Lehrstuhl für Analytics and Statistics
Prof. Dr. Oliver Grothe
SS 2022

Aufgaben zur Vorlesung

Multivariate Verfahren

Übungsblatt VI

---

1. Beantworten Sie folgende Fragen bzw. bearbeiten Sie folgende Arbeitsanweisungen:
   (a) Beschreiben Sie Sinn und Zweck einer Hauptkomponentenanalyse.
   (b) Skizzieren Sie das Vorgehen bei einer Hauptkomponentenanalyse.
   (c) Was sollten Sie wie entscheiden, wenn bei einer Hauptkomponentenanalyse die Varianz der einzelnen Merkmale innerhalb der Stichprobe stark unterschiedlich ist.
   (d) Warum sind möglichst hohe Korrelationen zwischen Merkmalen wichtig für eine Anwendung der Hauptkomponentenanalyse?
2. Leiten Sie die Formel für die Gewichtsvektoren w_i beim Locally Linear Embedding her.
3. Gegeben sei der Vektor $x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ und seine zwei nächsten Nachbarn $\eta_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ und $\eta_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$. Berechnen Sie den Gewichtsvektor w nach dem Locally Linear Embedding Ansatz und stellen Sie den Vektor x als Linearkombination aus seinen zwei nächsten Nachbarn dar.
4. Sie haben einen Auszug von Daten aus einer Pädiatrischen Längsschnittstudie (Institut für Soziale Pädiatrie, München). In der Studie wurden mehr als 1000 Kinder, die im Zeitraum 1971-1973 geboren wurden, über die ersten 5 Lebensjahre hinweg hinsichtlich vieler Kriterien beobachtet.

   Der Auszug besteht aus einer Stichprobe vom Umfang 69 und zeigt 5 Merkmale, die in Spalten angeordnet sind: (1) Größe, (2) Gewicht und (3) Kopfumfang (jeweils bei der Geburt) sowie die Größen von Mutter (4) und Vater (5).

   Folgende Größen lassen sich aus dem Datensetz errechnen.

   Empirische Kovarianzmatrix:

$$\begin{pmatrix} 4.3331 & 0.78123 & 1.9768 & 3.2917 & 1.2836 \\ 0.78123 & 0.22644 & 0.54275 & 0.46313 & 0.21301 \\ 1.9768 & 0.54275 & 2.2506 & 1.9987 & 1.4052 \\ 3.2917 & 0.46313 & 1.9987 & 24.838 & 7.1437 \\ 1.2836 & 0.21301 & 1.4052 & 7.1437 & 35.985 \end{pmatrix}$$

Empirische Korrelationsmatrix:

$$\begin{pmatrix} 1 & 0.78869 & 0.63301 & 0.3173 & 0.10279 \\ 0.78869 & 1 & 0.76029 & 0.19528 & 0.074623 \\ 0.63301 & 0.76029 & 1 & 0.26733 & 0.15614 \\ 0.3173 & 0.19528 & 0.26733 & 1 & 0.23895 \\ 0.10279 & 0.074623 & 0.15614 & 0.23895 & 1 \end{pmatrix}$$

Matrix der normierten Eigenvektoren $\mathbf{C}$ und zugehörige Diagonalmatrix $\mathbf{\Lambda}$ der Eigenwerte:

$$\mathbf{\Lambda} = \begin{pmatrix} 0.16246 & 0 & 0 & 0 & 0 \\ 0 & 0.36223 & 0 & 0 & 0 \\ 0 & 0 & 0.74502 & 0 & 0 \\ 0 & 0 & 0 & 1.1078 & 0 \\ 0 & 0 & 0 & 0 & 2.6225 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 0.49866 & -0.6579 & 0.042084 & 0.1387 & 0.54544 \\ -0.77475 & -0.079831 & -0.13303 & 0.2547 & 0.55751 \\ 0.36926 & 0.73882 & -0.13564 & 0.11125 & 0.53574 \\ -0.11836 & 0.077487 & 0.76594 & -0.55882 & 0.28468 \\ -0.027245 & -0.094537 & -0.61275 & -0.76892 & 0.15368 \end{pmatrix}$$

(a) Soll eine Hauptkomponentenanalyse eher mit der empirischen Kovarianzmatrix oder mit der empirischen Korrelationsmatrix durchgeführt werden?

(b) Finden Sie die Hauptkomponenten des Modells.

(c) Wie viele Komponenten würden Sie benutzen, um die Kinder hinreichend zu charakterisieren?

(d) Interpretieren Sie diese Komponenten soweit möglich.

5. Rechnerübung: Betrachten Sie den Datensatz *pisa*[1] aus Ilias. Dieser enthält die Werte für ,,Lesekompetenz", ,,Mathematische Grundbildung" und ,,Naturwissenschaftliche Grundbildung" der Teilnehmerstaaten der PISA-Studie. Führen Sie eine Hauptkomponentenanalyse durch:

   (a) Soll eine Hauptkomponentenanalyse eher mit der empirischen Kovarianzmatrix oder mit der empirischen Korrelationsmatrix durchgeführt werden?

   (b) Finden Sie die Hauptkomponenten des Modells und interpretieren Sie sie soweit möglich.

   (c) Wie viele Komponenten benötigen Sie?

[1] vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S. 107, 173,229