Hauptkomponenten-analyse

Die Hauptkompontenanalyse zählt zu den bedeutendsten Verfahren der Faktorenanalyse. Anhand eines beinahe aktuellen Beispiels wird gezeigt, was eine Hauptkomponente ausmacht und wie sie bestimmt wird.

Johannes Lüken, Diplom-Psychologe, ist Leiter des Bereichs Data Sciences bei IfaD.

jlueken@ifad.de

Prof. Dr. Heiko Schimmelpfennig, Diplom-Kaufmann, ist Projektleiter für Data Sciences bei IfaD.

hschimmelpfennig@ifad.de

Grundidee einer Hauptkomponente

Die Tabelle enthält ausgewählte Daten der Qualifikationsspiele der deutschen Fußball-Nationalmannschaft zur EM 2020/21. Ballbesitz, Zweikampfquote und Anzahl Ecken sollen – wenn möglich – zu einem Merkmal „Dominanz" zusammengefasst werden. Da die Ecken auf einer anderen Skala als die anderen beiden Variablen erhoben werden, bietet sich zunächst eine Standardisierung an. Anschließend wird die Dominanz durch einen gewichteten Mittelwert bestimmt:

Dominanz = a_1 · Ballbesitz (standardisiert) + a_2 · Zweikampfquote (standardisiert) + a_3 · Anzahl Ecken (standardisiert)

Ballbesitz variiert deutlich mehr als Zweikampfquote, das heißt differenziert besser zwischen den Spielen (allgemein zwischen den Fällen). Daraus leitet sich das Ziel ab, die Gewichte a_1, a_2 und a_3 so zu bestimmen, dass die Werte der Dominanz möglichst große Varianz aufweisen. Allerdings steigt die Varianz grundsätzlich mit höheren Gewichten an, da die Beträge der resultierenden Werte dadurch zwangsläufig größer werden. Insofern ist eine Normierung erforderlich. Üblicherweise wird gefordert, dass die Summe der quadrierten Gewichte gleich eins ist.

Extraktion der Hauptkomponenten

Die Varianz der Dominanz kann auch mithilfe der Korrelationsmatrix der drei Variablen berechnet werden. Sie ist gleich

$$(a_1 \quad a_2 \quad a_3) \cdot \begin{pmatrix} 1 & 0{,}373 & 0{,}917 \\ 0{,}373 & 1 & 0{,}293 \\ 0{,}917 & 0{,}293 & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

Die Gewichte a_1, a_2 und a_3 sind folglich so zu bestimmen, dass dieses Produkt maximal wird. Unter der Nebenbedingung $a_1^2 + a_2^2 + a_3^2 = 1$ führt die Optimierungsaufgabe zu einem Eigenwertproblem. Der Eigenvektor zu dem größten Eigenwert der Korrelationsmatrix ist die erste Hauptkomponente und entspricht den gesuchten Gewichten. Diese sind $a_1 = 0{,}659$, $a_2 = 0{,}389$ und $a_3 = 0{,}644$. Der Eigenwert 2,116 selbst ist gleich der Varianz der Werte der Dominanz, die der letzten Spalte der Tabelle zu entnehmen sind. Die Dominanz der deutschen Fußballnationalmannschaft war im Heimspiel gegen Estland am höchsten und in den beiden Spielen gegen die Niederlande am geringsten. Der durch die erste extrahierte Hauptkomponente erklärte Anteil der Varianz der drei Variablen beträgt 2,116 / 3 = 70,5 %.

Die Korrelation zwischen den Werten einer Variable und denen der Hauptkomponente kennzeichnet die Ladung der Variable auf die Hauptkomponente. Diese ist für den Ballbesitz gleich 0,958, für die Zweikampfquote gleich 0,566 und für die Anzahl Ecken gleich 0,936. In den Ladungen spiegelt sich die hohe Korrelation zwischen Ballbesitz und Anzahl Ecken wider. Beide laden höher als die Zweikampfquote auf die erste Hauptkomponente. Dennoch ist die Ladung der Zweikampfquote hoch genug, dass auch sie zu dieser Hauptkomponente gehört und mit zur Dominanz beiträgt – wenn auch etwas weniger als die anderen beiden Variablen.

Prinzipiell können so viele Hauptkomponenten extrahiert werden, wie beobachtete Variablen in die Analyse eingehen. Die weiteren Hauptkomponenten werden sukzessive so bestimmt, dass sie möglichst viel von der noch nicht erklärten Varianz erklären und orthogonal zu den übrigen Hauptkomponenten sind, das heißt die Werte der Hauptkomponenten nicht miteinander korrelieren.

In Ausgabe 3/2020:
Hauptkomponentenanalyse zur Datenreduktion

Literatur

Handl, A.: *Multivariate Analysemethoden*, 2. Auflage, Berlin, Heidelberg, 2010, S. 115-147.

Daten der Qualifikationsspiele der deutschen Nationalmannschaft zur Fußball-EM 2020/21

	Ballbesitz	Zweikampfquote	Anzahl Ecken	Ballbesitz (standardisiert)	Zweikampfquote (standardisiert)	Anzahl Ecken (standardisiert)	1. Hauptkomponente: Dominanz
Niederlande (A)	0,46	0,55	1	-1,660	0,374	-1,557	-1,950
Weißrussland (A)	0,75	0,47	8	0,605	-1,335	0,338	0,097
Estland (H)	0,82	0,58	11	1,152	1,014	1,151	1,895
Niederlande (H)	0,49	0,46	2	-1,425	-1,548	-1,286	-2,370
Nordirland (A)	0,74	0,59	7	0,527	1,228	0,068	0,869
Estland (A)	0,69	0,55	6	0,137	0,374	-0,203	0,105
Weißrussland (H)	0,71	0,53	11	0,293	-0,053	1,151	0,913
Nordirland (H)	0,72	0,53	8	0,371	-0,053	0,338	0,441
Varianz	0,016	0,002	13,643	1	1	1	2,116