.

.

Übungen 5 Lösungen

1. **Aufgabe:** Worin besteht der Unterschied zwischen einer PCA und einer Faktorenanalyse (FA)?

   (a) Die PCA bestimmt nur die Hauptachsen von Ellipsoiden, während die FA auf psychologisch sinnvolle Faktoren (= latente Dimensionen) zielt. ☐

   (b) Die PCA ist nur eine Anwendung von Theoremen der Linearen Algebra, während die FA ein psychologisches Modell ist. ☐

   (c) Bei der FA werden gemeinsame latente Variablen (Faktoren) einerseits und spezifische Faktoren bzw. Fehler separat geschätzt, bei der PCA wird eine solche Trennung zunächst nicht vorgenommen ☐**X**

   (d) Bei der FA wird die multivariate Normalverteilung der Daten angenommen, bei der PCA nicht. ☐

2. **Aufgabe:** Ein gängiges Argument gegen die PCA ist, dass die latenten Dimensionen (Hauptachsen) nicht rotiert werden können, ohne die Unabhängigkeit der latenten Variablen zu zerstören.

   (a) Macht nichts, unrotiert sind Hauptachsen am schönsten. ☐

   (b) Da Rotationen beliebig sind, gewinnt man auch nichts, wenn man rotiert. ☐

   (c) Man kann *entweder* die Faktorwerte *oder* die Ladungen rotieren und dabei die Orthogonalität erhalten (vergl. Satz 3.1 im Skriptum "Einführung in die Hauptkomponenten...") ☐**X**

   (d) Die Bedeutung von Rotationen wird weithin überschätzt. ☐

3. **Aufgabe:** Gegeben sei eine multiple Regression $\mathbf{y} = X\mathbf{b} + \mathbf{e}$, X die Matrix der Prädiktorvariablen (-> Spalten). Welcher Zusammenhang besteht zwischen der Schätzung $\hat{\mathbf{b}}$ und den Eigenwerten λ_k von X?

   (a) Der Vektor $(\lambda_1, \ldots, \lambda_n)$ gibt die Richtung an, in die die besten Schätzungen $\hat{\mathbf{y}}$ zeigen. ☐

   (b) Je größer die λ_k, desto weniger Komponenten von $\mathbf{b}$ müssen geschätzt werden. ☐

   (c) Es gibt keinen Zusammenhang zwischen $\hat{\mathbf{b}}$ und den λ_k, da die Schätzungen $\hat{\mathbf{b}}$ nur von der Voraussagekraft der Prädiktoren abhängen. ☐

(d) Korrelationen zwischen den Prädiktoren implizieren die Existenz kleiner Eigenwerte, die sich wiederum auf die Varianz der Komponenten von $\hat{\mathbf{b}}$ auswirken. ☐**X**

4. Bei einer PCA-Regression wird die SVD auf die Matrix X der Prädiktoren angewendet, $X = Q\Lambda^{1/2}T'$.

   (a) Die neuen Prädiktoren sind die Spalten von $A = T\Lambda^{1/2}$. ☐

   (b) Die neuen Prädiktoren sind die Spalten von $L = Q\Lambda^{1/2}$. ☐

   (c) Die neuen Prädiktoren sind die Spalten von Q, $\mathbf{u} = \Lambda^{1/2}T'\mathbf{b}$ ist der Vektor der neuen Regressionsgewichte. ☐**X**

5. **Aufgabe:** Eine Faktorenanalyse soll durch eine SVD approximiert werden, $X = Q\Lambda^{1/2}T'$. Welche Bedeutung hat die Matrix $A = T\Lambda^{1/2}$, in welcher Beziehung steht sie zur Matrix R der Korrelationen zwischen den Items?

