

Übungsaufgaben "Multivariate Verfahren"

Joram Soch

20. Oktober 2026

Inhaltsverzeichnis

I	Aufgaben	1
1	Einführung	2
1.1	Variablendefinition	2
1.2	Beispieldatensatz	2
1.3	Beispielstudie	2
1.4	Parameterschätzer	2
2	Matrizen	4
2.1	Spur	4
2.2	Determinante	4
2.3	Inverse	4
2.4	Probe	4
2.5	Matrixoperationen	4
3	Eigenanalyse	5
3.1	Eigenwerte	5
3.2	Eigenvektoren	5
3.3	KI-Prompt	5
4	Multivariate Normalverteilungen	6
4.1	Beispielverteilung	6
4.2	Marginale Normalverteilungen	6
4.3	Gemeinsame Normalverteilung	6
4.4	Korrelationskoeffizient	7
5	Multivariate Deskriptivstatistik	8
5.1	Beispielstatistiken	8
5.2	Stichprobenkorrelationsmatrix	8
5.3	Linear-affine Transformation	8
5.4	Mahalanobis-Distanz	9
6	Multivariate Varianzanalyse	10
6.1	Voraussetzungen	10
6.2	Beispielstudie	10
6.3	Beispielanalyse	10
6.4	R-Funktion	10
6.5	Spezialfall	10
6.6	KI-Frage	10
7	Kanonische Korrelation	11
7.1	Korrelation und Kovarianz	11
7.2	Stichprobenkorrelationen	11
7.3	Beispielstudie	11
7.4	Ergebnisinterpretation	12

	7.5	Variablenauswahl	12
8		Prädiktive Modellierung	13
	8.1	Beispielstudie	13
	8.2	Konfusionsmatrix	13
	8.3	Performanzmetriken	13
	8.4	Balanced Accuracy	13
II		Lösungen	14
1		Einführung	15
	1.1	Variablendefinition	15
	1.2	Beispieldatensatz	15
	1.3	Beispielstudie	15
	1.4	Parameterschätzer	16
2		Matrizen	17
	2.1	Spur	17
	2.2	Determinante	17
	2.3	Inverse	17
	2.4	Probe	17
	2.5	Matrixoperationen	18
3		Eigenanalyse	19
	3.1	Eigenwerte	19
	3.2	Eigenvektoren	20
	3.3	KI-Prompt	21
4		Multivariate Normalverteilungen	22
	4.1	Beispielverteilung	22
	4.2	Marginale Normalverteilungen	22
	4.3	Gemeinsame Normalverteilung	23
	4.4	Korrelationskoeffizient	23
5		Multivariate Deskriptivstatistik	24
	5.1	Beispielstatistiken	24
	5.2	Stichprobenkorrelationsmatrix	24
	5.3	Linear-affine Transformation	25
	5.4	Mahalanobis-Distanz	26
6		Multivariate Varianzanalyse	27
	6.1	Voraussetzungen	27
	6.2	Beispielstudie	28
	6.3	Beispielanalyse	28
	6.4	R-Funktion	28
	6.5	Spezialfall	29
	6.6	KI-Frage	30
7		Kanonische Korrelation	31
	7.1	Korrelation und Kovarianz	31
	7.2	Stichprobenkorrelationen	31
	7.3	Beispielstudie	31
	7.4	Ergebnisinterpretation	32
	7.5	Variablenauswahl	33
8		Prädiktive Modellierung	34
	8.1	Beispielstudie	34

8.2	Konfusionsmatrix	34
8.3	Performanzmetriken	34
8.4	Balanced Accuracy	35

Kapitel I

Aufgaben

1 Einführung

1.1 Variablendefinition

Erläutern Sie, was man in der psychologischen Forschung unter einer unabhängigen Variable bzw. einer abhängigen Variable versteht. Geben Sie jeweils ein Beispiel.

1.2 Beispieldatensatz

Betrachten Sie den in der ersten Seminarsitzung vorgestellten Datensatz in der Datei `FADE_SAME.csv` sowie die in ihm enthaltenen Variablen. Welche dieser Variablen würden eher als unabhängige und welche würden eher als abhängige Variablen betrachtet werden?

1.3 Beispielstudie

In einer Studie werden Menschen im Hinblick auf ihre Schlafqualität untersucht. Hierfür werden Angestellte mit Schichtarbeitsverhältnis sowie solche mit festen Tagesarbeitszeiten in die Kohorte aufgenommen. Von jeder Versuchsperson werden zunächst die durchschnittliche Schlafdauer, die durchschnittliche Wochenarbeitszeit und das letzte Nettomonatsgehalt erfragt. Anschließend werden im Schlaflabor die Anzahl der Schlafzyklen und die Gesamtdauer des REM-Schlafs in einer Nacht bestimmt. Welche dieser Variablen würden eher als unabhängige und welche würden eher als abhängige Variablen betrachtet werden?

1.4 Parameterschätzer

Gegeben sei ein Allgemeines Lineares Modell $y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ für n Versuchspersonen, wobei der Datenvektor $y \in \mathbb{R}^n$ gemessene Werte der Variable „durchschnittliche Schlafdauer“ beinhaltet und die Designmatrix $X \in \mathbb{R}^{n \times 3}$ gegeben ist als

$$X = \begin{pmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & \vdots \\ 1 & 1 & \vdots \\ \vdots & \vdots & \vdots \\ 1 & 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 3}, \quad (1)$$

wobei die erste Spalte von $X \in \mathbb{R}^n$ dem Einsvektor 1_n entspricht, die zweite Spalte von X genau dann 1 ist, wenn die zugehörige Versuchsperson ein Schichtarbeitsverhältnis hat (und 0 anderenfalls), und die dritte Spalte korrespondierende Werte der (Mittelwert-zentrierten) Variable „durchschnittliche Wochenarbeitszeit“ beinhaltet. Mit dem Betaparametervektor

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^3 \quad (2)$$

ergibt sich für einen einzelnen Datenpunkt die Gleichung

$$y_i = \beta_0 + z_i \beta_1 + x_i \beta_2 + \varepsilon_i \quad \text{mit} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{und} \quad z_i \in \{0, 1\} \quad , \quad (3)$$

wobei $z_i = 1$ ein Schichtarbeitsverhältnis anzeigt (und 0 keines).

Im Rahmen der Modellschätzung ergeben sich folgende Parameterschätzer:

$$\hat{\beta} = \begin{pmatrix} 8.01 \\ -1.17 \\ -0.08 \end{pmatrix} . \quad (4)$$

Wie sind diese Parameterschätzer zu interpretieren?

2 Matrizen

Gegeben sei die Matrix

$$A = \begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}. \quad (5)$$

2.1 Spur

Berechnen Sie die Spur von A .

2.2 Determinante

Berechnen Sie die Determinante von A .

2.3 Inverse

Berechnen Sie die Inverse von A .

2.4 Probe

Überprüfen Sie die Inverse auf ihre Richtigkeit.

2.5 Matrixoperationen

Gegeben seien eine Datenmatrix $Y \in \mathbb{R}^{n \times 2}$, deren n Zeilen n experimentellen Einheiten (z.B. Versuchspersonen) entsprechen und deren zwei Spalten zwei gemessenen Variablen entsprechen. Gegeben sei des Weiteren die Matrix

$$M = \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & -1 & \frac{1}{2} \end{pmatrix} \in \mathbb{R}^{2 \times 3}. \quad (6)$$

Beschreiben Sie in Ihren eigenen Worten die Bedeutung der Matrix YM . Worin bestehen die Zeilen dieser Matrix? Worin bestehen die Spalten dieser Matrix?

3 Eigenanalyse

Gegeben sei die Matrix

$$A = \begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}. \quad (7)$$

3.1 Eigenwerte

Berechnen Sie die Eigenwerte von A .

3.2 Eigenvektoren

Berechnen Sie die zugehörigen Eigenvektoren.

