Reader – Teil 7: Hauptkomponentenanalyse

Dr. Katharina Best

Die Hauptkomponentenanalyse, auch PCA von *principal component analysis*, dient der Komplexitätsreduktion. Sie hilft, umfangreiche vieldimensionale, sagen wir r, Datensätze zu strukturieren und zu vereinfachen. Diese können dann leichter veranschaulicht und auch untersucht werden.

1 Grundidee

Wir projezieren die Daten in den k-dimensionalen Unterraum. Um zu gewährleisten, dass auf diese Weise möglichst wenig Information verloren geht, wollen wir nicht einen beliebigen Unterraum verwenden. Stattdessen suchen wir uns denjenigen aus, der die meiste Information noch in sich trägt.

Mathematisch benötigen wir erst einmal eine Basistransformation des Vektorraums, welche die Daten dekorreliert. Im Anschluss werden dann die Koordinaten mit der geringsten Varianz fallen gelassen.

2 Verfahren

Wir betrachten eine Stichprobe $x_1, \ldots, x_n$ der Länge n, jede Beobachtung sei r-dimensional. Im weiteren wollen wir eine Transformation der Daten an Hand der Varianzen vornehmen. Zu diesem Zweck nutzen wir die empirische Kovarianzmatrix

$$S := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)(x_i - \overline{x}_n)^T, \tag{1}$$

welche wir später diagonalisieren wollen.

2.1 Diagonalisierung von Matrizen

Da die empirische Kovarianzmatrix S quadratisch, symmetrisch und positiv semidefinit ist, können wir sie diagonalisieren, d. h. es existiert eine eine Diagonalmatrix Λ, welche die Eigenwerte von S geordnet

enthält, d. h.

$$\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r), \tag{2}$$

wobei

$$\lambda_1 > \ldots > \lambda_r, \tag{3}$$

und eine Orthogonalmatrix $\boldsymbol{T}$, die aus den Eigenvektoren $v_1, \ldots, v_r$ von $\boldsymbol{S}$ besteht. Weiterhin gilt

$$\boldsymbol{S} = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^T, \tag{4}$$

da $\boldsymbol{T}^T = \boldsymbol{T}^{-1}$ ist.

2.2 Hauptkomponenten von S

Die Richtungen zu den Daten $x_1, \ldots x_n$ sind durch die gefundenen Eigenvektoren $v_1, \ldots, v_r$ gegeben. Sie werden auch als *Hauptkomponenten* bezeichnet. Die zugehörige Abbildung

$$y(z) = T^T(z - \overline{x}_n) \tag{5}$$

für alle $z \in \mathbb{R}^r$, welche die Daten transformiert, nennen wir *Hauptkomponententransformation*. Zu einem Vektor $z \in \mathbb{R}^r$ haben wir mit $y(z)$ seine Koordinaten bezüglich des Koordinatensystems mit Mittelpunkt $\overline{x}_n$ und der Orthonormalbasis $v_1, \ldots, v_r$.

2.3 Streuung

Wenn wir uns die transformierten Daten $y(x)$ ansehen, so erhalten wir für das Stichprobenmittel jeder Koordinate

$$(\overline{y}_n)_i = \frac{1}{n}\sum_{j=1}^{n} y_i(x_j) = \frac{1}{n}\sum_{j=1}^{n} v_i^T(x_j - \overline{x}_n) = \frac{1}{n} v_i^T(\overline{x}_n - \overline{x}_n) = 0 \tag{6}$$

wegen der Verschiebung des Koordinatensystems in den Punkt $\overline{x}_n$. Das vereinfacht uns die Berechnung der empirischen Kovarianzmatrix, die

$$\boldsymbol{S}_y = \operatorname{diag}(\lambda_1, \ldots, \lambda_r) \tag{7}$$

ist wegen

$$\begin{aligned}
\frac{1}{n-1}\sum_{j=1}^{n} y_i^2(x_j) &= \frac{1}{n-1}\sum_{j=1}^{n} v_i^T(x_j - \overline{x}_n) v_i^T(x_j - \overline{x}_n) \\
&= \frac{1}{n-1}\sum_{j=1}^{n} v_i^T(x_j - \overline{x}_n)(x_j - \overline{x}_n)^T v_i \\
&= v_i^T \left(\frac{1}{n-1}\sum_{j=1}^{n} (x_j - \overline{x}_n)(x_j - \overline{x}_n)^T \right) v_i \\
&= v_i^T \boldsymbol{S} v_i = \left(\boldsymbol{T}^T \boldsymbol{S} \boldsymbol{T} \right)_i = (\boldsymbol{D})_i = \lambda_i
\end{aligned} \tag{8}$$

auf der Diagonalen für $j = 1, \ldots, r$ und

$$\frac{1}{n-1}\sum_{j=1}^{n} y_i(x_j) y_k(x_j) = 0 \tag{9}$$

in den Kovarianzen analog.

