.

U. Mortensen 23. 06. 2017

Die wesentlichen Ergebnisse der Linearen Algebra für Anwendungen in der multivariaten Statistik.

Vorschlag: Statt sich durch das Skriptum über Vektoralgebra von Seite zu Seite zu quälen und dabei nicht zu wissen, was man für die Klausur braucht oder nicht, kann es besser sein, sich vom Hauptergebnis (die Singularwertzerlegung (SVD)) zurückzuhangeln. Gegeben sei eine $(m \times n)$-Datenmatrix, m Zeilen (zB Personen), n Spalten (Variable, Tests, Items, ...). Man ist meistens daran interessiert, die Korrelationen zwischen den Variablen zu "erklären". Dazu die SVD

$$X = Q\Lambda^{1/2}T' \tag{1}$$

X habe den Rang $\mathrm{rg}(X) = r \leq \min(m, n)$, Q ist eine $(m \times r)$-Matrix, $\Lambda^{1/2}$ ist eine $(r \times r)$-Diagonalmatrix, T ist eine $(n \times r)$-Matrix. Es ist $\mathrm{rg}(Q) = \mathrm{rg}(Q) = r$, Q enthält orthonormale Spaltenvektoren, T enthält orthonormale Spaltenvektoren.

Angenommen, Sie haben den Eindruck, nicht so recht zu verstehen, was diese Matrixgleichung eigentlich bedeutet[1] Sie könnten wie folgt vorgehen:

Vorbemerkung zum Skriptum über Lineare Algebra: Es liefert die mathematische Basis der multivariaten Statistik, soweit diese in der Vorlesung behandelt wurde. Der Zweck des Skriptums ist, nicht nur die nötigten Begriffe wie Vektor, Skalarprodukt, Matrix, Rang einer Matrix, Matrixmultiplikation, Eigenvektor, Eigenwert etc vorzustellen, sondern die logischen Verbindungen zwischen den Begriffen darzustellen. Man kann argumentieren, dass Psychologen über keine derartigen Kenntnisse verfügen müssen. Der Punkt ist aber, dass reines Auswendiglernen einiger Begriffsdefinitionen und eine eher rechnerische Annäherung an die Verfahren (PCA, Diskriminanzanalyse, multiple Regression etc) das unangenehme Gefühl erzeugt, nicht zu verstehen, was man da eigentlich macht und wie die Ergebnisse einer Datenanalyse zu deuten sind. Das Skriptum ist so aufgebaut,

[1] Zur Bedeutung der SVD:
1. Sie zeigt, dass die Spaltenvektoren von X als Linearkombinationen der linear unabhängigen (sogar orthogonalen) Spaltenvektoren von Q bzw. $L = Q\Lambda^{1/2}$ dargestellt werden können, und die Zeilenvektoren von X können als Linearkombinationen der linear unabhängigen (orthogonalen) Zeilenvektoren von T' bzw. von $A' = \Lambda^{1/2}T'$ dargestellt werden.
2. Die SVD zeigt den Zusammenhang von R- und Q-Analysen (Cattell) auf; R-Analysen fokussieren auf die Darstellung der Variablen (Tests, Items, etc), Q-Analysen auf die Darstellung der Personen ("Typen").
3. Die SVD ist eine Approximation an das Modell der Faktorenanalyse.

dass die logischen Beziehungen zwischen den Begriffen deutlich gemacht werden. Es ist durchaus möglich, dass die Lektüre von der ersten bis zur letzten Seite frustrierend ist, weil zunächst nicht deutlich wird, welche Aussagen ("Sätze") von Bedeutung für die Statistik sind und welche dieser Aussagen nur logische Verbindungen herstellen. Deswegen kann man mit einer Aussage wie (1) beginnen und sich zu den Definitionen und Sätzen zurückarbeiten. Auf diese Weise stellt sich das *Verstehen* der Zusammenhänge ebenfalls ein, – viell.eicht nicht unmittelbar, aber doch nach und nach, und das Gefühl der Frustration verschwindet, – ebenfalls nach und nach ... □

Beispiel: (1), d.h. $X = Q\Lambda^{1/2}T'$ bedeutet, dass die Datenmatrix als Produkt von Matrizen repäsentiert wird.

Also kann man nachschlagen, was ein Matrixprodukt bedeutet. Ergebnis: Matrixprodukte repräsentieren Linearkombinationen, in diesem Fall werden die Spaltenvektoren $\mathbf{x}_j$ von X als Linearkombinationen der Spalten von Q dargestellt, gleichzeitig werden die Zeilenvektoren als Linearkombinationen der Zeilenvektoren von $\Lambda^{1/2}T'$ daergestellt.

Wozu Linearkombinationen? Hier: die Datenvektoren werden als Linearkombinationen von *Basisvektoren* dargestellt. Was sind Basisvektoren? Nachschlagen im Skript liefert eine Antwort. Sich die Beziehung zwischen Basisvektoren und den "Faktoren" in der Faktorenanalyse klarmachen: die Faktoren sind "latente" (weil nicht unmittelbar beobachtete) Vektoren mit der Eigenschaft, unabhängig voneinander zu sein, – also nachschlagen, was *lineare Unabhängigkeit* (bzw. lineare Abhängigkeit) bedeutet. Immer, wenn ein Begriff unklar ist, kann bzw sollte man im Index des Skripts nachschlagen und die jeweiligen Charakterisierungen nachlesen. Auf diese Weise wird das Netzwerk der Begriffe allmählich klar und es stellt sich der Eindruck des Verstehens ein. Und zu verstehen, was man tut ist ja letztlich der Zweck der ganzen Übung.

Es ist zB wichtig, sich die Bedeutung des Begriffs der Ladung von Variablen (auf den latenten Variablen) klar zu machen: warum betrachtet man dazu die Elemente a_{jk} von $A = T\Lambda^{1/2}$? a_{jk} ist die "Ladung" der j-ten Variablen auf der k-ten latenten Dimension. a_{jk} ist die Korrelation zwischen der k-ten latenten Variablen und der j-ten Variablen; dieser Sachverhalt kann hilfreich bei der Interpretation sein. Im Skriptum gibt es eine kompakte Darstellung der Beziehungen zwischen Ladungen, Faktorwerten, und Eigenwerten der Kovarianz- bzw. Korrelationsmatrix. (Was waren nochmal Eigenwerte und Eigenvektoren? Welche Bezieungen gibt es zwischen Faktorwerten und Ladungen?) Man findet leicht, dass

$$XT = L = Q\Lambda^{1/2}, \quad X'Q = A = T\Lambda^{1/2} \tag{2}$$

Faktorwerte L und Ladungen A ergeben sich also als Linearkombinationen einerseits der Spaltenvektoren der Datenmatrix, andererseits der Spaltenvektoren von X', was auf eine Beziehung zwischen Ladungen und Faktorwerten verweist (s. a. R- und Q-Analysen von Catell ...).