3.3 KI-Prompt

Entwickeln Sie einen Prompt für ein Large Language Model (z.B. OpenAI ChatGPT oder Mistral Vibe), mit dem Sie eine weitere Übung wie in den Aufgaben 2.1–2.4 bzw. 3.1–3.2 erzeugen können. Die zu bearbeitende Matrix soll nicht symmetrisch sein und ganzzahlige Eigenwerte haben.

4 Multivariate Normalverteilungen

4.1 Beispielverteilung

Gegeben sei der dreidimensionale Zufallsvektor $\xi \in \mathbb{R}^3$, wobei ξ_1 dem Gewicht eines Menschen bei Geburt [kg], ξ_2 dem Intelligenzquotienten des Menschen im Alter von 10 Jahren [IQ] und ξ_3 dem Nettomonatseinkommen des Menschen im Alter von 25 Jahren [T€] entsprechen soll.

Der Zufallsvektor $\xi \in \mathbb{R}^3$ sei multivariat normalverteilt

$$\xi \sim \mathcal{N}(\mu, \Sigma) \quad (8)$$

mit dem Erwartungswertparameter

$$\mu = \begin{pmatrix} 3.5 \\ 100 \\ 2.783 \end{pmatrix} \in \mathbb{R}^3 \quad (9)$$

und dem Kovarianzmatrixparameter

$$\Sigma = \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 225 & 4.04 \\ 0 & 4.04 & 1.27 \end{pmatrix} \in \mathbb{R}^{3 \times 3}. \quad (10)$$

Welche Informationen lassen sich hieraus ableiten?

4.2 Marginale Normalverteilungen

Geben Sie die marginalen Verteilungen von Geburtsgewicht, Intelligenzquotient und Nettomonatseinkommen an, die sich aus der Definition der Beispielverteilung in der vorherigen Aufgabe ergeben. Wenden Sie dazu das Theorem zu marginalen Normalverteilungen aus der Vorlesung an.

4.3 Gemeinsame Normalverteilung

Gegeben seien die Zufallsvariablen $X \in \mathbb{R}$ und $Y \in \mathbb{R}$, wobei X die Körpergröße eines Menschen [kg] und Y das Körpergewicht des Menschen [cm] darstellen soll.

Körpergröße sei marginal normalverteilt mit

$$X \sim \mathcal{N}(170, 10^2) \quad (11)$$

und Körpergewicht sei bedingt normalverteilt als

$$Y \sim \mathcal{N}(X - 100, 12^2). \quad (12)$$

Ermitteln Sie die gemeinsame Verteilung der Zufallsvariablen X und Y , d.h. Verteilung des Zufallsvektors $(X, Y)^T$. Wenden Sie dazu das Theorem zu gemeinsamen Normalverteilungen aus der Vorlesung an.

4.4 Korrelationskoeffizient

Berechnen Sie die Korrelation von Körpergröße und Körpergewicht, die sich aus der gemeinsamen Verteilung in der vorherigen Aufgabe ergibt. Wenden Sie dazu (i) das Theorem zur Kovarianzmatrix der multivariaten Normalverteilung, (ii) das Theorem zu den Eigenschaften der Kovarianzmatrix und (iii) die Definition der Korrelationsmatrix eines Zufallsvektors an.

5 Multivariate Deskriptivstatistik

5.1 Beispielstatistiken

Für das in Aufgabe 4.1 skizzierte Beispiel (Geburtsgewicht, Intelligenzquotient, Nettomonatseinkommen) sei $y_1, \dots, y_n \in \mathbb{R}^m$ ein empirischer Datensatz.

Das multivariate Stichprobenmittel beträgt

$$\bar{y} = \begin{pmatrix} 3.610 \\ 103.407 \\ 2.911 \end{pmatrix} \in \mathbb{R}^3 \quad (13)$$

und die Stichprobenkovarianzmatrix beträgt

$$C = \begin{pmatrix} 0.21 & 0.81 & -0.06 \\ 0.81 & 197.77 & 3.82 \\ -0.06 & 3.82 & 1.57 \end{pmatrix} \in \mathbb{R}^{3 \times 3}. \quad (14)$$

Welche Informationen lassen sich hieraus ableiten?

5.2 Stichprobenkorrelationsmatrix

Ermitteln Sie die Stichprobenkorrelationsmatrix, die sich aus der Stichprobenkovarianzmatrix in der vorherigen Aufgabe ergibt. Wenden Sie dazu das Theorem zu Datenmatrix und multivariaten Deskriptivstatistiken aus der Vorlesung an.

5.3 Linear-affine Transformation

Es sei $\{y_1, \dots, y_n\}$ mit $y_i \in \mathbb{R}^m$ für $i = 1, \dots, n$ ein multivariater Datensatz und es sei $\{\tilde{y}_1, \dots, \tilde{y}_n\}$ mit $\tilde{y}_i = Ay_i + b$ für $i = 1, \dots, n$ eine linear-affine Transformation dieses Datensatzes mit einer $k \times m$ Matrix A und einem k -dimensionalen Vektor b .

Zeigen Sie, dass das multivariate Stichprobenmittel derselben Transformation unterliegt:

$$\tilde{\bar{y}} = A\bar{y} + b. \quad (15)$$

Begründen Sie weiterhin, weshalb sich die Stichprobenkorrelationsmatrix unter linear-affiner Transformation nicht verändert.

5.4 Mahalanobis-Distanz

Gegeben seien zwei Zufallsvektoren $x, y \in \mathbb{R}^m$ und eine Kovarianzmatrix $C \in \mathbb{R}^{m \times m}$. Zeigen Sie, dass die Mahalanobis-Distanz von x und y hinsichtlich C mit der quadrierten Euklidischen Distanz von x und y identisch ist, wenn C der Einheitsmatrix I_m entspricht:

$$D = \left(\sqrt{\sum_{j=1}^m (x_j - y_j)^2} \right)^2. \quad (16)$$

6 Multivariate Varianzanalyse

6.1 Voraussetzungen

Gegeben sei das Modell der multivariaten Varianzanalyse (MANOVA) mit $m = 2$ Variablen und $p = 3$ Gruppen. Stellen Sie mithilfe von 2D-Konturplots zwei Szenarien dar: eines, in dem die Annahmen der MANOVA erfüllt sind, und eines, in dem sie verletzt werden. Erläutern Sie, welche Annahme im zweiten Szenario nicht erfüllt ist.

6.2 Beispielstudie

Für das in Aufgabe 4.1 skizzierte Beispiel (Geburtsgewicht, Intelligenzquotient, Nettomonatseinkommen) wurde ein empirischer Datensatz für zwei Gruppen (47 Psychologie-Studierende, 68 Mathematik-Studierende) erhoben. Geben Sie n , m und p an.

6.3 Beispielanalyse

Für den in der vorherigen Aufgabe beschriebenen Datensatz wird die Nullhypothese der multivariaten Varianzanalyse verworfen ($\Lambda = 0.89$, $\tau = 4.57$, $p = 0.005$). Was lässt sich daraus schlussfolgern?

6.4 R-Funktion

Entwickeln Sie eine R-Funktion, die mithilfe des Theorems zu approximativen H_0 -Verteilungen von Λ -Transformationen aus einer Wilks'- Λ -Statistik den p-Wert für eine multivariate Varianzanalyse berechnet.

6.5 Spezialfall

Gegeben sei das Szenario der multivariaten Varianzanalyse (MANOVA) mit $m = 1$ Variable und $p \geq 2$ Gruppen. Zeigen Sie, dass die Wilks'- Λ -Statistik in diesem Fall mit der F-Statistik der univariaten Varianzanalyse identisch ist. Wenden Sie dazu das Theorem zu speziellen H_0 -Verteilungen von Λ -Transformationen an.

