

Literatur

Everitt und Hothorn: An Introduction to Applied Multivariate Analysis with R. Springer 2011, Kap. 3

- Karl Pearson (1901): Lines and planes of closest fit to data
- Hotelling (1933): Verwendung im Zusammenhang mit der Faktorenanalyse
- Verfahren zur Dimensionsreduktion und zur explorativen Analyse der Korrelationsstruktur
- Anwendungen in der Psychologie zur Bildung von Scores
- Relation zur Faktorenanalyse: Konzeptionell verschieden, kann aber als Lösungsstrategie im Faktormodell gesehen werden

Fragestellung

Wie kann die Dimension von \mathbf{x} mit möglichst geringem Informationsverlust reduziert werden?

Information \Leftrightarrow Varianz

\Rightarrow Finde Linear-Kombination $y = a^T x$ mit $V(y) = a^T \Sigma a$ maximal
 y ist dann die 1. Hauptkomponente

Löse

$a^T \Sigma a$ maximal unter der Nebenbedingung $a^T a = 1$

$$L(a) = a^T \Sigma a - \lambda(a^T a - 1)$$

$$\frac{\partial L}{\partial a} = 2\Sigma a - 2\lambda a$$

$$\Rightarrow (\Sigma - \lambda I)a = 0$$

$\Rightarrow \lambda$ Eigenwert zu Σ , a Eigenvektor

Es folgt

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

Maximierungsproblem wird für den größten Eigenwert λ_1 gelöst.

\Rightarrow Die 1. Hauptkomponente ist $y_1 = a_1^T x$, wobei a_1 der Eigenvektor zum größten Eigenwert von Σ ist.

Beispiel zur Berechnung der ersten Hauptkomponente

$$\Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

$$\lambda_1 = 3/2 \quad \lambda_2 = 1/2$$

$$a_1^T = \left(1/\sqrt{2}, 1/\sqrt{2} \right)$$

$$y_1 = 1/\sqrt{2} x_1 + 1/\sqrt{2} x_2$$

$$V(y_1) = \frac{3}{2}$$

Zweite Hauptkomponente

Falls ein Vektor nicht ausreicht, um die Information der Daten zu repräsentieren

⇒ Wahl einer weiteren Linearkombination

$$\begin{aligned} y_2 &= a_2^T x && \text{mit } \text{Cov}(y_1, y_2) = 0 \\ &&& \Leftrightarrow \text{Cov}(a_1^T x, a_2^T x) = 0 \\ &&& \Leftrightarrow a_1^T a_2 = 0 \\ &&& \text{und } a_2^T a_2 = 1 \end{aligned}$$

⇒ Löse $a_2^T \Sigma a_2$ maximal unter obigen Nebenbedingungen:

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

$$\frac{\partial L}{\partial a_2} = 2(\Sigma - \lambda_2 I)a_2 - \delta a_1 = 0.$$

...

$$\Rightarrow (\Sigma - \lambda_2 I)a_2 = 0$$

⇒ λ_2 ist der zweitgrößte Eigenwert und a_2 der dazugehörige Eigenvektor.

Hauptkomponentenzerlegung

Fortsetzung des Verfahrens bis zur p-ten Hauptkomponente:

Man erhält den Vektor der Hauptkomponenten mittels Spektralzerlegung von Σ .

Sei A die Matrix der normierten Eigenvektoren von Σ ,

Λ die Matrix der der Größe nach geordneten Eigenwerte von Σ

$$\begin{aligned} A &= (a_1, a_2, \dots, a_p) \\ AA^T &= I \\ \Sigma &= A\Lambda A^T \\ \Lambda &= A^T \Sigma A \\ Y &= A^T x \end{aligned}$$

Es gilt für die Hauptkomponenten:

$$\begin{aligned} \text{Var}(y_k) &= a_k^T \Sigma a_k = \lambda_k a_k^T a_k = \lambda_k \\ \sum_{k=1}^p \text{Var}(y_k) &= \sum_{k=1}^p \lambda_k = \text{Spur}(\Lambda) \\ &= \text{Spur}(A^T \Sigma A) = \text{Spur}(\Sigma A^T A) = \text{Spur}(\Sigma) = \sum_{k=1}^p \text{Var}(x_k) \end{aligned}$$

⇒ Anteil der Varianz, der durch die ersten s Hauptkomponenten erklärt wird:

$$\frac{\sum_{j=1}^s \lambda_j}{\sum_{k=1}^p \lambda_k}$$

Geometrische Interpretation

Hauptkomponenten haben die Richtung der Hauptachsen der zu Σ gehörigen

Ellipsen:

$$x^T \Sigma^{-1} x = c$$

Die Größe der Eigenwerte ist proportional zu den Längen der Hauptachsen.

Richtungen mit kurzen Hauptachsen sind evtl. vernachlässigbar

Zahl der nötigen Hauptkomponenten

Anzahl festlegen mit

- Varianzaufklärung wichtiges Kriterium:
Eine Hauptkomponente sollte mindestens genauso viel zur Varianzaufklärung beitragen wie eine einzelne Variable im Durchschnitt. Im Fall der normierten Variablen (Korrelationsmatrix) sind dies alle Hauptkomponenten mit Eigenwert > 1 .
- Graphische Darstellung der Eigenwerte gegen die Nummer der Hauptkomponente (Scree-Plot) und beim Ellenbogenaufhören

Voraussetzung: Skalierung

- Die Verwendung der Hauptkomponentenanalyse setzt gleiche Skalierung voraus. Variablen mit höherer Varianz erhalten meist ein höheres Gewicht.
- Häufig werden die Variablen vorher normiert. Dies entspricht der Analyse der **Korrelationsmatrix** statt der Kovarianzmatrix.

Rotation von Hauptkomponenten

- Häufig sind bei Lösungen mit mehreren Hauptkomponenten Interpretationen problematisch
- Versuch, eine bessere Interpretation durch Rotation der Achsen innerhalb der gewählten Hauptkomponenten zu bekommen.
- Es entstehen neue Hauptkomponenten, die die gleiche Varianzaufklärung haben.

- Daten als Scatterplot für die ersten beiden Hauptkomponenten
- Gewichte der einzelnen Variablen als Scatterplot (Jede Variable entspricht einem Punkt)
- Beides zusammen in einen Biplot

Probleme bei der Anwendung

- Exploratives Verfahren
- Gewichte der Hauptkomponenten nur durch Korrelation sollten auf inhaltliche Probleme geprüft werden
- Bei Verwendung mehrerer Hauptkomponenten ist die Annahme der Unabhängigkeit bisweilen problematisch (Unabhängigkeit inhaltlich teilweise nicht plausibel)
- Vorsicht bei Hinzunahme sehr ähnlicher Variablen
- Struktur der Gewichte liefert häufig interessante Aussagen zur Zusammenhangsstruktur der Daten
- Auf Skalierung achten! In der Regel arbeitet man mit der Korrelationsmatrix