Gewinnung von Standardkoordinaten

1. Aufgabe

Gegeben ist eine Kontingenztabelle, deren Inhalt durch die Matrix $\underline{N} = (n_{ij})$ mit i = 1, ..., I und j = 1, ..., J und den eingetragenen Häufigkeiten n_{ij} dargestellt wird. Es ist

(1.1)
$$\underline{N} = \begin{array}{c} \\ x_1 \\ . \\ . \\ . \\ x_I \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & . & . & . \end{array} \quad y_J \\ \left[\begin{array}{cccc} n_{11} & . & . & . \end{array} \quad n_{1J}\right] \\ \left[\begin{array}{cccc} . & & & \end{array} \quad .\right] \\ \left[\begin{array}{cccc} . & & & \end{array} \quad .\right] \\ \left[\begin{array}{cccc} . & & & \end{array} \quad .\right] \\ \left[\begin{array}{cccc} n_{I1} & . & . & . \end{array} \quad n_{IJ}\right] \\ \begin{array}{cccc} n_{.1} & . & . & . \end{array} \quad n_{.J} \end{array} \begin{array}{c} \\ n_{1.} \\ . \\ . \\ . \\ n_{I.} \\ \\ \end{array}$$

mit den Randhäufigkeiten $n_{i.}$ und $n_{.j}$ und $n = \sum_i \sum_j n_{ij}$. In $\underline{N}$ kommen weder Nullzeilen noch Nullspalten vor. Den Zeilen sollen Skalenwerte oder *Koordinaten* x_1, ..., x_I, den Spalten Skalenwerte oder *Koordinaten* y_1, ..., y_J zugeordnet werden. Dadurch wird einem Fall, der zur Häufigkeit n_{ij} beiträgt, das Paar (x_i, y_j) zugeordnet.

Durch die Koordinaten sollen die Inhalte der Datentafel so gut wie möglich repräsentiert werden.

Verhältnismäßig einfach lassen sich diese Koordinaten bestimmen, wenn man einem Ansatz von Kristof folgt, in dem zunächst davon ausgegangen wird, daß *ein Satz von Koordinaten bereits gegeben ist*.

2. Bestimmung von Koordinaten y_j bei gegebenen Koordinaten x_i

$$(2.1)\quad \underline{N} = \begin{array}{c} \\ x_1 \\ . \\ . \\ . \\ x_I \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & . & . & . \end{array} \quad y_J \\ \begin{bmatrix} n_{11} & . & . & . \end{bmatrix} \quad n_{1J} \\ \begin{bmatrix} . & & & \end{bmatrix} \quad . \\ \begin{bmatrix} . & & & \end{bmatrix} \quad . \\ \begin{bmatrix} . & & & \end{bmatrix} \quad . \\ \begin{bmatrix} n_{I1} & . & . & . \end{bmatrix} \quad n_{IJ} \\ \begin{array}{cccc} n_{.1} & . & . & . \end{array} \quad n_{.J} \end{array} \begin{array}{c} \\ n_{1.} \\ . \\ . \\ . \\ n_{I.} \\ \\ \end{array}$$

Aufgabe: Die x_i seien *gegeben*, die y_j seien gesucht.

Die gesuchten Koordinaten sind geschrieben als Zeilenvektor

$$(2.2)\quad \underline{x}' = (x_1, ..., x_I) \qquad \text{und} \qquad \underline{y}' = (y_1, ..., y_J),$$

entsprechend gibt es die Variablen x und y.

Man definiert

$$(2.3)\quad \underline{D}_z = \begin{pmatrix} n_{1.} & & 0 \\ & \ddots & \\ 0 & & n_{I.} \end{pmatrix} \quad \text{und} \quad \underline{D}_s = \begin{pmatrix} n_{.1} & & 0 \\ & \ddots & \\ 0 & & n_{.J} \end{pmatrix}.$$

Man geht davon aus, daß die x_i *standardisiert* vorgegeben sind, daher ist

$$(2.4)\quad \sum_i n_{i.} x_i = 0 \qquad \text{bzw.} \quad \underline{1}_I' \underline{D}_z \underline{x} = 0 \quad \text{und}$$

$$(2.5)\quad \sum_i n_{i.} x_i^2 = n \qquad \text{bzw.} \quad \underline{x}' \underline{D}_z \underline{x} = n.$$

Die y_j sind standardisiert gesucht, also

$$(2.6)\quad \sum_j n_{.j} y_j = 0 \qquad \text{bzw.} \quad \underline{1}_J \underline{D}_s \underline{y} = 0 \quad \text{und}$$

(2.7) $\sum_j n_{.j} y_j^2 = n$ bzw. $\underline{y}' \underline{D}_s \underline{y} = n$

Durch die Standardisierung ist die Korrelation zwischen den Variablen x und y gleich der Kovarianz, also

(2.8) $\text{corr}(x, y) = \frac{1}{n} \sum_i \sum_j n_{ij} x_i y_j = \frac{1}{n} \underline{x}' \underline{N} \underline{y}.$

Die eigentliche Aufgabe kann nun formuliert werden als Bestimmung standardisierter y_j bei gegebenen x_i, so daß die Korrelation zwischen x und y *maximal* wird.