6.6 KI-Frage

Stellen Sie einem Large Language Model (z.B. OpenAI ChatGPT oder Mistral Vibe) folgende Frage und geben Sie die Antwort in Ihren eigenen Worten wieder: „Aus der Ablehnung einer MANOVA-Nullhypothese kann man ableiten, dass sich der Erwartungswert mindestens einer Variable zwischen mindestens zwei Gruppen unterscheidet. Aber dadurch weiß ich ja noch gar nichts! Warum kann ich nicht eine separate univariate Varianzanalyse für jede Variable oder einen separaten Zweistichproben-T-Test für jede Variable und jedes Gruppenpaar durchführen? Ich bitte um eine kurze Antwort.“

7 Kanonische Korrelation

7.1 Korrelation und Kovarianz

Gegeben seien zwei Zufallsvariablen ξ und v mit der Kovarianz $\mathbb{C}(\xi, v) = -7.48$ und der Korrelation $\rho(\xi, v) = -0.67$. Berechnen Sie Kovarianz und Korrelation von ξ und $\zeta = 100 - v$. Wenden Sie dazu das Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen an.

7.2 Stichprobenkorrelationen

Gegeben seien Datensätze $\{(x_1, y_1), \dots, (x_n, y_n)\}$ mit folgenden Stichprobenumfängen, Stichprobenvarianzen und Stichprobenkovarianzen. Berechnen Sie die zugehörigen Stichprobenkorrelationen.

Nr.	n	c_{xy}	s_x^2	s_y^2	r_{xy}
1	12	5.37	4.21	9.43	
2	88	-0.87	1.21	1.57	
3	176	13.78	98.70	66.25	
4	768	-8.99	55.78	36.01	

7.3 Beispielstudie

Gegeben sei eine Stichprobe ($n = 768$) von Versuchspersonen mit (i) Testscores zur Erfassung der *Big Five* (OCEAN) mittels des *Big Five Inventory 2* (BFI-2) sowie (ii) Messwerten für objektiven Berufserfolg in Form von (a) Bruttomonatseinkommen, (b) Anzahl bisheriger Beförderungen und (c) Anzahl unterstellter Mitarbeiter.

Beurteilen Sie, welcher dieser Variablensätze eher den unabhängigen Variablen (UV) und welcher dieser Variablensätze eher den abhängigen Variablen (AV) im Rahmen einer Kanonischen Korrelationsanalyse entspricht.

Identifizieren Sie dann m_x und m_y sowie die Einträge von $x \in \mathbb{R}^{m_x}$ und $y \in \mathbb{R}^{m_y}$. Schreiben Sie weiterhin die Grundgleichung dieser Kanonischen Korrelationsanalyse mit den zugrundeliegenden Linearkombinationen auf.

7.4 Ergebnisinterpretation

Eine Kanonische Korrelationsanalyse zu der in der vorherigen Aufgabe beschriebenen Beispielstudie resultiert in folgender maximaler kanonischer Korrelation und folgenden geschätzten kanonischen Koeffizientenvektoren:

```
rho_hat_1 : 0.42
a_hat_1   : 0.10  0.66  0.41 -0.10 -0.36
b_hat_1   : 0.55  0.42  0.32
```

Was lässt sich aus diesen Ergebnissen ableiten?

7.5 Variablenauswahl

Betrachten Sie den in der ersten Seminarsitzung vorgestellten Datensatz in der Datei `FADE_SAME.csv`. Wenn man die Variablen zum Hippocampus-Volumen (`V.HC...`, `V.left...` sowie `V.right...`) einmal außen vor lässt – wie könnte eine passende Kanonische Korrelationsanalyse aussehen?

8 Prädiktive Modellierung

8.1 Beispielstudie

Ein Forschungsteam möchte mithilfe von Fitness-Tracker-Daten vorhersagen, ob eine Person mit SARS-CoV-2 infiziert ist. Hierzu werden die durchschnittliche Schlafdauer, die tägliche Schrittzahl, die Ruheherzfrequenz und die nächtliche Atemfrequenz erhoben. Für 453 Personen liegen zusätzlich die Ergebnisse eines PCR-Tests im Sinne des tatsächlichen Infektionsstatus vor.

Identifizieren Sie für diese Studie die Bedeutung der Labelvariable y , die Einträge des Featurevektors x , die Anzahl der Featurevariablen m und die Anzahl der Datenpunkte n .

8.2 Konfusionsmatrix

Für die in der vorherigen Aufgabe beschriebene Beispielstudie liegt ein binärer Klassifikationsdatensatz vor. Durch prädiktive Modellierung mittels *leave-one-out cross-validation* (LOO-CV) ergeben sich prädizierte Labels $\hat{y}_i = f(x_i)$ für alle Datenpunkte $i = 1, \dots, n$. Vervollständigen Sie die folgende Konfusionsmatrix:

	$f(x_i) = 0$	$f(x_i) = 1$	
$y_i = 0$		45	
$y_i = 1$			402
	15		

8.3 Performanzmetriken

Gegeben sei die vervollständigte Konfusionsmatrix aus der vorherigen Aufgabe. Berechnen und interpretieren Sie auf Grundlage dieser Zahlen die *true positive rate* (TPR) a.k.a. „Sensitivität“, die *true negative rate* (TNR) a.k.a. „Spezifizität“, den *positive predictive value* (PPV) und den *negative predictive value* (NPV).

8.4 Balanced Accuracy

Gegeben sei die vervollständigte Konfusionsmatrix aus der vorherigen Aufgabe. Berechnen Sie auf Grundlage dieser Zahlen die Klassifikationsgenauigkeit (ACC) und die balancierte Klassifikationsgenauigkeit (BAC) a.k.a. „balanced accuracy“. Erklären Sie am Beispiel des vorliegenden Falls, warum die Klassifikationsgenauigkeit (ACC) zu irreführenden Ergebnissen führen kann und wie die balancierte Klassifikationsgenauigkeit (BAC) ein plausibleres Maß für prädiktive Performanz darstellt.

Kapitel II

Lösungen

1 Einführung

1.1 Variablendefinition

Unter einer unabhängigen Variable versteht man eine Größe, die Experimentierende typischerweise unter ihrer Kontrolle haben und die im Verlaufe des Experiments variiert wird, um ihren Einfluss auf andere Größen zu ermitteln. Ein Beispiel hierfür ist „experimentelle Bedingung“, z.B. kongruente vs. inkongruente Distraktoren im Eriksen-Flanker-Test.

Unter einer abhängigen Variable versteht man eine Größe, die Experimentierende typischerweise messen und die das primäre Ergebnismaß des Experiments darstellt, das ggf. von anderen Größen beeinflusst wird. Ein Beispiel hierfür ist „Reaktionszeit“, d.h. Zeitdifferenz zwischen Stimuluspräsentation und Tastendruck durch eine Versuchsperson.

1.2 Beispieldatensatz

Unabhängige Variablen dieser Studie sind „Alter“ bzw. „Altersgruppe“ (weil gezielt Versuchspersonen aus den Altersgruppen 18-35 bzw. 50+ rekrutiert wurden), „Geschlecht“ (indem annähernd gleich viele weibliche und männliche Versuchspersonen gemessen wurden) und „Scanner“ (da Experimente mit Versuchspersonen an zwei verschiedenen MRT-Scannern durchgeführt wurden).

Abhängige Variablen dieser Studie sind die fMRT-Scores „novelty-FADE“, „novelty-SAME“, „memory-FADE“ und „memory-SAME“, kognitive Leistungsmaße wie „memory“ (Gedächtnisleistung) und „MWT-B“ (kristalline Intelligenz) sowie sämtliche Hippocampus-Volumen-Variablen.

Die Einordnung der Variable „Abitur“ ist diskussionswürdig. Zwar wäre es prinzipiell möglich, das Vorliegen einer Hochschulzugangsberechtigung als unabhängige Variable bei der Versuchspersonenrekrutierung zu kontrollieren – insbesondere um für ein ausgeglichenes Bildungsniveau zwischen Altersgruppen zu sorgen –, faktisch ist dies aber nicht erfolgt, sodass hier eher von einer abhängigen Variable auszugehen ist.

