

6.Tutorium Multivariate Verfahren

- Hauptkomponentenanalyse -

Andreas Hölzl:

23.06.2014 und 30.06.2014

Shuai Shao:

26.06.2014 und 03.07.2014

Institut für Statistik, LMU München

Gliederung

- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten
- 3 Lösung des Eigenwertproblems
- 4 Geometrische Interpretation
- 5 Anzahl nötiger Hauptkomponenten

Gliederung

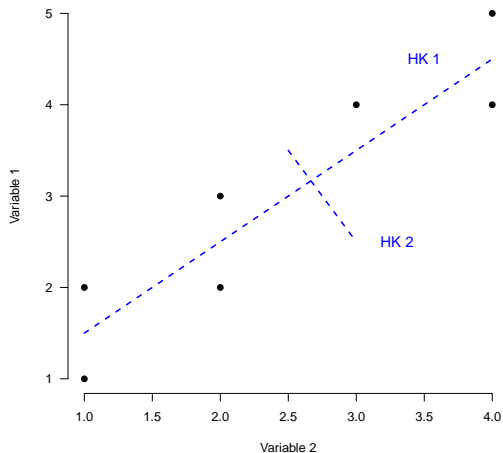
- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten
- 3 Lösung des Eigenwertproblems
- 4 Geometrische Interpretation
- 5 Anzahl nötiger Hauptkomponenten

Problemstellung

- Instrument zur Erklärung der **Variabilität** der Daten
- Ziel: einfache Interpretierbarkeit
- Vereinfachung und Veranschaulichung der Daten
- Reduktion der Daten auf wenige **möglichst aussagekräftige** Hauptkomponenten
⇒ **Fragestellung:** Wie kann die Dimension mit möglichst geringem Informationsverlust reduziert werden?

Information \Leftrightarrow Varianz

Graphische Veranschaulichung



Gliederung

- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten**
- 3 Lösung des Eigenwertproblems
- 4 Geometrische Interpretation
- 5 Anzahl nötiger Hauptkomponenten

1. Hauptkomponente

- Ausgangsgrößen:
 - Zufallsvektor: $\mathbf{x} = (x_1, \dots, x_p)^\top$
 - Erwartungswertvektor: $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$
 - empirische Kovarianzmatrix: $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{x})$
- Finde die Linear-Kombination $y_1 = \mathbf{a}_1^\top \mathbf{x}$ mit Vektor $\mathbf{a}_1 = (a_{11}, \dots, a_{1p})^\top$, sodass $\text{Var}(y_1) = \mathbf{a}_1^\top \boldsymbol{\Sigma} \mathbf{a}_1$ maximal wird
- $y_1 \hat{=} 1$. Hauptkomponente
- **Nebenbedingung:**

$$\|\mathbf{a}_1\|^2 = \mathbf{a}_1^\top \mathbf{a}_1 = 1$$

→ ansonsten ist das Problem nicht eindeutig lösbar!

2. Hauptkomponente

- Finde im **zweiten** Schritt die Linear-Kombination $y_2 = \mathbf{a}_2^\top \mathbf{x}$ mit Vektor $\mathbf{a}_2 = (a_{21}, \dots, a_{2p})^\top$, sodass $\text{Var}(y_2) = \mathbf{a}_2^\top \boldsymbol{\Sigma} \mathbf{a}_2$ maximal wird
- **Nebenbedingungen:**
 - ① $\|\mathbf{a}_2\|^2 = \mathbf{a}_2^\top \mathbf{a}_2 = 1$ (wie vorher)
 - ② $\text{Cov}(y_1, y_2) = 0 \Leftrightarrow \mathbf{a}_1 \perp \mathbf{a}_2$
- **Merke:** Die ersten beiden Hauptkomponenten sind unkorreliert und stehen somit aufeinander senkrecht!

Gliederung

- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten
- 3 Lösung des Eigenwertproblems**
- 4 Geometrische Interpretation
- 5 Anzahl nötiger Hauptkomponenten

Maximierung unter Nebenbedingung (1. HK):

Lagrange:

$$\varphi(\mathbf{a}_1) = \mathbf{a}_1^T \mathbf{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}_1^T \mathbf{a}_1 - 1) \xrightarrow{\mathbf{a}_1} \max$$

$$\frac{\partial \varphi}{\partial \mathbf{a}_1} = 2\mathbf{\Sigma} \mathbf{a}_1 - 2\lambda \mathbf{a}_1 \stackrel{!}{=} 0$$

$$\Leftrightarrow \mathbf{\Sigma} \mathbf{a}_1 = \lambda \mathbf{a}_1 \quad \hat{=} \quad \text{Eigenwertproblem!}$$

