.

10. 12. 2016 Multivariate Verfahren, WS 2016/2017

Übungen 4 – Formale Grundlagen der PCA

1. **Aufgabe:** Gegeben sei eine $(m \times n)$-Datenmatrix: $m = 200$ Fälle, $n = 50$ Variablen. Was bedeutet es, wenn für diese Matrix gesagt werden kann, dass die 200 Zeilenvektoren in einem 5-dimensionalen Teilraum des $\mathbb{R}^{200}$ liegen? Ist es in diesem Fall möglich, dass die 50 200-dimensionalen Spaltenvektoren in einem 5-dimensionalen Teilraum liegen?

2. **Aufgabe:** Es sei A eine (10×3)-Matrix. $\mathbf{x}$ und $\mathbf{y}$ seien Vektoren. Betrachten sie das Gleichungssystem $\mathbf{y} = A\mathbf{x}$.
   (a) Welche notwendige Bedingung muß erfüllt sein, damit dine Lösungsvektor $\vec{x}$ existiert?
   (b) Wieviele Dimensionen hat der Vektorraum, in dem $\mathbf{x}$ liegt, und wieviele Dimensionen hat der Vektorraum, in dem $\mathbf{y}$ liegt?
   (b) Die Matrix A und der Vektor $\mathbf{y}$ seien vorgegeben. Läßt sich $\mathbf{x}$ dann berechnen? Wenn ja, Welche Voraussetzung muß erfüllt sein?
   (c) Existiert eine zu A inverse Matrix A^{-1}?

3. **Aufgabe:** Es seien A und B zwei nicht quadratische Matrizen.
   (a) Welche Voraussetzungen müssen erfüllt sein, damit Sie ein Produkt $C = AB$ berechnen können, und gilt dann $AB = BA$?
   (b) Angenommen, das Produkt C kann berechnet werden. Sind die Spaltenvektoren von C dann Linearkombinationen der Spalten von A oder der Spalten von B?
   (c) A habe den Rang $\text{rg}(A) = r_A$ und B habe den Rang $\text{rg}(B) = r_B < r_A$. Welche allgemeine Aussage läßt sich über den Rang von C machen, falls C berechenbar ist?
   (d) Welche Bedingung muß *notwendig* erfüllt sein, damit die Spaltenvektoren von C Linearkombinationen der Zeilenvektoren von B sind? Die Bedingung muß nicht hinreichend sein. (Hinweis: Die Anzahlen der Zeilen und Spalten von C hängen von den Anzahlen der Zeilen und Spalten von A und B ab!)

4. **Aufgabe:** Die $(m \times n)$-Datenmatrix X habe den Rang r. Bekanntlich läßt sich X dann stets als Produkt zweier Matrizen U und V schreiben, d.h. $X = UV$.
   (a) Welche Aussage läßt sich dann einerseits über die Anzahl der Zeilen von U und andererseits über die Zeilen und Spalten von V machen?
   (b) Was läßt sich über die Ränge von U und V sagen?
   (c) Müssen die Spaltenvektoren von U notwendig orthogonal sein?

(d) Welche Bezienung besteht zwischen der SVD von X und der Beziehung $X = UV$?

5. **Aufgabe:** Eine $(m \times n)$-Datenmatrix X habe den Rang r.
   (a) Ist die Möglichkeit, die Spalten von X als Linearkommbinationen von Basisvektoren darzustellen, an Annahmen über die Wahrscheinlichkeitsverteilung der Elemente x_{ij} gekoppelt?
   (b) Welche Dimensionalität haben die Basisvektoren, und welche Dimensionalität hat die lineare Hülle (i) der Spaltenvektoren von X, (ii) der Zeilenvektoren von X?
   (c) Wieviele Möglichkeiten haben Sie, Basisvektoren für den Teilraum des $\mathbb{R}^m$ zu wählen, in dem die Spaltenvektoren von X liegen?

6. **Aufgabe:** Für die Datenmatrix wird der Ansatz $X = LT'$ gemacht, wobei die Transformation von Vektoren $\mathbf{x}$ gemäß $\mathbf{y} = T\mathbf{x}$ eine Rotation der $\mathbf{x}$ bedeuten soll.
   (a) Welche Beziehung besteht zwischen den Zeilenvektoren $\tilde{\mathbf{x}}_i$ von X und den Zeilenvektoren $\tilde{\mathbf{L}}_i$ $(i = 1, \ldots, m)$ von L?
   (b) Die Punktekonfiguration der Fälle wird durch die Endpunkte der Vektoren $\tilde{\mathbf{x}}_i$ definiert. Warum liegt der Endpunkt jedes Vektors $\tilde{\mathbf{x}}_i$ auf einem Ellipsoid $\mathcal{E}_i$ und warum haben alle diese Ellipsoide dieselbe Orientierung?
   (c) Impliziert die unter (b) genannte Beziehung zwischen Datenpunkten und Ellipsoiden, dass die Daten multivariat normalverteilt sind?
   (d) Liegen die Variablen, die durch die Endpunkte der Spaltenvektoren $\mathbf{x}_j$ von X repräsentiert werden, ebenfalls auf Ellipsoiden?