1.3 Beispielstudie

Eine mögliche Einteilung wäre etwa:

unabhängige Variablen	abhängige Variablen
Schichtarbeitsverhältnis (ja/nein)	durchschnittliche Schlafdauer
durchschnittliche Wochenarbeitszeit	Anzahl der Schlafzyklen
letztes Nettomonatsgehalt	Gesamtdauer des REM-Schlafs

1.4 Parameterschätzer

Als Vorhersage für einen einzelnen Datenpunkt ergibt sich:

$$\hat{y}_i = \hat{\beta}_0 + z_i \hat{\beta}_1 + x_i \hat{\beta}_2 \quad \text{mit} \quad z_i \in \{0, 1\} . \quad (1)$$

Für das in der Aufgabenstellung beschriebene Modell entspricht also

- $\hat{\beta}_0 = 8.01$ der geschätzten mittleren Schlafdauer einer Versuchsperson mit mittlerer Wochenarbeitszeit ($x_i = 0$) in der Referenzgruppe ($z_i = 0$) derjenigen ohne Schichtarbeitsverhältnis (durchschnittlich 8.01 h Schlaf pro Nacht ohne Schichtarbeit);
- $\hat{\beta}_1 = -1.17$ der geschätzten veränderten Schlafdauer der Versuchspersonen mit mittlerer Wochenarbeitszeit ($x_i = 0$) in der Nicht-Referenzgruppe ($z_i = 1$) derjenigen mit Schichtarbeitsverhältnis (durchschnittlich 1.17 h weniger Schlaf pro Nacht bei Schichtarbeit);
- $\hat{\beta}_2 = -0.08$ dem geschätzten Einfluss der mittleren Wochenarbeitszeit auf die mittlere Schlafdauer unabhängig von der Zugehörigkeit zu einer Versuchspersonengruppe (durchschnittlich 0.08 h weniger Schlaf pro 1 h mehr Wochenarbeitszeit).

2 Matrizen

Wir bezeichnen die Einträge der Matrix A mit

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (2)$$

2.1 Spur

$$\operatorname{tr}(A) = a_{11} + a_{22} = 6 + 2 = 8. \quad (3)$$

2.2 Determinante

$$\det(A) = a_{11}a_{22} - a_{12}a_{21} = 6 \cdot 2 - 1 \cdot (-3) = 12 + 3 = 15. \quad (4)$$

2.3 Inverse

$$\begin{aligned} A^{-1} &= \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \\ &= \frac{1}{6 \cdot 2 - 1 \cdot (-3)} \begin{pmatrix} 2 & -1 \\ 3 & 6 \end{pmatrix} = \frac{1}{15} \begin{pmatrix} 2 & -1 \\ 3 & 6 \end{pmatrix} \\ &= \begin{pmatrix} 2/15 & -1/15 \\ 3/15 & 6/15 \end{pmatrix} = \begin{pmatrix} 2/15 & -1/15 \\ 1/5 & 2/5 \end{pmatrix}. \end{aligned} \quad (5)$$

2.4 Probe

Für die Inverse muss gelten, dass $AA^{-1} = I_n$. Wir überprüfen:

$$\begin{aligned} AA^{-1} &= \begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 2/15 & -1/15 \\ 3/15 & 6/15 \end{pmatrix} \\ &= \begin{pmatrix} 12/15 + 3/15 & -6/15 + 6/15 \\ -6/15 + 6/15 & 3/15 + 12/15 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2. \end{aligned} \quad (6)$$

Analog ergibt sich $A^{-1}A = I_n$. Damit ist die Inverse korrekt.

2.5 Matrixoperationen

Da Y eine $n \times 2$ Matrix ist und M eine 2×3 Matrix ist, ist das Matrixprodukt YM definiert und eine Matrix der Größe $n \times 3$. Evaluation des Matrixprodukts ergibt:

$$\begin{aligned}
 YM &= \begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & -1 & \frac{1}{2} \end{pmatrix} \\
 &= \begin{pmatrix} (y_{11} + y_{12}) & (y_{11} - y_{12}) & (\frac{1}{2}y_{11} + \frac{1}{2}y_{12}) \\ \vdots & \vdots & \vdots \\ (y_{n1} + y_{n2}) & (y_{n1} - y_{n2}) & (\frac{1}{2}y_{n1} + \frac{1}{2}y_{n2}) \end{pmatrix} \\
 &= \begin{pmatrix} (y_{11} + y_{12}) & (y_{11} - y_{12}) & \frac{y_{11}+y_{12}}{2} \\ \vdots & \vdots & \vdots \\ (y_{n1} + y_{n2}) & (y_{n1} - y_{n2}) & \frac{y_{n1}+y_{n2}}{2} \end{pmatrix}
 \end{aligned} \tag{7}$$

Entsprechend gilt: Die Zeilen der Matrix YM entsprechen den n experimentellen Einheiten der Datenmatrix. Die Spalten der Matrix YM beinhalten (i) die Summe der beiden Variablen (erste Spalte), (ii) die Differenz der beiden Variablen (zweite Spalte) und (iii) dem Durchschnitt der beiden Variablen (dritte Spalte).

3 Eigenanalyse

Wir bezeichnen die Einträge der Matrix A mit

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (8)$$

3.1 Eigenwerte

Das charakteristische Polynom lautet

$$\begin{aligned} \chi_A(\lambda) &= |A - \lambda I_n| = \left| \begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} 6-\lambda & 1 \\ -3 & 2-\lambda \end{pmatrix} \right| \\ &= (6-\lambda)(2-\lambda) - 1 \cdot (-3) = 12 - 6\lambda - 2\lambda + \lambda^2 + 3 \\ &= \lambda^2 - 8\lambda + 15. \end{aligned} \quad (9)$$

Das heißt, wir müssen folgende quadratische Gleichung lösen:

$$0 = \lambda^2 - 8\lambda + 15. \quad (10)$$

Lösung der Gleichung mittels p-q-Formel ergibt

$$\begin{aligned} \lambda_{1,2} &= -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} \\ &= -\frac{-8}{2} \pm \sqrt{\left(\frac{-8}{2}\right)^2 - 15} \\ &= 4 \pm \sqrt{(-4)^2 - 15} \\ &= 4 \pm \sqrt{1} \\ &= 4 \pm 1. \end{aligned} \quad (11)$$

Damit sind die Eigenwerte $\lambda_1 = 5$ und $\lambda_2 = 3$.

3.2 Eigenvektoren

Für $\lambda_1 = 5$ ergibt sich:

$$\begin{aligned}
 (A - \lambda_1 I_n)v_1 = 0_n &\Leftrightarrow \left(\begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} - 5 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 &\Leftrightarrow \begin{pmatrix} 1 & 1 \\ -3 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 &\Leftrightarrow \begin{aligned} 1x + 1y &= 0 \\ -3x - 3y &= 0 \end{aligned} \\
 &\Rightarrow v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \text{ ist eine Lösung .}
 \end{aligned} \tag{12}$$

Für $\lambda_2 = 3$ ergibt sich:

$$\begin{aligned}
 (A - \lambda_2 I_n)v_2 = 0_n &\Leftrightarrow \left(\begin{pmatrix} 6 & 1 \\ -3 & 2 \end{pmatrix} - 3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 &\Leftrightarrow \begin{pmatrix} 3 & 1 \\ -3 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
 &\Leftrightarrow \begin{aligned} 3x + 1y &= 0 \\ -3x - 1y &= 0 \end{aligned} \\
 &\Rightarrow v_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ -3 \end{pmatrix} \text{ ist eine Lösung .}
 \end{aligned} \tag{13}$$

Für die Eigenvektoren gilt

$$||v_1|| = ||v_2|| = 1 \quad \text{sowie} \quad Av_1 = \lambda_1 v_1 \quad \text{und} \quad Av_2 = \lambda_2 v_2 . \tag{14}$$