$\Rightarrow \mathbf{a}_1$ ist der Eigenvektor zum größten Eigenwert λ von $\mathbf{\Sigma}$

Maximierung unter Nebenbedingung (2. HK):

Lagrange:

$$\varphi(\mathbf{a}_2) = \mathbf{a}_2^T \mathbf{\Sigma} \mathbf{a}_2 - \underbrace{\lambda(\mathbf{a}_2^T \mathbf{a}_2 - 1)}_{1. \text{ NB}} - \underbrace{\varphi(\mathbf{a}_1^T \mathbf{a}_2)}_{2. \text{ NB}} \xrightarrow{\mathbf{a}_2} \max$$

$\Rightarrow \mathbf{a}_2$ ist der Eigenvektor zum zweitgrößten Eigenwert λ von $\mathbf{\Sigma}$

Merke: Es können maximal p Hauptkomponenten bestimmt werden, was der Anzahl an betrachteten Variablen entspricht!

Spektralzerlegung

- Man erhält den Vektor der Hauptkomponenten mittels Spektralzerlegung von $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \quad \text{mit}$$

$\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_p)$ Eigenvektoren von $\mathbf{\Sigma}$
(entsprechen den gesuchten $(\mathbf{a}_1, \dots, \mathbf{a}_p)$)

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \quad \text{Eigenwerte } \lambda_1 \geq \dots \geq \lambda_p$$

- Kovarianz der Hauptkomponentenanalyse:

$$\text{Cov}(\mathbf{y}) = \text{Cov}(\mathbf{P}^T \mathbf{x}) = \mathbf{P}^T \mathbf{\Sigma} \mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

Gliederung

- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten
- 3 Lösung des Eigenwertproblems
- 4 Geometrische Interpretation**
- 5 Anzahl nötiger Hauptkomponenten

Geometrische Interpretation

Darstellung einer Beobachtung \mathbf{x} im Koordinatensystem:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \dots + x_p \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

d.h. x_1, \dots, x_p sind die Koordinaten von \mathbf{x} zur **kanonischen** Basis

$$\mathbf{y} = \mathbf{P}^\top \mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{P} \mathbf{y}$$

$$\mathbf{x} = (\mathbf{p}_1 \dots \mathbf{p}_p) \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = y_1 \mathbf{p}_1 + \dots + y_p \mathbf{p}_p$$

d.h. y_1, \dots, y_p sind die Koordinaten von \mathbf{x} zur Basis $\mathbf{p}_1, \dots, \mathbf{p}_p$

Bemerkungen

- Die Hauptkomponenten stellen neue Basisvektoren dar
- Die Hauptkomponenten haben die Richtung der Hauptachsen der zu Σ gehörigen Ellipsen
- Die Größe der Eigenwerte entspricht der Varianz der Hauptkomponenten und ist somit proportional zur Länge der Basisvektoren

Merke:

Bei Skalenänderung ändern sich die Eigenwerte und Eigenvektoren!
⇒ meist Verwendung **standardisierter** Daten bzw. der Korrelationsmatrix!

Gliederung

- 1 Idee der Hauptkomponentenanalyse
- 2 Bestimmung der Hauptkomponenten
- 3 Lösung des Eigenwertproblems
- 4 Geometrische Interpretation
- 5 Anzahl nötiger Hauptkomponenten

Erklärte Varianz der Hauptkomponenten:

- Gesamtvarianz:

$$\text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^p \text{Var}(x_i)$$

- Anteil erklärter Varianz durch r Hauptkomponenten

$$\text{eV}(r) = \frac{\sum_{i=1}^r \text{Var}(y_i)}{\sum_{i=1}^p \text{Var}(y_i)} = \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i}$$

Der **Anteil der Varianz**, den die Hauptkomponenten erklären, ist ein ausschlaggebendes Kriterium für die **Anzahl** zu verwendender Hauptkomponenten!

Mögliche Kriterien

- 1 Verwende so viele Hauptkomponenten bis ein gewisser Anteil $x\%$ der Gesamtvarianz der Daten erklärt ist
- 2 Eine Hauptkomponente sollte mindestens genauso viel beitragen wie eine einzelne Variable im Durchschnitt. Im Fall normierter Variablen (Korrelationsmatrix) sind dies alle Hauptkomponenten mit Eigenwert $\lambda > 1$.
- 3 Betrachte eine graphische Darstellung der Eigenwerte (Scree-Plot). Verwende so viele Hauptkomponenten bis zum Knick des Graphen (Ellbogen).