7. **Aufgabe:** Es gelte wieder $X = LT'$. Die Annahmen, dass (i) die Spaltenvektoren von L sind orthognal, und (ii) T repräsentiert eine Rotation implizieren, dass die Hauptachsen der unter (c) der vorangegangenen Aufgabe genannten Ellipsoide als neues Koordinatensystem betrachtet werden, dessen Achsen unkorrelierte latente Variablen repräsentieren.
   (a) Wie werden die Koordinaten der Datenpunkte auf diesen Achsen berechnet?
   (b) Durch welche Eigenschaft sind die Eigenvektoren von $X'X$ charakterisiert?
   (c) in welcher Beziehung stehen die Varianzen der Koordinaten der Datenpunkte (Fälle) auf den Hauptachsen der Ellipsoide zu den Eigenwerten von $X'X$?
   (d) Für die zentrierten oder standardisierten Messwerte x_{ij} gilt gemäß dem Ansatz $X = LT'$

$$x_{ij} = \begin{cases} q_{i1}a_{j1} + q_{i2}a_{j2} + \cdots + q_{in}a_{jn}, \text{ oder} \\ L_{i1}t_{j1} + L_{i2}t_{j2} + \cdots + L_{in}t_{jn}. \end{cases} \quad (1)$$

   Von welcher Umformulierung von $X = LT'$ wurde hier Gebrauch gemacht und welche Zielsetzungen liegt Ihre Entscheidung für eine der beiden Möglichkeiten zugrunde?

(e) Ein Sozialpsychologe berichtet, er habe bei der Analyse einer Datenmatrix X gefunden, dass die Ränge von X und L gleich r seien, T aber einen Rang $s < r$ habe. Ein pädagogischer Psychologe stimmt dem Befund zu, weil er bei einer Datenanalyse gefunden habe, X und T hätten denselben Rang r gehabt, L aber habe einen Rang $s > r$ gehabt; die Anzahl der Dimensionen für die Fälle einerseits und die Variablen andererseits könnten also verschieden sein. Können Sie Bedingungen angeben, unter denen die beiden Forscher recht haben?

8. **Aufgabe:** Die rechte Seite der Gleichungen (1) enthält lauter unbekannte Grössen.
   (a) Sie werden mit der Methode der Kleinsten Quadrate geschätzt.
   (b) Sie lassen sich direkt aus den Daten ausrechnen.
   (c) Sie lassen sich aus Hypothesen über die Beziehungen zwischen den Variablen herleiten.

9. **Aufgabe:** Für eine gegebene Datenmatrix X berechnen Sie die Ladungen der Variablen.
   (a) In welcher Beziehung stehen die Vektoren $\mathbf{a}_j$ und $\mathbf{a}_k$ der Ladungen für die Variablen j und k zu den Korrelationen r_{jk}?
   (b) Welche allgemeine Interpretation für die Ladung a_{jk} (j-te Variable, k-te latente Dimension) kennen Sie?
   (c) Welche Implikation hat die Tatsache, dass $r_{jj} = 1$ für alle Variablen $j = 1, \ldots, n$, für die Position der Endpunkte der Ladungsvektoren $\mathbf{a}_j$?
   (d) Der Rang von X sei $r < n$; welchem geometrischen Gebilde entsprechen die Endpunkte der $\mathbf{a}_j$ in diesem Fall?

10. **Aufgabe:** X sei wieder eine $(m \times n)$-Datenmatrix. Welche der folgenden Aussagen ist korrekt, und wenn ja, warum?, und wenn nicht, warum nicht?
    (a) Die SVD $X = Q\Lambda^{1/2}T'$ läßt sich nur berechnen, wenn $m > n$.
    (b) Die Matrix A der Ladungen ist stets quadratisch.
    (c) Welcher Matrix entspricht das Produkt AA', – können Sie Ihre Antwort herleiten?
    (d) Welcher Matrix entspricht das Produkt $A'A$, – können Sie Ihre Antwort herleiten?