3.3 KI-Prompt

Ein solcher KI-Prompt könnte z.B. wie folgt lauten: „Bitte gib mir eine 2x2-Matrix, für die dann a) Spur, b) Determinante, c) Inverse, d) Probe der Inverse, e) Eigenwerte und f) Eigenvektoren berechnet werden sollen. Die Matrix soll folgende Bedingungen erfüllen: 1. Sie soll nicht symmetrisch sein. 2. Sie soll nur ganzzahlige Einträge haben. 3. Ihre Eigenwerte sollen ganzzahlig sein. (Die Eigenvektoren sollen die Länge 1 haben, dies kann aber mithilfe einer Wurzelnormierung dargestellt werden.) Bitte gib mir die Aufgabenstellung, aber noch nicht die Lösung, bis ich die Aufgabe selbst bearbeitet habe.“

4 Multivariate Normalverteilungen

4.1 Beispielverteilung

Aus den Verteilungsparametern lässt sich ablesen:

- $\mu_1 = 3.5$: Der Erwartungswert von Geburtsgewicht ist 3.5 (3500 g).
- $\mu_2 = 100$: Der Erwartungswert von Intelligenzquotient ist 100 (100 IQ-Punkte).
- $\mu_3 = 2.783$: Der Erwartungswert von Nettomonatseinkommen ist 2.783 (2783 €).
- $\sigma_{11} = 0.25$: Die Varianz von Geburtsgewicht ist 0.25, d.h. die Standardabweichung ist $\sqrt{0.25} = 0.5$ (500 g).
- $\sigma_{22} = 225$: Die Varianz von Intelligenzquotient ist 225, d.h. die Standardabweichung ist $\sqrt{225} = 15$ (15 IQ-Punkte).
- $\sigma_{33} = 1.27$: Die Varianz von Nettomonatseinkommen ist 1.27, d.h. die Standardabweichung ist $\sqrt{1.27} \approx 1.127$ (1127 €).
- $\sigma_{12} = \sigma_{21} = 0$: Die Kovarianz von Geburtsgewicht und Intelligenzquotient ist 0, d.h. diese Variablen sind unkorreliert.
- $\sigma_{13} = \sigma_{31} = 0$: Die Kovarianz von Geburtsgewicht und Nettomonatseinkommen ist 0, d.h. diese Variablen sind unkorreliert.
- $\sigma_{23} = \sigma_{32} = 4.04$: Die Kovarianz von Geburtsgewicht und Intelligenzquotient ist 4.04, d.h. diese Variablen sind positiv korreliert.

4.2 Marginale Normalverteilungen

Mit dem Theorem zu marginalen Normalverteilungen ergibt sich:

$$\begin{aligned}\xi_1 &\sim \mathcal{N}(3.5, 0.25) \\ \xi_2 &\sim \mathcal{N}(100, 225) \\ \xi_3 &\sim \mathcal{N}(2.783, 1.27) .\end{aligned}\tag{15}$$

4.3 Gemeinsame Normalverteilung

Das Theorem zu gemeinsamen Normalverteilungen besagt:

$$\begin{aligned}
 \xi &\sim \mathcal{N}(\mu_\xi, \Sigma_{\xi\xi}) \\
 v|\xi &\sim \mathcal{N}(A\xi + b, \Sigma_{vv}) \\
 \Rightarrow \begin{pmatrix} \xi \\ v \end{pmatrix} &\sim \mathcal{N}(\mu_{\xi,v}, \Sigma_{\xi,v}) \\
 \text{mit } \mu_{\xi,v} &= \begin{pmatrix} \mu_\xi \\ A\mu_\xi + b \end{pmatrix} \\
 \text{und } \Sigma_{\xi,v} &= \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\xi}A^T \\ A\Sigma_{\xi\xi} & \Sigma_{vv} + A\Sigma_{\xi\xi}A^T \end{pmatrix}.
 \end{aligned} \tag{16}$$

In unserem Fall gilt:

$$\xi = X, \quad v = Y, \quad A = 1, \quad b = -100, \quad \mu_\xi = 170, \quad \Sigma_{\xi\xi} = 10^2 = 100, \quad \Sigma_{vv} = 12^2 = 144. \tag{17}$$

Mithin ergibt sich:

$$\begin{aligned}
 \begin{pmatrix} X \\ Y \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 170 \\ 1 \cdot 170 + (-100) \end{pmatrix}, \begin{pmatrix} 100 & 100 \cdot 1 \\ 1 \cdot 100 & 144 + 1 \cdot 100 \cdot 1 \end{pmatrix}\right) \\
 \begin{pmatrix} X \\ Y \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 170 \\ 70 \end{pmatrix}, \begin{pmatrix} 100 & 100 \\ 100 & 244 \end{pmatrix}\right).
 \end{aligned} \tag{18}$$

4.4 Korrelationskoeffizient

Die Kovarianzmatrix der gemeinsamen Verteilung von X und Y ist

$$\Sigma_{XY} = \begin{pmatrix} 100 & 100 \\ 100 & 244 \end{pmatrix}. \tag{19}$$

Die Einträge der Kovarianzmatrix haben folgende Entsprechungen:

$$\Sigma_{XY} = \begin{pmatrix} \mathbb{C}(X, X) & \mathbb{C}(X, Y) \\ \mathbb{C}(Y, X) & \mathbb{C}(Y, Y) \end{pmatrix} = \begin{pmatrix} \mathbb{V}(X) & \mathbb{C}(X, Y) \\ \mathbb{C}(Y, X) & \mathbb{V}(Y) \end{pmatrix}. \tag{20}$$

Mithin ergibt sich für den Korrelationskoeffizienten:

$$\rho_{XY} = \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}} = \frac{100}{\sqrt{100}\sqrt{244}} \approx 0.64. \tag{21}$$

5 Multivariate Deskriptivstatistik

5.1 Beispielstatistiken

Aus den Deskriptivstatistiken lässt sich ablesen:

- $\bar{y}_1 = 3.610$: Das durchschnittliche Geburtsgewicht ist 3.610 (3610 g).
- $\bar{y}_2 = 103.407$: Der durchschnittliche Intelligenzquotient ist etwa 103.4 (ca. 103 IQ-Punkte).
- $\bar{y}_3 = 2.911$: Das durchschnittliche Nettomonatseinkommen ist 2.911 (2911 €).
- $c_{11} = 0.21$, $c_{22} = 197.77$, $c_{33} = 1.57$: Die Stichprobenvarianzen von Geburtsgewicht, Intelligenzquotient und Nettomonatseinkommen sind 0.21, 197.77 und 1.57.
- $c_{11} = 0.21$, $c_{22} = 197.77$, $c_{33} = 1.57$: Die entsprechenden Stichprobenstandardabweichungen sind $\sqrt{0.21} \approx 0.458$ (458 g), $\sqrt{197.77} \approx 14.06$ (ca. 14 IQ-Punkte) und $\sqrt{1.57} \approx 1.253$ (1253 €).
- $c_{12} = 0.81$, $c_{13} = -0.06$, $c_{23} = 3.82$: Die Stichprobenkovarianz von Geburtsgewicht und Intelligenzquotient ist 0.81; diejenige von Geburtsgewicht und Nettomonatseinkommen ist -0.06 ; diejenige von Intelligenzquotient und Nettomonatseinkommen ist 3.82.