Zur Durchführung dieser Aufgabe sind einige Definitionen hilfreich. Es sei

(2.9) $\underline{\eta} = \frac{1}{\sqrt{n}} \underline{D}_s^{1/2} \underline{y}, \qquad \underline{y} = \sqrt{n}\, \underline{D}_s^{-1/2} \underline{\eta}.$

Damit wird unter Berücksichtigung von (2.7)

(2.10) $\underline{\eta}' \underline{\eta} = \frac{1}{n} \underline{y}' \underline{D}_s \underline{y} = \frac{1}{n} n = 1.$

Folglich handelt es sich bei dem Vektor $\underline{\eta}$ um einen *Einheitsvektor*. Ferner gilt wegen (2.6)

(2.11) $\underline{1}_J' \underline{D}_s^{1/2} \underline{\eta} = 0.$

Die unter (2.8) angegebene Korrelation läßt sich unter Verwendung von (2.9) schreiben als

(2.12) $\text{corr}(x,y) = \frac{1}{\sqrt{n}} \underline{x}' \underline{N} \underline{D}_s^{-1/2} \underline{\eta}.$

Die Koordinaten x_i sind bereits gegeben, ebenso die Matrizen $\underline{N}$ und $\underline{D}_s^{-1/2}$. Also ist die Korrelation durch Wahl des Einheitsvektors $\underline{\eta}$ zu maximieren. Dieses Maximum ist offensichtlich genau dann gegeben, wenn für $\underline{\eta}$ der zu $\underline{x}'\ \underline{N}\ \underline{D}_s^{-1/2}$ parallele Einheitsvektor eingesetzt wird.

Folglich wird

$$(2.13)\quad \underline{\eta} = \frac{\underline{D}_s^{-1/2}\underline{N}'\underline{x}}{\sqrt{\underline{x}'\underline{N}\underline{D}_s^{-1}\underline{N}'\underline{x}}}$$

Nach (2.9) läßt sich nun das maximierende y schreiben als

$$(2.14)\quad \underline{y} = \frac{\sqrt{n}\underline{D}_s^{-1}\underline{N}'\underline{x}}{\sqrt{\underline{x}'\underline{N}\underline{D}_s^{-1}\underline{N}'\underline{x}}}.$$

Nach (2.12) wird das Korrelationsmaximum

$$(2.15)\quad \lambda = \sqrt{\frac{1}{n}\underline{x}'\underline{N}\underline{D}_s^{-1}\underline{N}'\underline{x}} \quad \text{bzw.} \quad \lambda^2 = \frac{1}{n}\underline{x}'\underline{N}\underline{D}_s^{-1}\underline{N}'\underline{x}.$$

Daher läßt sich nun $\underline{y}$ auch ausdrücken durch

$$(2.16)\quad \underline{y} = \frac{1}{\lambda}\underline{D}_s^{-1}\underline{N}'\underline{x}.$$

Zusammenfassung:

Sind Zahlen x_i standardisiert vorgegeben, dann lassen sich nach (1.14) standardisierte y_j bestimmen, so daß corr (x, y) maximal wird. Dieses Maximum ist durch (1.15) gegeben und erfordert keine explizite Berechnung der y_j. Es läßt sich relativ einfach nachweisen, daß die so gewonnenen y_j tatsächlich standardisiert sind.

3. Erweiterung des vorgestellten Verfahrens

Eine Erweiterung des Verfahrens besteht darin, die x_i so vorzugeben, daß das erreichbare Korrelationsmaximum λ wiederum maximal wird. Gesucht sind die x_i, die zugehörigen y_j sowie das entsprechende λ.

Das Korrelationsmaximum wird nach (2.15) bestimmt. Ferner wird definiert

(3.1) $\underline{\xi} = \frac{1}{\sqrt{n}} \underline{D_z}^{1/2} \underline{x}$, also $\underline{x} = \sqrt{n} \underline{D_z}^{-1/2} \underline{\xi}$.

