08. 06. 2013

Multivariate Methoden – Aufgaben II + Lösungen

1. **Aufgabe:** Unter welchen Bedingungen gilt

$$\mathbf{x}'\mathbf{y} = \cos\theta = r_{xy},$$

r_{xy} der Produkt-Moment-Korrelationskoeffizient zwischen zwei zufälligen Veränderlichen X und Y, θ der Winkel zwischen den Vektoren $\mathbf{x}$, $\mathbf{y}$, deren Komponenten die Messwerte für X und Y für eine gegebene Stichprobe sind?

2. **Aufgabe:** Es seien A und B zwei psychische Merkmale, und $\mathbf{a}$ und $\mathbf{b}$ seien zwei n-dimensionale Vektoren, die die "wahren" Werte der Ausprägungen von A und B bei einer Stichprobe von n Personen enthalten mögen. Es gelte die Beziehung $\mathbf{b} = \alpha\mathbf{a}$, dh $\mathbf{a}$ und $\mathbf{b}$ bilden einen 1-dimensionalen Teilraum des n-dimensionalen Vektorraums. $\mathbf{x} = \mathbf{a} + \mathbf{e}_x$ und $\mathbf{y} = \mathbf{b} + \mathbf{e}_y$ seien die Vektoren der Messwerte für A und B bei der betrachteten Stitchprobe, wobei $\mathbf{e}_x$ und $\mathbf{e}_y$ die Vektoren der Messfehler seien. Erläutern Sie bitte die Aussage, dass $\mathbf{x}$ und $\mathbf{y}$ linear unabhängige Vektoren sind.

3. **Aufgabe:** Der Rang r einer $m \times n$-Matrix X ist die Anzahl der linear unabhängigen (l.u.) Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_r$, die notwendig sind, um die Spaltenvektoren von X als Linearkombinationen der $\mathbf{a}_1, \ldots, \mathbf{a}_r$ darzustellen; r ist auch gleich der Anzahl der l. u. Vektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$, die notwendig sind, um die Zeilenvektoren als Linearkombinationen der $\mathbf{b}_1, \ldots, \mathbf{b}_r$ darzustellen. X sei insbesondere eine Matrix von Messwerten. Warum findet man im Allgemeinen, dass der Rang von X gleich $\min, (m, n)$ ist?

4. **Aufgabe:** Es sei X eine $m \times n$-Matrix von Messwerten. Mit welcher Begründung versucht man, die Vektoren von X durch Linearkombinationen von $r < \min(m, n)$ orthogonalen und damit linear unabhängigen Vektoren zu approximieren? Warum sucht man für diese Approximation insbesondere orthogonale und nicht einfach linear unabhängige Vektoren? (Es genügt ja, l. u. Vektoren zu wählen, um den ganzen n- und m-dimensionalen Vektorraum aufzuspannen!)

5. **Aufgabe:** Die Aufgabe, eine Matrix X durch "latente Variablen" oder "latente Vektoren" zu "erklären" wird bei der Hauptachsentransformation (PCA) ja gelöst, indem man die Vektoren von X um einen bestimmten Winkel rotiert.

   (a) Die Komponenten des Zeilenvektors $\mathbf{x}_{(i)}$ von X sind die Messwerte x_{ij}der i-ten Person für die betrachteten n Variablen. Es werde zur Vereinfachung vorausgesetzt, dass die Komponenten zentriert seien,

d.h. es gelte $x_{ij} = X_{ij} - \bar{x}_j$, X_{ij} der "Rohwert" der Messung. die Komponenten von $\mathbf{x}_{(i)}$ können als Koordinaten der i-ten Person in einem n-dimensionalen Vektorraum interpretiert werden. Was geschieht mit den $\mathbf{x}_{(i)}$ bei einer PCA?

(b) Es sei $\mathbf{L}_1$ der Vektor, dessen Komponenten die Koordinaten der Personen auf der ersten latenten Variablen sind. Was läßt sich über die Varianz dieser Koordinaten relativ zu den Komponenten des Vektors $\mathbf{L}_2$ sagen, der zu $\mathbf{L}_1$ orthogonal ist?

(c) $\mathbf{L}_1$ und $\mathbf{L}_2$ seien die Vektoren, die die beiden ersten, unkorrelierten latenten Variablen repräsentieren. Die Vektoren $\mathbf{P}_1$ und $\mathbf{P}_2$ repräsentieren die n gemessenen Variablen auf den latenten Variablen. Sie möchten $\mathbf{P}_1$ und $\mathbf{P}_2$ um einen Winkel θ rotieren, um zu einer besseren Interpretation zu kommen. Dies bedeutet, dass Sie $\mathbf{L}_1$ und $\mathbf{L}_2$ ebenfalls um diesen Winkel rotieren müssen, so dass den Personen ebenfalls neue Koordinaten zugewiesen werden. Sind die latenten Variablen, die diesen neuen Vektoren entsprechen, ebenfalls unkorreliert?