5.2 Stichprobenkorrelationsmatrix

Die Diagonalmatrix der inversen Stichprobenstandardabweichungen ist

$$\begin{aligned}
 D &= \text{diag} \left(\frac{1}{\sqrt{c_{11}}}, \dots, \frac{1}{\sqrt{c_{mm}}} \right) \\
 &= \text{diag} \left(\frac{1}{\sqrt{0.21}}, \frac{1}{\sqrt{197.77}}, \frac{1}{\sqrt{1.57}} \right) \\
 &= \begin{pmatrix} \frac{1}{\sqrt{0.21}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{197.77}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{1.57}} \end{pmatrix}.
 \end{aligned} \tag{22}$$

Also ergibt sich die Stichprobenkorrelationsmatrix als

$$\begin{aligned}
 R &= DCD \\
 &= \begin{pmatrix} \frac{1}{\sqrt{0.21}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{197.77}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{1.57}} \end{pmatrix} \begin{pmatrix} 0.21 & 0.81 & -0.06 \\ 0.81 & 197.77 & 3.82 \\ -0.06 & 3.82 & 1.57 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{0.21}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{197.77}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{1.57}} \end{pmatrix} \\
 &= \begin{pmatrix} \frac{0.21}{\sqrt{0.21}\sqrt{0.21}} & \frac{0.81}{\sqrt{0.21}\sqrt{197.77}} & \frac{-0.06}{\sqrt{0.21}\sqrt{1.57}} \\ \frac{0.81}{\sqrt{197.77}\sqrt{0.21}} & \frac{197.77}{\sqrt{197.77}\sqrt{197.77}} & \frac{3.82}{\sqrt{197.77}\sqrt{1.57}} \\ \frac{-0.06}{\sqrt{1.57}\sqrt{0.21}} & \frac{3.82}{\sqrt{1.57}\sqrt{197.77}} & \frac{1.57}{\sqrt{1.57}\sqrt{1.57}} \end{pmatrix} = \begin{pmatrix} 1 & 0.13 & -0.10 \\ 0.13 & 1 & 0.22 \\ -0.10 & 0.22 & 1 \end{pmatrix}.
 \end{aligned} \tag{23}$$

5.3 Linear-affine Transformation

Beweis 1: Mit der Definition des multivariaten Stichprobenmittels gilt:

$$\begin{aligned}
 \bar{\tilde{y}} &= \frac{1}{n} \sum_{i=1}^n \tilde{y}_i \\
 &= \frac{1}{n} \sum_{i=1}^n (Ay_i + b) \\
 &= \frac{1}{n} \sum_{i=1}^n Ay_i + \frac{1}{n} \sum_{i=1}^n b \\
 &= A \left(\frac{1}{n} \sum_{i=1}^n y_i \right) + \frac{1}{n} n \cdot b \\
 &= A\bar{y} + b .
 \end{aligned} \tag{24}$$

Beweis 2: Alternativ gilt mit der Datenmatrix $Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix} \in \mathbb{R}^{m \times n}$, mit der linear-affin transformierten Datenmatrix $\tilde{Y} = AY + b \mathbf{1}_n^T \in \mathbb{R}^{k \times n}$ und mit dem Theorem zu Datenmatrix und multivariaten Deskriptivstatistiken:

$$\begin{aligned}
 \bar{\tilde{y}} &= \frac{1}{n} \tilde{Y} \mathbf{1}_n \\
 &= \frac{1}{n} (AY + b \mathbf{1}_n^T) \mathbf{1}_n \\
 &= \frac{1}{n} AY \mathbf{1}_n + \frac{1}{n} b \mathbf{1}_n^T \mathbf{1}_n \\
 &= A \frac{1}{n} Y \mathbf{1}_n + \frac{1}{n} b n \\
 &= A\bar{y} + b .
 \end{aligned} \tag{25}$$

Begründung: Die Einträge der Stichprobenkorrelationsmatrix sind paarweise Stichprobenkorrelationen. Da sich die Stichprobenkorrelation zweier Variablen unter linear-affiner Transformation nicht ändert (bzw. lediglich ihr Vorzeichen wechselt, vom absoluten Betrag her also gleich bleibt), bleiben auch die Einträge der Stichprobenkorrelationsmatrix unverändert.

5.4 Mahalanobis-Distanz

Die Mahalanobis-Distanz ist definiert als

$$D = (y_1 - y_2)^T C^{-1} (y_1 - y_2) \quad (26)$$

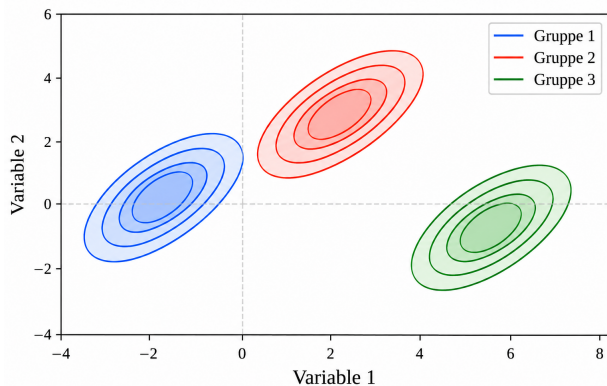
Für das beschriebene Szenario ergibt sich:

$$\begin{aligned} D &= (x - y)^T I_m^{-1} (x - y) \\ &= (x - y)^T I_m (x - y) \\ &= (x - y)^T (x - y) \\ &= \begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_m - y_m \end{pmatrix}^T \begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_m - y_m \end{pmatrix} \\ &= \begin{pmatrix} x_1 - y_1 & \dots & x_m - y_m \end{pmatrix} \begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_m - y_m \end{pmatrix} \\ &= \sum_{j=1}^m (x_j - y_j)^2 \\ &= \left(\sqrt{\sum_{j=1}^m (x_j - y_j)^2} \right)^2. \end{aligned} \quad (27)$$

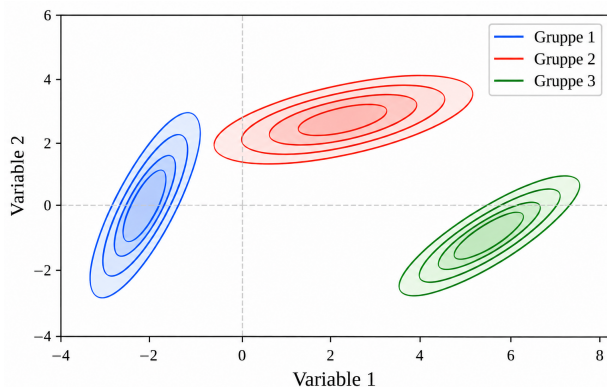
6 Multivariate Varianzanalyse

6.1 Voraussetzungen

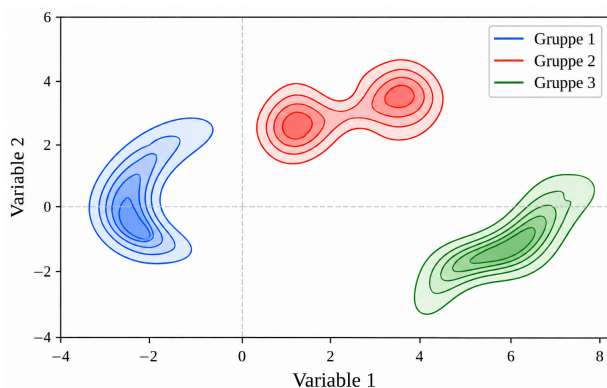
Die folgende Abbildung zeigt eine Situation, in der die Voraussetzungen der multivariaten Varianzanalyse (multivariate Normalverteilung, nicht notwendigerweise identische Erwartungswertparameter, identische Kovarianzmatrixparameter) erfüllt sind:



Die folgende Abbildung zeigt eine Situation, in der die Annahme eines für alle Gruppen identischen Kovarianzmatrixparameters verletzt ist:



Die folgende Abbildung zeigt eine Situation, in der die Annahme der multivariaten Normalverteilung für Datenpunkte des MANOVA-Szenarios verletzt ist:



6.2 Beispielstudie

Es sind:

- $n = n_1 + n_2 = 47 + 68 = 115$.
- $m = 3$ (Anzahl der Variablen).
- $p = 2$ (Anzahl der Gruppen).

6.3 Beispielanalyse

Die Nullhypothese der multivariaten Varianzanalyse lautet:

Für jede Variable gilt: Der Erwartungswertparameter ist für alle Gruppen gleich.