   (a) $R = A'A$, wobei $\text{diag}(A'A) = \text{diag}(\lambda_1, \ldots, \lambda_n)$. ☐

   (b) $R = AA'$, $A'A = \text{diag}(\lambda_1, \ldots, \lambda_n)$. ☐**X**

   (c) A ist die Matrix der Ladungen der Variablen. Da sie hypothetisch ist, steht sie in keiner direkten Beziehung zur empirisch gegeben Matrix R. ☐

   (d) $X = A'A$, da man die Daten aus den Ladungen vorhersagen kann. ☐

6. **Aufgabe:** Es sei $\mathbf{a}_j = (a_{1j}, a_{2j}, \ldots, a_{nj})'$ der Vektor der Ladungen der j-ten Variablen auf den latenten Dimensionen.

   (a) $\mathbf{a}_j'\mathbf{a}_j = 1$ für alle j. ☐**X**

   (b) $\mathbf{a}_j\mathbf{a}_j'$ ist als dyadisches Produkt eine orthonormale Matrix, da die $\mathbf{a}_j$ orthonormal sind. ☐

   (c) $\mathbf{a}_j'\mathbf{a}_{j'} = 0$, da die Faktoren orthonormal sind. ☐

   (d) $\mathbf{a}_j = \vec{0}$, wenn die j-te Variable linear abhängig von den übrigen Variablen ist. ☐

7. **Aufgabe:** Sie sollen einen aus dichotomen Items bestehenden Fragebogen analysieren und dazu die latenten Variablen bestimmen, die den befragten Merkmalen unterliegen. Sie berechnen deshalb alle Paare von Items die korrespondierenden ϕ-Koeffizienten.

   (a) ϕ-Koeffizienten sind keine richtigen Korrelationskoeffizienten, da die Antworten nur als entweder falsch (0) oder korrekt (1) kodiert werden, die Merkmale also nicht auf einer Intervallskala gemessen werden. ☐

   (b) $\phi_{ij} = .5$ für alle Variablen $i \neq j$ bedeutet, dass es bei n Items n/s latente Dimensionen gibt. ☐

(c) Im Fall von $\phi_{ij} \approx .5$ für alle $i \neq j$ gibt es bei hinreichend großem Wert von n nur einen Eigenwert größer als 1. ☐**X**

(d) Die Randverteilungen $p_i \neq p_j$ für Itempaare $(i \neq j)$ sind wesentlich, um irrelevante Schwierigkeitsfaktoren zu vermeiden. ☐

Kommentar: Die Antwortalternativen sind leider suboptimal gewählt worden, weil nicht $\phi \approx .5$ eine wichtige Bedingung für eine sinnvolle Schätzung des Korrelationskoeffizienten ρ durch ϕ ist, sondern sondern $p_x = p_y = .5$; in diesem Fall kann ϕ jeden Wert zwischen -1 und +1 annehmen! Sorry!

8. **Aufgabe:** Verurteilte Straftäter sollen als entweder rückfallgefährdet oder nicht rückfallgefährdet klassifiziert werden. Für eine Gruppe von Straftätern hat man p Symptome (Prädiktoren) auf Intervallsskalen gemessen, so dass $\mathbf{y} = u_1\mathbf{x}_1 + u_2\mathbf{x}_2 + \cdots + u_p\mathbf{x}_p$ und der Wert von $\mathbf{y}$ soll eine möglichst fehlerfreie Klassifikation erlauben. Gesucht sind die "Gewichte" u_k, $k = 1, \ldots, p$, um eine optimale Klassifikation zu gewährleisten.

   (a) Der Ansatz ist schon deswegen falsch, weil kein Fehlerterm $\mathbf{e}$ in der Regressionsgleichung für $\mathbf{y}$ auftritt. ☐

   (b) Der Ansatz ist schon deswegen falsch, weil sich menschliches Vedhalten nicht berechnen läßt. ☐

   (c) Der Ansatz ist falsch, weil $\mathbf{y}$ notwendig eine (0, 1)-Skala sein müsste (nicht rückfällig, rückfällig), bei kontinuierlichen Prädiktoren $\mathbf{x}_k$ muß abr $\mathbf{y}$ ebenfalls kontinuierlich sein. ☐

   (d) Man muß $\mathbf{u} = (u_1, \ldots, u_p)'$ so bestimmen, dass die Varianz der $\mathbf{y}$ zwischen den beiden Gruppen (rückfällig versus nicht rückfällig) maximal wird relativ zur Varianz innerhalb der Gruppen. ☐**X**

9. **Aufgabe:** Es werde noch einmal der Ansatz von Aufgabe 8 betrachtet.

