA aber über Matrizen
- Eigenvektoren und -werte
- Transformation mittels Pillai-Bartlett Trace, Hotelling's T^2 , Wilk's Lambda, Roy's Largest Root
- Grundannahmen, insbesondere homogene Varianz-Kovarianz Matrizen und multivariate Normalverteilung
- Folgeanalysen mittels ANOVA oder Diskriminantenanalyse (Siehe Moodle)
- Aufwand und Fehlerpotential nicht unterschätzen
- Robust: Wilcox (2005)