Dem entsprechend lautet die dazugehörige Alternativhypothese:

Für mindestens eine Variable gilt: Es gibt mindestens zwei Gruppen, die sich hinsichtlich des Erwartungswerts dieser Variable unterscheiden.

Also lässt sich aus dem Ergebnis schlussfolgern:

Geburtsgewicht unterscheidet sich zwischen Psychologie- und Mathematik-Studierenden

oder

Intelligenzquotient unterscheidet sich zwischen Psychologie- und Mathematik-Studierenden

oder

Nettomonatseinkommen unterscheidet sich zwischen Psychologie- und Mathematik-Studierenden.

6.4 R-Funktion

```
MANOVA_pval = function(L,n,m,p) {

  # p-Wert berechnen
  w      = n-1-(1/2)*(m+p)                # w
  t      = sqrt((m^2*(p-1)^2-4)/(m^2+(p-1)^2-5))  # t
  nu_1   = m*(p-1)                        # \nu_1
  nu_2   = w*t-(1/2)*(m*(p-1)-2)          # \nu_2
  tau    = ((1-L^(1/t))/L^(1/t))*(nu_2/nu_1)  # \tau
  pval   = 1-pf(tau,nu_1,nu_2)            # p-Wert

  # p-Wert zurückgeben
  return(pval)
}
```

6.5 Spezialfall

Gemäß der dritten Zeile der Tabelle im Theorem zu speziellen H_0 -Verteilungen von Λ -Transformationen errechnet sich die Teststatistik in diesem Szenario als

$$F = \frac{1 - \Lambda}{\Lambda} \cdot \frac{n - p}{p - 1} \quad (28)$$

und folgt unter der Nullhypothese einer f -Verteilung:

$$F \sim f(p - 1, n - p) . \quad (29)$$

Wilks'- Λ -Statistik ist:

$$\Lambda = \frac{|W|}{|B + W|} . \quad (30)$$

Im Fall von $m = 1$ sind $B \in \mathbb{R}^{m \times m}$ und $W \in \mathbb{R}^{m \times m}$ Skalare: B ist identisch mit der *between sum of squares*, W ist identisch mit der *within sum of squares* und ihre Summe ist die *total sum of squares*:

$$\begin{aligned} B &= \sum_{i=1}^p n_i (\bar{y}_i - \bar{y}) (\bar{y}_i - \bar{y})^T \\ &= \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 = \text{SQB} \\ W &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) (y_{ij} - \bar{y}_i)^T \\ &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \text{SQW} \\ B + W &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}) (y_{ij} - \bar{y})^T \\ &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \text{SQT} . \end{aligned} \quad (31)$$

Einsetzen von (31) in (30) und von (30) in (28) ergibt:

$$\begin{aligned}
 F &= \frac{1 - \frac{SQW}{SQB+SQW}}{\frac{SQW}{SQB+SQW}} \cdot \frac{n-p}{p-1} \\
 &= \frac{\frac{SQT}{SQT} - \frac{SQW}{SQT}}{\frac{SQW}{SQT}} \cdot \frac{n-p}{p-1} \\
 &= \frac{\frac{1}{SQT}}{\frac{1}{SQT}} \cdot \frac{SQT - SQW}{SQW} \cdot \frac{n-p}{p-1} \\
 &= \frac{SQB}{SQW} \cdot \frac{n-p}{p-1} \\
 &= \frac{SQB/(p-1)}{SQW/(n-p)} .
 \end{aligned} \tag{32}$$

Hierbei handelt es sich um die Teststatistik der einfaktoriellen Varianzanalyse.

6.6 KI-Frage

Eine mögliche Antwort könnte folgende sein:

Doch, **kannst du** – aber nicht ohne Korrektur für multiples Testen.

Wenn du nach einer signifikanten MANOVA einfach für jede Variable und jedes Gruppenpaar einzelne ANOVAs bzw. t-Tests mit jeweils $\alpha = 0.05$ durchführst, **inflationiert der Fehler 1. Art**: Je mehr Tests, desto größer die Wahrscheinlichkeit mindestens eines falsch-positiven Ergebnisses.

Die MANOVA hat dabei zwei Vorteile: Sie testet **eine gemeinsame globale Nullhypothese** und berücksichtigt die **Korrelationen zwischen den abhängigen Variablen**.

Nach einer signifikanten MANOVA sind univariate ANOVAs und anschließend paarweise t-Tests durchaus typische **Follow-up-Tests** – aber mit geeigneter Korrektur für multiples Testen (z.B. Holm). Kurz:

MANOVA sagt: „Irgendwo im multivariaten Mittelwertsprofil gibt es einen Unterschied.“ **Follow-up-Tests sagen**: „Wo genau liegt er?“

Es ist also durchaus erlaubt, im Anschluss an eine MANOVA-Analyse – **die einen Gruppenunterschied statistisch belegt** – weitere ANOVA-Analysen bzw. T-Tests durchzuführen – **um die Quelle der Gruppenunterschiede zu illustrieren**.

Mit „Holm“ ist hier die Bonferroni-Holm-Prozedur zur Korrektur für multiples Testen gemeint.

7 Kanonische Korrelation

7.1 Korrelation und Kovarianz

Gemäß dem Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen gilt für zwei Zufallsvariablen ξ und v mit Skalaren $\alpha, \beta, \gamma, \delta \in \mathbb{R}$:

$$\begin{aligned}\mathbb{C}(\alpha\xi + \beta, \gamma v + \delta) &= \alpha\gamma\mathbb{C}(\xi, v) \\ |\rho(\alpha\xi + \beta, \gamma v + \delta)| &= |\rho(\xi, v)|.\end{aligned}\tag{33}$$

In unserem Fall gilt $\alpha = 1$ und $\beta = 0$ (da ξ unverändert bleibt) sowie $\gamma = -1$ und $\delta = 100$ (da $\zeta = 100 - v$). Mithin erhalten wir für die Kovarianz:

$$\mathbb{C}(\xi, \zeta) = 1 \cdot (-1) \cdot \mathbb{C}(\xi, v) = 1 \cdot (-1) \cdot (-7.48) = 7.48.\tag{34}$$

Die Korrelation von ξ und ζ bleibt vom Betrag her gleich, ändert aber ihr Vorzeichen, da auch die Kovarianz von ξ und ζ ihr Vorzeichen ändert:

$$\rho(\xi, \zeta) = -\rho(\xi, v) = -(-0.67) = 0.67.\tag{35}$$

7.2 Stichprobenkorrelationen

Nr.	n	c_{xy}	s_x^2	s_y^2	r_{xy}
1	12	5.37	4.21	9.43	0.852
2	88	-0.87	1.21	1.57	-0.631
3	176	13.78	98.70	66.25	0.170
4	768	-8.99	55.78	36.01	-0.201

7.3 Beispielstudie

Als unabhängige Variablen bietet sich der Variablensatz „Persönlichkeit“ an:

- x_1 = Testscore „Openness“;
- x_2 = Testscore „Conscientiousness“;
- x_3 = Testscore „Extraversion“;
- x_4 = Testscore „Agreeableness“;
- x_5 = Testscore „Neuroticism“.

Als abhängige Variablen bietet sich der Variablensatz „Berufserfolg“ an:

- y_1 = Bruttomonatseinkommen;
- y_2 = Anzahl bisheriger Beförderungen;
- y_3 = Anzahl unterstellter Mitarbeiter.

Insgesamt haben wir also:

- $x \in \mathbb{R}^5 \Rightarrow m_x = 5$ (Anzahl unabhängiger Variablen);
- $y \in \mathbb{R}^3 \Rightarrow m_y = 3$ (Anzahl abhängiger Variablen);
- $z \in \mathbb{R}^8 \Rightarrow m = 8$ (Anzahl aller Variablen).