Daraus folgt, dass die Daten $x_1, \ldots, x_n$ ihre größte Streuung bezüglich ihrer ersten Hauptkomponente haben, die zweitgrößte bezüglich der zweiten Hauptkomponente usw.

Überlegen Sie sich, wie Sie den Beweis angehen würden.

2.4 Dimensionsreduktion

Mit diesen Ergebnissen begegnen wir wieder der ursprünglichen Fragestellung. Wir wollen den r-dimensionalen Raum in einen k-dimensionalen Raum projezieren, möglichst wenig Information verloren geht.

Betrachten wir einen Eigenwert λ_{k+1}. Aus (8) wissen wir, dass die Hauptkomponenten $y_{k+1}(x_j)$ mit $j = 1, \ldots, n$ nur sehr wenig streuen. Wegen (3) können wir das gleiche über $y_i(x_j)$ mit $j = 1, \ldots, n$ für alle $i = k+1, \ldots, r$ sagen. Die transformierten Daten $y(x_j)$ mit $j = 1, \ldots, n$ sind demnach in den letzten $r - k$ Koordinaten nahezu Null. Die Abbildung $w : n\mathbb{R}^r \to n\mathbb{R}^k$ mit

$$w(z) = \big(y_1(z), \ldots, y_k(z)\big)^T = \Big(v_1^T(z - \overline{x}_n), \ldots, v_k^T(z - \overline{x}_n)\Big)^T \tag{10}$$

liefert uns eine fast verlustfreie Dimensionsreduktion.

Was uns noch fehlt, ist ein Kriterium zur Bestimmung von k. Dazu wird die Gesamtstreuung

$$\frac{1}{n-1} \sum_{i=1}^{n} \|x_j - \overline{x}_n\|_2^2 = \text{Spur } \boldsymbol{S} = \text{Spur } \boldsymbol{D} = \sum_{i=1}^{r} \lambda_r, \tag{11}$$

auch *Gesamtvariation resp. total variation* genannt, mit der in den k Dimensionen erhaltenen verglichen,

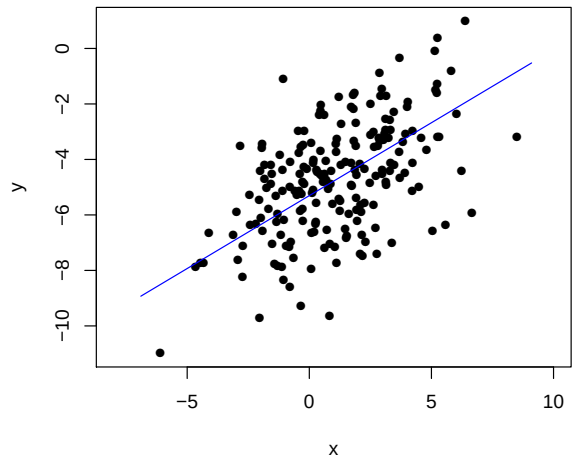
$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_r} \tag{12}$$

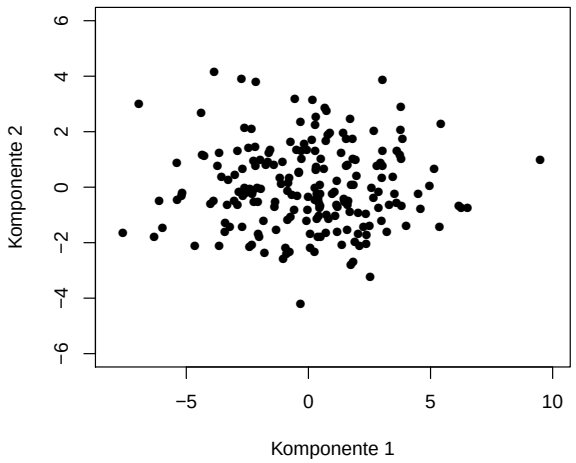
und k ensprechend eines selbst vorgegebenen Prozentsatzes bestimmt.

Eine weitere Möglichkeit bietet der Scree-test, welcher die Eigenwerte gegen ihre Indizes angibt. Damit ergibt sich eine visuelle Möglichkeit, einen *natürlich sich ergebenden* Wert k zu ermitteln.

3 Geometrische Interpretation

Im zweidimensionalen Fall schätzt die Hauptkomponentenanalyse eine Gerade durch die Daten, die erste Hauptkomponente.





Ein uns bereits bekanntes Verfahren, die Regressionsanalyse, liefert ebenfalls eine Gerade.

Machen Sie sich Gedanken darüber, wo hierbei die Ähnlichkeiten und Unterschiede zur Regression sind.