08. 06. 2013

Multivariate Methoden – Aufgaben I + Lösungen

1. **Aufgabe:** Worin liegt der Sinn, im Zusammenhang mit der multiplen Regression von korrelierenden zu nicht korrelierenden Prädiktoren überzugehen? Es gebe n Prädiktorvariablen, und es werden m Fälle untersucht, dh die Regressionsparameter werden anhand der Daten von diesen Fällen geschätzt. Wieviele Basisvektoren zur Darstellung der Prädiktorvektoren benötigen Sie maximal, und welche Dimensionalität haben die Basisvektoren?

   Antwort: Korrelierende Prädiktoren bedeuten den gefürchteten Fall der *Kollinearität*[1], und Kollinearität bedeutet, dass (i) die Regressionskoeffizienten schwer oder gar nicht interpretiert werden können, weil die Schätzungen eine große Varianz haben und überdies negativ miteinandere korreliert sind (ist $\hat{b}_j$ positiv, so ist $\hat{b}_{j+1}$ negativ, etc), und (ii) die Vorhersage des Wertes von Y, der Kriteriumsvariable, sehr ungenau wird, obwohl der Goodnes-of-fit-Test einen guten Fit des Modells an die Daten signalisiert[2]. Hat man z.B. einen Fragebogen für eine bnestimmte klinische Situation konstruiert und sind die Prädiktoren die Fragen (Items), so wird die Diagnose anhand des Fragebogens eine höchst wacklige Angelegenheit, weil die "Vorhersage" (= Diagnose) $\hat{y}$ sehr ungenau wird. *Eine* Möglichkeit, des Problems der Multikollinearität Herr zu werden, ist eine PCA der Prädiktoren, da die PCA auf orthogonale und damit unkorrelierte Prädiktoren führt. Unter Umständen kann man dazu diejenigen Items auswählen, die (fast) nur auf einer latenten Variablen eine Ladung ungleich Null haben, und alle Items, die Ladungen ungleich Null auf mehreren latenten Dimensionen haben, aus dem Fragebogen entfernt. Wenn der Fragebogen ursprünglich n Items enthält, gibt es maximal n latente Dimensionen (die PCA besteht ja nur in einer Drehung der Achsen). Über den Scree-Test kann man zu einer Abschätzung der wirklich relevanten $r < n$ Dimensionen kommen.

2. **Aufgabe:** Gegeben sei die Matrix

$$X = \begin{pmatrix} 2 & 1 & 3 \\ 1 & .5 & 1.5 \\ 3 & 1.5 & 4.5 \end{pmatrix}.$$

[1] Warum nicht einfach von korrelierenden Prädiktoren gesprochen wird, ist nicht so ganz klar.

[2] Ein einfaches Beispiel für die Schätzung eines Parameters ist der Mittelwert $\bar{x}$ für den Erwartungswert $\mathbb{E}(x)$ (Mittelwert über *alle möglichen* x-Werte): die Varianz der Schätzung $\bar{x}$ ist $Var(\bar{x}) = \sigma^2/n$, σ^2 die Varianz der zufälligen Variablen X. Für kleine Stichprobenumfänge n ist $Var(\bar{x})$ groß, und für $n \to \infty$ geht sie gegen Null.

Sind die Spaltenvektoren von X linear abhängig oder liner unabhängig? Schreiben Sie zumindest den Ansatz auf, nach dem Sie vorgehen würden, um die Frage zu beantworten.