Die Grundgleichung dieser Kanonischen Korrelationsanalyse lautet dann:

$$\begin{aligned} \xi &= a^T x = \begin{pmatrix} a_1 \\ \vdots \\ a_5 \end{pmatrix}^T \begin{pmatrix} x_1 \\ \vdots \\ x_5 \end{pmatrix} = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5 \\ v &= b^T y = \begin{pmatrix} b_1 \\ \vdots \\ b_3 \end{pmatrix}^T \begin{pmatrix} y_1 \\ \vdots \\ y_3 \end{pmatrix} = b_1 y_1 + b_2 y_2 + b_3 y_3 . \end{aligned} \tag{36}$$

7.4 Ergebnisinterpretation

Es lässt sich konstatieren:

- Die erste kanonische Korrelation, d.h. die höchstmögliche Korrelation zwischen einer Linearkombination von Persönlichkeitsmerkmalen $x_1, \dots, x_5$ und einer Linearkombination von Berufserfolgsindikatoren $y_1, \dots, y_3$, ist 0.42. Die resultierende Linearkombination von Persönlichkeitsmerkmalen (der beste Prädiktor) erklärt also $0.42^2 = 0.1764 \approx 17.6\%$ der Varianz der resultierenden Linearkombination von (des am besten prädizierbaren Kriteriums).
- Auf der Seite der Persönlichkeitsmerkmale wird die erste kanonische Variate $\hat{\xi}_1$ vor allem durch hohe Gewissenhaftigkeit, hohe Extraversion und niedrigen Neurotizismus charakterisiert, da entsprechende Einträge von $\hat{a}_1$ von ihrem Betrag her am höchsten sind.
- Auf der Seite der Berufserfolgsindikatoren wird die erste kanonische Variate $\hat{v}_1$ durch hohes Einkommen, aber auch durch Beförderungen und Führungsverantwortung geprägt, da dies die Reihenfolge der Einträge von $\hat{b}_1$ ist, wenn man sie absteigend nach ihrem Betrag sortiert.

7.5 Variablenauswahl

Die fMRT-Scores „novelty-FADE“, „novelty-SAME“, „memory-FADE“ und „memory-SAME“ und die als „memory“ gemessene Gedächtnisleistung sind klarerweise abhängige Variablen und könnten als Variablensatz „neuronale Gedächtnisprozesse“ konzeptionalisiert werden. Als unabhängige Variablen im Rahmen einer Kanonischen Korrelationsanalyse kämen z.B. Alter in Jahren (als Indikator eines voranschreitenden biologischen Prozesses) und die Punktzahl im MWT-B (als Maß des Intelligenzniveaus einer Person) in Frage. In diesem Fall ergäbe sich $m_y = 5$ und $m_x = 2$.

Die Variablen „Geschlecht“ (female/male), „Scanner“ (Skyra/Verio) und „Abitur“ (yes/no) kommen für eine Kanonische Korrelationsanalyse nur eingeschränkt in Betracht, da sie als diskrete Zufallsvariablen mit wenigen möglichen Werten und nur bedingt zur Bildung von Linearkombinationen geeignet sind, für die Korrelationskoeffizienten sinnvoll berechnet werden können.

8 Prädiktive Modellierung

8.1 Beispielstudie

Es ergibt sich:

- $y = 1$: eine Person ist gemäß PCR-Ergebnis mit SARS-CoV-2 infiziert („positiver Fall“);
- $y = 0$: eine Person ist gemäß PCR-Ergebnis nicht mit SARS-CoV-2 infiziert („negativer Fall“);
- $x_1 \in \mathbb{R}$: durchschnittliche Schlafdauer;
- $x_2 \in \mathbb{R}$: durchschnittliche tägliche Schrittzahl;
- $x_3 \in \mathbb{R}$: durchschnittliche Ruheherzfrequenz;
- $x_4 \in \mathbb{R}$: durchschnittliche nächtliche Atemfrequenz;
- $m = 4$: Anzahl der Feature-Variablen;
- $n = 453$: Anzahl der Datenpunkte.

8.2 Konfusionsmatrix

Die folgende Tabelle beschreibt den Rechenweg:

	$f(x_i) = 0$	$f(x_i) = 1$	
$y_i = 0$	$51 - 45 = 6$	45	$453 - 402 = 51$
$y_i = 1$	$15 - 6 = 9$	$438 - 45 = 393$	402
	15	$453 - 15 = 438$	$n = 453$

Die folgende Tabelle enthält die Ergebnisse:

	$f(x_i) = 0$	$f(x_i) = 1$	
$y_i = 0$	6	45	51
$y_i = 1$	9	393	402
	15	438	453

8.3 Performanzmetriken

Wir berechnen:

$$\begin{aligned}
 \text{Sensitivität} = \text{TPR} &= \frac{\text{TP}}{\text{P}} = \frac{393}{402} = 97.76\% \\
 \text{Spezifizität} = \text{TNR} &= \frac{\text{TN}}{\text{N}} = \frac{6}{51} = 11.76\% \\
 \text{PPV} &= \frac{\text{TP}}{\text{PP}} = \frac{393}{438} = 89.72\% \\
 \text{NPV} &= \frac{\text{TN}}{\text{PN}} = \frac{6}{15} = 40.00\% .
 \end{aligned} \tag{37}$$

Wir interpretieren:

- TPR: Von denjenigen, die mit SARS-CoV-2 infiziert sind, werden 97.76% richtigerweise als solche erkannt.
- TNR: Von denjenigen, die nicht mit SARS-CoV-2 infiziert sind, werden $100 - 11.76 = 88.24\%$ fälschlicherweise als infiziert eingestuft.
- PPV: Unter denjenigen, die als SARS-CoV-2-positiv klassifiziert werden, ist diese Zuschreibung bei 89.73% korrekt.
- NPV: Unter denjenigen, die als SARS-CoV-2-negativ klassifiziert werden, ist diese Zuschreibung bei $100 - 40.00 = 60.00\%$ nicht korrekt.

8.4 Balanced Accuracy

Klassifikationsgenauigkeit und balancierte Klassifikationsgenauigkeit sind:

$$\begin{aligned} \text{ACC} &= \frac{\text{TN} + \text{TP}}{T} = \frac{6 + 393}{453} = 88.08\% \\ \text{BAC} &= \frac{\text{TPR} + \text{TNR}}{2} = \frac{1}{2} \left(\frac{393}{402} + \frac{6}{51} \right) = 54.76\% . \end{aligned} \tag{38}$$

Die Klassifikationsgenauigkeit (ACC) gibt den Anteil korrekt klassifizierter Fälle an allen Datenpunkten an. Im vorliegenden Fall beträgt sie 88.08 %, was zunächst auf eine gute Klassifikationsleistung schließen lässt. Allerdings sind die beiden Klassen stark ungleich verteilt: 402 von 453 Personen (88.74 %) sind tatsächlich positiv. Ein trivialer Klassifikator, der ausnahmslos SARS-CoV-2 vorhersagt, würde somit sogar eine höhere ACC von 88.74 % ($= (402 + 0)/453$) erreichen. LOO-CV ist ein Instrument, um die Vorhersagegüte eines Klassifikators außerhalb des Trainingsdatensatzes zu schätzen, löst aber nicht das Problem unbalancierter Klassen.

Die balancierte Klassifikationsgenauigkeit (BAC) berücksichtigt dieses Problem, indem sie den Mittelwert aus Sensitivität (TPR) und Spezifität (TNR) bildet. Dadurch gehen beide Klassen unabhängig von ihrer Häufigkeit mit gleichem Gewicht in die Bewertung ein. Im vorliegenden Fall beträgt die BAC lediglich 54.76 %, da die hohe Sensitivität (97.76 %) durch eine sehr geringe Spezifität (11.76 %) relativiert wird. Eine BAC von 50 % entspricht dem Erwartungswert eines Klassifikators, der unabhängig von der tatsächlichen Klasse zufällig vorhersagt. Der vorliegende Klassifikator liegt mit 54.76 % nur wenig darüber. Die BAC zeigt somit, dass die prädiktive Performanz trotz hoher ACC nur geringfügig über dem Zufallsniveau liegt.