# Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (PCA = Principal Component Analysis) ist ein multivariates Analyseverfahren mit dem Ziel, die Zeilen einer Ausgangsdatenmatrix in einem kartesischen Koordinatensystem geringer Dimensionalität im Sinn der „kleinsten Quadrate“ darzustellen.

Bei den Ausgangsdaten handelt es sich um Objekte, für die jeweils mehrere Merkmale – üblicherweise auf metrischem Skalenniveau - erhoben wurden. Die Objekte werden zeilenweise, die Merkmale spaltenweise angeordnet.

Beispiel: „Sicherheitsmängel bei Autos der unteren Mittelklasse“, HAZ vom 11.08.1990, S. 21:

| | Fahrgastzelle | Lenkrad | Armaturenbrett | Fußraum | Seitenverkleid. | Vordersitze | Rücksitze |
|---|---|---|---|---|---|---|---|
| **Renault 19 TR 1,4** | 2,1 | 2,6 | 2,0 | 2,5 | 3,8 | 2,3 | 3,3 |
| **Toyota Corolla Compact 1,3 XLi** | 1,6 | 3,0 | 2,3 | 3,5 | 3,0 | 2,8 | 2,2 |
| **VW Golf CL 1,3** | 2,6 | 2,3 | 2,8 | 2,8 | 3,0 | 2,2 | 2,7 |
| **Opel Vektra GL 1,6** | 2,9 | 2,0 | 2,1 | 2,5 | 3,5 | 2,2 | 2,3 |
| **Fiat Tipo 1400** | 3,8 | 4,1 | 3,5 | 3,5 | 3,8 | 3,3 | 3,3 |
| **Ford Escord C 1,4 i** | 4,6 | 3,6 | 3,1 | 3,0 | 3,5 | 2,7 | 3,2 |
| **Opel Kadett LS 1,4 i** | 2,0 | 2,8 | 2,1 | 2,8 | 3,3 | 2,8 | 3,3 |

Die Ausgangsdatenmatrix wird mit $\underline{X}$ bezeichnet und ist vom Format (I, J):

$$(1.1) \qquad \underline{X} = (x_{ij}) = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}$$

Bei den Zeilen von $\underline{X}$ handelt es sich um die sog. Zeilenprofile der Objekte. Man betrachtet zunächst das mittlere Zeilenprofil, also das arithmetische Mittel für jede Spalte. Die Matrix $\overline{\underline{X}}$, die I-mal untereinandergeschrieben das mittlere Zeilenprofil enthält, ist

$$(1.2) \qquad \underline{\overline{X}} = \begin{pmatrix} \underline{\overline{x}} & \rightarrow \\ \vdots & \\ \underline{\overline{x}} & \rightarrow \end{pmatrix}.$$

Man kann diese Matrix rechnerisch einfach erhalten, indem man eine Matrix $\underline{1}_{II}$ vom Format (I, I) bildet, die nur aus Einsen besteht. Dann ist

$$(1.3) \qquad \underline{\overline{X}} = \frac{1}{I}\underline{1}_{II}\underline{X}.$$

Nun wird die Matrix $\underline{A}$ benötigt, die die Abweichungen der Zeilenprofile vom mittleren Zeilenprofil enthält:

$$(1.4) \qquad \underline{A} = \underline{X} - \underline{\overline{X}}.$$

Die Aufgabe besteht nun darin, eine möglichst gute Darstellung der Zeilen von $\underline{A}$ in einem kartesischen Koordinatensystem geringer Dimensionalität im Sinn der „kleinsten Quadrate“ zu finden.

$\underline{e}$ sei ein Einheitsvektor mit J Komponenten. Dann ist $\underline{Ae}$ der Spaltenvektor (mit J Komponenten), der die Projektionen der Zeilenvektoren von $\underline{A}$ auf $\underline{e}$ enthält. Die Summe der quadrierten Projektionen ist

$$(1.5) \qquad (\underline{Ae})'\underline{Ae} = \underline{e}'\underline{A}'\underline{Ae}.$$

$\underline{A}'\underline{A}$ ist eine symmetrische Matrix vom Format (J, J) mit Rang r und positiv-semidefinit[1]. Die Zahl $\underline{e}'\underline{A}'\underline{Ae}$ ist durch die Wahl des Einheitsvektors $\underline{e}$ zu maximieren. Hierbei handelt es sich um ein Standardproblem der linearen Algebra, das auf eine Eigenwertzerlegung von $\underline{A}'\underline{A}$ führt. Die Eigenwertzerlegung liefert

[1] Eine positiv-semidefinite Matrix enthält nur nicht-negative Eigenwerte.

$$(1.6) \qquad \underline{A}'\underline{A} = \underline{E\Lambda E}' = \begin{pmatrix} \downarrow & & \downarrow \\ \underline{e}_1 & \cdots & \underline{e}_J \\ \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_J \end{pmatrix} \begin{pmatrix} \rightarrow & \underline{e}_1 & \rightarrow \\ & \vdots & \\ \rightarrow & \underline{e}_J & \rightarrow \end{pmatrix}.$$

Die Eigenwerte $\lambda_1, ..., \lambda_J$ werden in absteigender Reihenfolge angeordnet. Ist $r < J$, so erhält man $(J - r)$ Eigenwerte der Größe Null bzw. r Eigenwerte > 0.