Antwort: Inspektion der Spalten der Matrix zeigt, dass die zweite und dritte Spalte Vielfache der ersten Spalte sind: $\mathbf{x}_2 = .5\mathbf{x}_1$, $\mathbf{x}_3 = 1.5\mathbf{x}_1$, dh die Vektoren unterscheiden sich alle nur hinsichtlich ihrer Länge, nicht aber hinsichtlich ihrer Orientierung. Deshalb liegen sie alle in einem 1-dimensionalen Teilraum des 3-dimensionalen Vektorraums. Der Grundsätzliche Ansatz zur Überprüfung der linearen Abhängigkeit besteht darin, die Vektorgleichung

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \lambda_3\mathbf{x}_3 = \vec{0}$$

zu betrachten. Diese Gleichung ist äquivalent einem System von drei Gleichungen mit drei Unbekannten:

$$\begin{aligned} \lambda_1 2 + \lambda_2 1 + \lambda_3 3 &= 0 \\ \lambda_1 1 + \lambda_2 .5 + \lambda_3 1.5 &= 0 \\ \lambda_1 3 + \lambda_2 1.5 + \lambda_3 4.5 &= 0 \end{aligned}$$

Man löst es, indem man zB die erste Gleichung nach λ_1 auflöst:

$$\lambda_1 = -\frac{1}{2}\lambda_2 1 - \frac{1}{2}\lambda_3 3$$

und diesen Ausdruck für λ_1 in die zweite und dritte Gleichung einsetzt; diese beiden Gleichungen bilden dann ein System von zwei Gleichungen mit zwei Unbekannten, das man in analoger Weise löst, indem man eine Gleichung nach λ_2 auflöst und den entstehenden Ausdruck in die andere Gleichung einsetzt, etc. Diese Schreibübung muß man hier nicht durchführen. Linear unabhängig sind die Vektoren, wenn es nur die Lösing $\lambda_1 = \lambda_2 = \lambda_3 = 0$ gibt. Aber dass die λ-Werte ungleich Null sind, wissen wir ja schon.

3. **Aufgabe:** Gegeben seien die Vektoren

$$\vec{x}_1 = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ .5 \\ 1.5 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 1.5 \\ 4.5 \end{pmatrix}$$

Liegen diese Vektoren in einem 1- oder einem 2-dimensionalen Teilraum, oder spannen Sie den 3-dimensionalen Raum auf?

Antwort: Wie in der Antwort zur vorangegangenen Aufgabe schon gezeigt wurde, liegen die Vektoren in einem 1-dimensionalen Teilraum, und sie können dann eben nur diesen Teilraum aufspannen und keinen 2- oder

gar 3-dimensionalen (Teil-)Raum. Linearkombinationen dieser drei Vektoren führen stets nur zu einem Vektor, der sich nur in der Länge, nicht aber in der Orientierung von diesen Vektoren unterscheidet und die deshalb nicht aus diesem Teilraum hinausführen.

4. Gegeben sei eine symmetrische Matrix M:

   (a) $\mathbf{t}$ sei ein Eigenvektor von M. $M\mathbf{t} = \mathbf{y}$ ist wieder ein Vektor. Wie groß darf der Unterschied zwischen $\mathbf{t}$ und $\mathbf{y}$ in Bezug auf (i) die Länge, (ii) die Orientierung sein?

   Antwort: Da $\mathbf{t}$ ein Eigenvektor sein soll, muß $M\mathbf{t} = \lambda\mathbf{t}$, also $\mathbf{y} = \lambda\mathbf{t}$ gelten. Damit hat $\mathbf{y}$ dieselbe Orientierung wie $\mathbf{t}$, dh $\mathbf{y}$ und $\mathbf{t}$ dürfen sich hinsichtlich der Orientierung gar nicht unterscheiden. Der Wert von λ hängt von der Matrix M ab und ist insofern beliebig, dh $\mathbf{y}$ und $\mathbf{t}$ dürfen sich beliebig bis auf die Einschränkung $\lambda < \infty$ hinsichtlich der Länge unterscheiden.

   (b) $\mathbf{t}$ kann nur dann ein Eigenvektor von M sein, wenn M eine Korrelationsmatrix ist – richtig oder falsch?

   Antwort: Diese Behauptung ist falsch. Es kommt nur auf die Symmetrie von M an, unabhängig davon, was die Elemente von M bedeuten.

   (c) M sei eine $n \times n$-Matrix. Wieviele Dimensionen hat das Ellipsoid, das durch M definiert wird?