Anmerkung/Erläuterung: Nach der SVD gilt

$$X = Q\Lambda^{1/2}P'$$

und es sei $A = P\Lambda^{1/2}$ die Matrix der Faktorladungen. Rotation bedeutet Multiplikation mit einer Matrix T: $\tilde{A} = AT$[1]. Dabei gilt $T^{-1} = T'$, $T'T = TT' = I$, I die Einheitsmatrix. Dann enthält QT die Koordinaten der Personen auf den neuen latenten Variablen, denn

$$X = Q\Lambda^{1/2}P' = QA' = QTT'A'.$$

6. **Aufgabe:** Die SVD $X = Q\Lambda^{1/2}P'$ kann für eine gegebene Datenmatrix *stets* berechnet werden. *Folgt* daraus logisch, dass eine Interpretation der Messwerte gemäß der Gleichung

$$x_{ij} = q_{i1}a_{1j} + q_{i2}a_{2j} + \cdots + q_{in}a_{nj},$$

wobei die q_{ik}, $k = 1, \ldots, n$ Ausprägungen der Personen auf den latenten Dimensionen und a_{jk} die "Ladungen" der "Tests" (gemessene Variablen) sind, stets möglich und erlaubt ist?

7. **Aufgabe:** Die n Experten $R_1, R_2, \ldots, R_n$ beurteilen m Personen bezüglich einer bestimmten Variablen (zB Eignung für das Studium der Psychologie). Es entsteht eine Datenmatrix X, wobei das Element x_{ij} die Beurteiliung der i-ten Person durch die j-te Rater-Persönlichkeit ist. Die Rater behaupten, aufgrund ihrer Erfahrung absolut objektiv zu urteilen. Wie können Sie diese Behauptung testen?

[1] Zur Erinnerung: die Spaltenvektoren von $\tilde{A}$ werden bei der Rotation als Linearkombinationen der Spaltenvektoren von A dargestellt, – also $\tilde{A} = AT$ und nicht $\tilde{A} = TA$!

Antworten und Lösungen

1. **Antwort Aufgabe 1:** Es ist

$$\cos\theta = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\frac{1}{m}\mathbf{x}'\mathbf{y}}{\frac{1}{m}\|\mathbf{x}\|\|\mathbf{y}\|} = r_{xy},$$

   wenn $\mathbf{x}'\mathbf{y}/m = Kov(x,y)$ und $\|\mathbf{x}\|/m = s_x$, $\|\mathbf{y}\|/m = s_y$. Die Komponenten von $\mathbf{x}$ und $\mathbf{y}$ müssen also Abweichungen vom jeweiligen Mittelwert $\bar{x}$ bzw $\bar{y}$ sein.

2. **Antwort Aufgabe 2:** Damit $\mathbf{e}_x$ und $\mathbf{e}_y$ als Fehlervektoren gelten können, dürfen nicht die Beziehungen $\mathbf{e}_x = \lambda_x\mathbf{x}$ bzw. $\mathbf{e}_y = \lambda_y\mathbf{y}$ Beziehungen gelten, wären $\mathbf{e}_x$ und $\mathbf{e}_y$ exakt parallel zu $\mathbf{x}$ und $\mathbf{y}$ und wären damit Teile von $\mathbf{x}$ bzw. $\mathbf{y}$. Gelten diese Beziehungen also *nicht*, so sind sind $\mathbf{x}$ und $\mathbf{e}_x$ und $\mathbf{y}$ und $\mathbf{e}_y$ nicht parallel und damit sind die Fehlervektoren und die $\mathbf{a}$ und $\mathbf{b}$ linear unabhängig. (Vergl. die Betrachtungen zum Skalarprodukt und dem Begriff der linearen Unabhängigkeit im Skriptum zu Vektoren und Matrizen!).

3. **Antwort Aufgabe 3:** Die Antwort folgt sofort aus der zur vorangehenden Aufgabe: da jeder Datenvektor einen spezifischen Fehlervektor enthält, sind alle Vektoren der Matrix zumindest im numerischen Sinn linear unabhängig (numerisch bedeutet hier, dass das Gleichungdssystem

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_n\mathbf{x}_n = \vec{0}$$

   exakt nur die Lösung $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 0$ hat).

4. **Antwort Aufgabe 4** Die Annahme $r < n$ resultiert aus der Annahme, dass die anscheinende lineare Unabhängigkeit der n Spaltenvektoren (falls $n < m$) eben nur ein Resultat der Existenz von n Fehlervektoren in den Daten ist. Die Eigenwerte repräsentieren die Varianz, die durch eine latente Dimension erzeugt wird. Eine "große" Varianz (wie die für die erste, zweite latente Variable) resultiert, wenn außer den Fehlerkomponenten noch systematische Unterschiede zwischen den Probanden existieren. Gibt es diese systematischen Unterschiede nicht mehr, so sind die Unterschiede eben nur zufällige Unterschiede. Nimmt man an, dass die Varianz der Fehlerkomponenten für alle latenten Dimensionen von gleicher Größenordnung ist, so haben die "latenten" Dimensionen, die nur durch Fehler entstehen, vergleichbare und auf jeden Fall "kleine" Varianzen, und diese Dimensionen repräsentieren dann eben nur Fehler und keine inhaltlich systematische Information, können also vernachlässigt werden.