Zur Darstellung in einem K-Dimensionalen kartesischen Koordinatensystem (üblicherweise ist K = 2) werden nur die K größten Eigenwerte und die zugehörigen Eigenvektoren beibehalten. Die Koordinaten liefern $\underline{Ae}_1, ..., \underline{Ae}_K$, zusammen:

$$(1.7) \qquad \underline{AE}_K.$$

Da üblicherweise $K < r$, wird nicht die gesamte Information im K-Dimensionalen Koordinatensystem erfaßt. Um zu beurteilen, wie groß der Anteil der erfaßten Information ist, wird die Güte der Anpassung angegeben. Die Güte der Anpassung ist der Anteil der durch die K Dimensionen erfaßte Summe der quadrierten Längen der Zeilen von $\underline{A}$:

$$(1.8) \qquad \Gamma = \frac{\lambda_1 + ... + \lambda_K}{\lambda_1 + ... + \lambda_r}$$

Entsprechend läßt sich die Güte angeben, die jede einzelne Dimension (Achse) liefert.

Beispiel: „Sicherheitsmängel bei Autos der unteren Mittelklasse“, HAZ vom 11.08.1990, S. 21:

Ausgangsdatenmatrix $\underline{X}$

```
X =

    2.1000    2.6000    2.0000    2.5000    3.8000    2.3000    3.3000
    1.6000    3.0000    2.3000    3.5000    3.0000    2.8000    2.2000
    2.6000    2.3000    2.8000    2.8000    3.0000    2.2000    2.7000
    2.9000    2.0000    2.1000    2.5000    3.5000    2.2000    2.3000
    3.8000    4.1000    3.5000    3.5000    3.8000    3.3000    3.3000
    4.6000    3.6000    3.1000    3.0000    3.5000    2.7000    3.2000
    2.0000    2.8000    2.1000    2.8000    3.3000    2.8000    3.3000
```

Bestimmung der Matrix $\overline{\underline{X}}$

```
Xquer = (1/I)*ones(I,I)*X

    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
    2.8000    2.9143    2.5571    2.9429    3.4143    2.6143    2.9000
```

Bestimmung der Matrix $\underline{A}$

```
A =

   -0.7000   -0.3143   -0.5571   -0.4429    0.3857   -0.3143    0.4000
   -1.2000    0.0857   -0.2571    0.5571   -0.4143    0.1857   -0.7000
   -0.2000   -0.6143    0.2429   -0.1429   -0.4143   -0.4143   -0.2000
    0.1000   -0.9143   -0.4571   -0.4429    0.0857   -0.4143   -0.6000
    1.0000    1.1857    0.9429    0.5571    0.3857    0.6857    0.4000
    1.8000    0.6857    0.5429    0.0571    0.0857    0.0857    0.3000
   -0.8000   -0.1143   -0.4571   -0.1429   -0.1143    0.1857    0.4000
```

Eigenwertzerlegung der Matrix A´A

```
[E,L] = eig(A'*A)

E =

    0.0660    0.2447    0.1388    0.1718   -0.1472   -0.5756    0.7308
   -0.5050    0.1757   -0.4137    0.1969    0.1415    0.5227    0.4593
    0.1048   -0.5872   -0.1961   -0.6035   -0.2793    0.1028    0.3910
   -0.1297   -0.0273    0.8011    0.0674   -0.3296    0.4546    0.1446
   -0.0126   -0.7189    0.1259    0.5243    0.4156   -0.0813    0.1143
    0.8435    0.1370   -0.1223    0.2182    0.0541    0.4122    0.1855
    0.0354    0.1677    0.3141   -0.4904    0.7720    0.0406    0.1838


L =

    0.0770         0         0         0         0         0         0
         0    0.2025         0         0         0         0         0
         0         0    0.0000         0         0         0         0
         0         0         0    0.4286         0         0         0
         0         0         0         0    1.6471         0         0
         0         0         0         0         0    3.1419         0
         0         0         0         0         0         0   10.7829
```

Bestimmung der Koordinaten zu Darstellung in einem 2-dimensionalen kartesischen Koordinatensystem

```
K2 =

   -0.8785   -0.1647
   -0.9991    1.0442
   -0.5149   -0.3912
   -0.7669   -0.9859
    1.9694    0.6620
    1.9317   -0.5553
   -0.7416    0.3909
```

Bestimmung der Güte der Anpassung (erste Achse, zweite Achse, gesamt)

```
G1 =

    0.6623

G2 =

    0.1930

G =

    0.8553
```

# Graphische Darstellung