Lehrstuhl fuer Statistik, Oekonometrie und empirische Wirtschaftsforschung, Universitaet Tuebingen

Hauptkomponentenanalyse

Dr. S. Prohl

8. Juni 2007/ SoSe 2007

Hauptkomponentenanalyse

- Problemstellung der Hauptkomponentenanalyse
- Hauptkomponentenanalyse bei bekannter Varianz-Kovarianz-Matrix
- Hauptkomponentenanalyse bei unbekannter Varianz-Kovarianz-Matrix
  - Anzahl der Hauptkomponenten
  - Ueberpruefung der Guete der Anpassung
- Beispiel

Hauptkomponentenanalyse

Literatur:

- Handl, A.: Multivariate Verfahren: Theorie und Praxis multivariater Verfahren unter besonderer Beruecksichtigung von S-Plus, 2002, Kapitel 5.

Hauptkomponentenanalyse: Vorgehensweise

- Ausgangslage: hochdimensionaler Datensatz, der in einem niedrig-dimensionalen Raum dargestellt werden soll.
  - Alle Merkmale sind quantitativ.
  - Die Daten liegen als Varianz-Kovarianz-Matrix oder Korrelationsmatrix vor.
- Soll die Hauptkomponentenanalyse auf Basis der Varianz-Kovarianz-Matrix, oder auf Basis der Korrelationsmatrix durchgefuehrt werden?
- Man bestimmt die Eigenwerte und Eigenvektoren der Varianz-Kovarianz-Matrix bzw. der Korrelationsmatrix.
- Wie viele Hauptkomponenten benoetigt man?
- Die Hauptkomponenten werden interpretiert.

Hauptkomponentenanalyse: Ausgangslage

Ausgangsgroessen:
Zufallsvektor
$\mathrm{x}^T = (x_1, \ldots, x_p)$
$\mu = E(x)$
$\Sigma = Cov(x)$
bzw. Datenmatrix $\mathbf{x}_1, \ldots \mathbf{x}_n$ mit $\bar{x}$ und empirischer Kovarianzmatrix Σ.

Hauptkomponentenanalyse: Problemstellung

- Reduziere die Dimension von **x** mit moeglichst geringem Informationsverlust
- $\Rightarrow$ Es wird gesucht:
  - Linear-Kombination $y = a^T x$ mit groesstem $V(y) = a^T \Sigma a$

Hauptkomponentenanalyse: Loesung des Problems

Erste Hauptkomponente

- Finde die Loesung:

$$a^T \Sigma a \quad \text{unter der Nebenbedingung} \quad a^T a = 1$$

- Die Lagrange-Funktion lautet:

$$\begin{aligned} L(a,\lambda) &= a^T \Sigma a - \lambda(a^T a - 1) \\ \frac{\partial L(a,\lambda)}{da} &= 2\Sigma a - 2\lambda a \\ \frac{\partial L(a,\lambda)}{d\lambda} &= 1 - a^T a \\ \Rightarrow (\Sigma - \lambda)a &= 0 \end{aligned}$$

- λ ist Eigenwert zu Σ, a ist Eigenvektor

Hauptkomponentenanalyse: Loesung des Problems

Es folgt

$$a^T \Sigma a = a \lambda a = \lambda$$

Es handelt sich um ein Eigenwertproblem!

- Da wir eine Linearkombination mit der groessten Varianz suchen, waehlen wir den Eigenvektor, der zum groessten Eigenwert λ_1 gehoert.
- Erste Hauptkomponente ist $y_1 = a_1^T x$, und a_1 ist der Eigenvektor zum groessten Eigenwert von Σ.

Hauptkomponentenanalyse: Loesung des Problems

Zweite Hauptkomponente:

- Finde die Loesung:

  $a_2^T \Sigma a_2$ unter der Nebenbedingungen $a_1^T a_2 = 0$
- Die Lagrange-Funktion lautet:

$$L(a_2, \lambda, v) = a_2^T \Sigma a_2 - \lambda(a_2^T a_2 - 1) - v(a_2^T a_1)$$

$$\frac{\partial L(a_2, \lambda, v)}{da_2} = 2(\Sigma - \lambda)a_2 - va_1 = 0.$$

$\Rightarrow (\Sigma - \lambda)a_2 = 0.$

- a_2 ist orthogonal zu a_1.
- a_2 ist zweite Hauptkomponente.

Hauptkomponenteanalyse: Zahl der Hauptkomponenten

Wie viele Hauptkomponenten benoetigt man?

- Anteil der Gesamtstreuung, die durch Hauptkomponenten erklaert wird,
- Kaiser-Kriterium ,
- Jolliffe-Kriterium,
- Scree-Plot