   Zur "Erklärung" der Daten wählt man orthogonale Vektoren, weil linear Unabhängigkeit ja nur bedeutet, dass die Vektoren eine unterscheidliche Orientierung haben, – sie können orthogonal, müssen aber nicht orthogonal

sein. Begnügte man sich mit l.u. Vektoren, so würde es genügen, Vektoren zu wählen, die sich nur sehr wenig hinsichtlich ihrer Orientierung unterscheiden. Da hieße, dass sie sich auch hinsichtlich ihrer Bedeutung nur sehr wenig unterscheiden. Orthogonale Vektoren haben den Vorteil, dass sie Bedeutungen haben, die nicht auch von den jeweils anderen Vektoren repräsentiert wird, man greift den Raum der möglichen Bedeutungen sozusagen optimal ab.

Abgesehen davon haben orthogonale Vektoren den Vorteil, auf eine Lösung zu führen, – nämlich die Eigenvektoren von $X'X$ bzw. XX'.

5. **Antwort Aufgabe 5a** Die Vektoren werden in Vektoren $\mathbf{y}_{(i)}$ überführt; die Koordinaten der Endpunkte dieser Vektoren beziehen sich auf ein Koordinatensystem, dessen Orientierung mit der der Hauptachsen des durch $X'X$ definierten Ellipsoids zusammenfällt. Diese neuen Koordinaten repräsentieren nicht korrelierende ("latente") Merkmale.

6. **Antwort Aufgabe 5b:** Die Varianz der Koordinaten auf $\mathbf{L}_1$ ist maximal etc.

7. **Antwort Aufgabe 5c:** Die SVD von X ist $X = Q\Lambda^{1/2}P'$. Es sei $A = P\Lambda^{1/2}$. Die Vektoren von A zu rotieren heißt, durch Linearkombination der Spalten von A zu neuen Koordinaten $\tilde{A}$ überzugehen. Das geschieht, indem man A mit einer geeignet gewählten Matrix T multipliziert: $\tilde{A} = AT$. Da man die Rotation auch in umgekehrter Richtung durchführen kann, folgt $\tilde{A}T' = ATT' = A$, dh $TT' = I$ die Einheitsmatrix, dh T muß orthonormal sein.

   Es gilt aber
   $$Q\tilde{A}' = QT'\Lambda^{1/2} \neq Q\Lambda^{1/2}P' = X,$$
   es kann nur
   $$X = QTT'\Lambda^{1/2}P' = Q\Lambda^{1/2}P', \quad TT' = I$$
   gelten. Dies bedeutet, dass die Rotation $AT = \tilde{A}$ von der Rotation $QT = \tilde{Q}$ begleitet sein muß. $\tilde{Q}$ und $\tilde{A}$ enthalten neue Koordinaten für die Personen bzw. Variablen. Es gibt aber nur ein Koordinatensystem, das durch die Vektoren in Q und P gegeben ist, das unkorrelierte Koordinaten enthält. In jedem anderen Koordinatensystem erscheint die Punktekonfiguration der Personen als relativ zu den Achsen geneigt, – also korrelieren die so gewählten latenten Merkmale.

8. **Antwort Aufgabe 6:** Die Repräsentation der x_{ij} in der Form
   $$x_{ij} = q_{i1}a_{1j} + q_{i2}a_{2j} + \cdots + q_{in}a_{nj},$$
   ist sicherlich stets möglich und erlaubt, – wer will Ihnen eine solche Repräsentation verbieten? Wählt man die PCA und damit die SVD als Basis für

die Analyse der Daten, so akzepiert man diese Repräsentation, – aber sie ist keineswegs die einzig mögliche und kann durchaus falsch sein. Denn in einem speziellen Fall kann zB die Gleichung

$$x_{ij} = q_{i1}a_{1j} + q_{i2}a_{2j} + q_{i3}a_{1j}a_{2j}$$

die korrekte Beschreibung der Daten sein, dh die Messwerte können durch eine Wechselwirkung zwischen zwei latenten Variablen bestimmt sein. Die dritte latente Variable ist aber sicherlich nicht orthogonal zu den beiden ersten, da sie sich ja multiplikativ aus den beiden ersten zusammensetzt (die dritte korreliert also mit der ersten und zweiten latenten Variable). Die PCA kann eine solche multiplikative Verknüpfung gewissermaßen nicht erkennen und "erfindet" eventuell eine dritte orthogonale Variable, – die dann aber ein reines Artefakt ist. Man sieht es den gefundenen latenten Variablen nicht notwendig an, ob sie nur Artefakte oder aber real existierende latente Variable abbilden.

(Vergl. den Artikel von Rist, F., Glöckner-Rist, A., Demmel, R. (2009))