   (a) $\mathbf{y}$ ist eine Linearkombination der $\mathbf{x}_k$, $1 \leq k \leq p$ und definiert deshalb eine Gerade in einem maximal p-dimensionalen Raum, der linearen Hülle der $\mathbf{x}_k$. Der Ansatz $\mathbf{y} = \sum_k u_k \mathbf{x}_k$ kann also nur funktionieren, wenn sich das Merkmal Rückfälligkeit auf einer 1-dimensionalen Skala abbilden läßt. ☐

   (b) Man kann die Varianzen-Kovarianzen zwischen den Gruppen in einer Matrix B zusammenfassen, und die Varianzen-Kovarianzen innerhalb der Gruppen in einer Matrix W. Die Maximierung Der Varianz "zwischen" relativ zur Varianz "innerhalb"

      i. bedeutet die Maximierung von $\mathbf{u}'B\mathbf{u}/\mathbf{u}'W\mathbf{u} = \lambda$ bezüglich $\mathbf{u}$. ☐**X**

      ii. bedeutet die Bestimmung der $\mathbf{u}$ für $W^{-1}B\mathbf{u} = \lambda\mathbf{u}$. ☐**X**

      iii. $\mathbf{u}$ als Eigenvektor von $W^{-1}B$ kann nur berechnet werden, wenn $W^{-1}B$ eine symmetrische Matrix ist. ☐

(c) $W = PDP'$, P die Eigenvektoren von W und $D = \text{diag}(d_1, \ldots, d_n)$ die Eigenwerte von W. Dann $W^{-1} = PD^{-1}P'$. Welche Implikationen haben kleine Eigenwerte von W? Bedenken Sie: $W^{-1} = \sum_k \frac{\mathbf{p}_k \mathbf{p}_k'}{d_k}$.

   i. Kleine Eigenwerte deuten auf Kollinearitäten der Prädiktoren, W^{-1} bedeutet große Varianz der Schätzungen für $\mathbf{u}$. ☐**X**

   ii. Da $W^{-1}B\mathbf{u} = \lambda\mathbf{u}$ folgt, dass die Komponenten von $\mathbf{u}$ zu klein ausfallen. ☐

   iii. Da $W^{-1}B\mathbf{u} = \lambda\mathbf{u}$ folgt, dass die Komponenten von $\mathbf{u}$ zu groß ausfallen. ☐

(d) Die Anzahl der Lösungen $\mathbf{u}$ für $W^{-1}B\mathbf{u} = \lambda\mathbf{u}$ entspricht dem Rang von $W^{-1}B$.

   i. Da $\mathbf{y} = \sum_k u_k \mathbf{x}_k$ eine Gerade im p-dimensionalen Raum der Prädiktoren definiert, muß *eine* der Lösungen $\mathbf{u}_k$ ausgewählt werden, – man wählt die, deren zugehöriger Eigenwert λ maximal ist. ☐

   ii. Verschiedene Lösungen $\mathbf{u}_k$ können nicht orthogonal sein, da $W^{-1}B$ nicht notwendig symmetrisch ist[1]. ☐

   iii. Verschiedene Lösungen $\mathbf{u}_k$ definieren einen orthogonalen Teilraum der linearen Hülle der $\mathbf{x}_k$ mit den Achsen $\mathbf{y}_1, \ldots, \mathbf{y}_r$, $r \leq p$. ☐**X**

[1] Die Eigenvektoren *symmetrischer* Matrizen sind orthogonal.