   Antwort: Die Anzahl der Dimensionen wird durch den Rang r der Matrix bestimmt. Es ist stets $r \leq n$. Das Ellipsoid hat demnach maximal n Dimensionen. r ist die Anzahl der linear unabhängigen (l.u.) Vektoren, die zur Darstellung der Vektoren von M benötigt werden.

   (d) R sei eine $n \times n$-Korrelationsmatrix. Müssen die Daten multivariat normalverteilt sein, damit R ein Ellipsoid definiert?

   Antwort: Nein. Es ist eine Eigenschaft aller symmetrischen Matrizen, ein Ellipsoid zu charakterisieren (die Orientierungen der Eigenvektoren geben die Orientierungen der Hauptachsen des Ellipsoids an.

5. **Aufgabe:** Es sei X eine $m \times n$-Datenmatrix. Ein Statistiker erklärt Ihnen, er könne die Daten durch die Gleichung $X = LP'$, "erklären", wobei die Spaltenvektoren von L orthgonal sind. Er meint damit, dass die Spaltenvektoren von X als Linearkmombinationen

   (a) der Spaltenvektoren von P' gegeben sind.

   (b) der Spaltenvektoren von P gegeben sind.

   (c) der Spaltenvektoren von L' gegeben sind.

   (d) der Spaltenvektoren von L gegeben sind.

Die Matrizen L und P sind unbekannt. Welche Eigenschaft der Matrix L erlaubt es, die Matrix P zu berechnen? Warum kann man L berechnen, wenn P bekannt ist?

Antwort: Die Spaltenvektoren von X werden als Linearkombinationen der Spalten von L erklärt, die als orthogonal vorausgesetzt werden, so dass $L'L$ eine Diagnonalmatrix ist. Dann folgt $X'X = P(L'L)P' = P\Lambda P'$, $\Lambda = L'L$, woraus wiederum folgt, dass P die Matrix der Eigenvektoren von $X'X$ ist. Da $X'X$ symmetrisch ist, folgt, dass P orthonormal ist (die Spaltenvektoren von P sind orthogonal und normiert, dh haben die Länge 1). Deswegen folgt aus $X = LP'$ durch Multiplikation mit P von rechts

$$XP = LP'P = L.$$

Es ist also die Eigenschaft der Orthonormiertheit von P, die es erlaubt, die Matrix L zu berechnen.

6. **Aufgabe:** Was versteht man unter der Singularwertzerlegung (SVD) der Matrix X? In welchem Sinne korrespondiert die SVD zum Modell der Faktorenanalyse, und in Bezug auf welchen Aspekt der Faktorenanalyse unterscheidet sich die SVD vom Modell der Faktorenanalyse?

   Antwort: Die SVD folgt aus dem Ansatz $X = LP'$ mit L eine orthogonale Matrix. Durch Multiplikation von L von rechts mit $\Lambda^{-1/2}$ entsteht $Q = L\Lambda^{-1}$, d.h. Q ist orthonormal (rechnen Sie zur Übung explizit nach, dass auf diese Weise eine Normierung der Vektoren in L entsteht, bedenken Sie dabei, dass $\lambda_j = \mathbf{L}_j'\mathbf{L}_j = \|\mathbf{L}_j\|^2$). Also folgt die SVD

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

   Aus $X = LP'$ folgt

$$x_{ij} = p_{1j}L_{i1} + \cdots + p_{nj}L_{in}.$$

   Die $L_{i1}, \ldots, L_{in}$ können als "Scores" der i-ten Person auf den latenten Achsen interpretiert werden, die $p_{1j}, \ldots, p_{nj}$ als Koordinaten des j-ten "Tests" auf den latenten Dimensionen. Ist man an den Ladungen $A = P\Lambda^{1/2}$ interessiert, so erhält man die Darstellung

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

   Beide Darstellungen von x_{ij} entsprechen dem Prinzip der Faktorenanalyse, allerdings mit verschiedenem Fokus: der erste fokussiert auf die Personen, der zweite auf die Tests (Items, etc).