Wegen (2.5) liefert dies

(3.2) $\underline{\xi}' \underline{\xi} = \frac{1}{n} x' \underline{D_z} \underline{x} = 1.$

Daher ist $\underline{\xi}$ ein Einheitsvektor. Mit (3.1) liefert (2.15)

(3.3) $\lambda^2 = \underline{\xi}' \underline{D_z}^{-1/2} \underline{N} \underline{D_s}^{-1} \underline{N}' \underline{D_z}^{-1/2} \underline{\xi}.$

Der rechts stehende Ausdruck ist nun durch Wahl von $\underline{\xi}$ zu maximieren. Durch Definition von

(3.4) $\underline{A} = \underline{D_z}^{-1/2} \underline{N} \underline{D_s}^{-1/2}$ wird

(3.5) $\lambda^2 = \underline{\xi}' \underline{A} \underline{A}' \underline{\xi}.$

Dieser Ausdruck soll durch die Wahl des Einheitsvektors $\underline{\xi}$ maximiert werden. Folglich steht man vor dem Eigenwertproblem

(3.6) $(\underline{A} \underline{A}' - \lambda \underline{I}) \underline{\xi} = 0.$

Die beiden Ausdrücke (3.5) und (3.6) sind gleichbedeutend. Es wird zunächst gesetzt

(3.7) $\underline{\xi}_0 = \frac{1}{\sqrt{n}} \underline{D}_z^{1/2} \underline{1}_l.$

$\underline{\xi}_0$ ist tatsächlich ein Einheitsvektor, da

(3.8) $\underline{\xi}_0' \underline{\xi}_0 = \frac{1}{n} \underline{1}_l' \underline{D}_z \underline{1}_l = \frac{1}{n} n = 1.$

Wird nun $\underline{\xi}_0$ in (3.3) eingesetzt, dann ergibt sich

(3.9) $\lambda^2 = \frac{1}{n} \underline{1}_l' \underline{N} \underline{D}_s^{-1} \underline{N}' \underline{1}_l = \frac{1}{n} n = 1.$

Dies ist der maximale Wert, den ein (quadrierter) Korrelationskoeffizient erreichen kann. Daher ist $\lambda^2 = 1$ der größte Eigenwert in (3.6), der zugehörige Eigenvektor ist $\underline{\xi}_0$. Allerdings ergibt sich diese Lösung *immer* – unabhängig vom Inhalt der Ausgangsdatenmatrix $\underline{N}$. Aus diesem Grund ist dies nicht die geforderte Lösung. Nach (3.1) und (3.6) würde der $\underline{\xi}_0$ entsprechende Vektor $\underline{x}$

(3.10) $\underline{x}_0 = \sqrt{n}\, \underline{D}_z^{-1/2} \underline{\xi}_0 = \underline{1}_l.$

Dieser Vektor erfüllt jedoch die Bedingung (2.4) nicht; er hätte dort $\underline{N} = \underline{0}$ zur Folge!

Ein Lösungsvektor ξ, für den der Vektor $\underline{x}$ die Bedingung (2.4) erfüllt, ist

(3.11) $\underline{1}_l' \underline{D}_z^{1/2} \xi = 0.$

Wegen (3.7) und (2.4) folgt

(3.12) $\underline{\xi}_0' \xi = 0,$

also muß jeder akzeptable Lösungsvektor $\underline{\xi}$ senkrecht auf $\underline{\xi}_0$ stehen. Praktisch bedeutet dies, daß der Lösungsvektor $\underline{\xi}_1$ der Eigenvektor zum größten Eigenwert λ_1^2 von $\underline{AA}'$ mit $\lambda_1^2 < 1$ ist. Der gesuchte Vektor $\underline{x}_1$ der Standardkoordinaten der Zeilen ist wegen (3.1)

(3.13) $\underline{x}_1 = \sqrt{n}\,\underline{D}_z^{-1/2}\underline{\xi}_1.$

Nach (2.14), (3.1) und (3.4) wird

(3.14) $\underline{y}_1 = \frac{\sqrt{n}}{\lambda_1}\underline{D}_s^{-1/2}\underline{A}'\underline{\xi}_1.$

Schließlich lassen sich (3.13) und (3.14) unter Verwendung von (3.4) zusammenfassen, so daß

(3.15) $\underline{y}_1 = \frac{1}{\lambda_1}\underline{D}_s^{-1}\underline{N}'\underline{x}_1$ und analog

(3.16) $\underline{x}_1 = \frac{1}{\lambda_1}\underline{D}_z^{-1}\underline{N}'\underline{y}_1.$

Damit ist die Aufgabe gelöst. Die Inhalte der Zahlentafel lassen sich nun unter Verwendung der Standardkoordinaten x_1 und y_1 im eindimensionalen Raum, d.h. auf einer Geraden, darstellen.

Für die Darstellung der Zeilen und Spalten im zweidimensionalen Raum benötigt man einen weiteren Satz von Standardkoordinaten. Analog zu (3.13) findet man den Vektor $\underline{x}_2$ der Standardkoordinaten, indem man für $\underline{\xi}_2$ den zugehörigen Eigenvektor zum zweitgrößten Eigenwert λ_2^2 von $\underline{AA}'$ einsetzt, wobei wiederum der Eigenwert der Größe Eins unberücksichtigt bleibt. Für eine Darstellung im dreidimensionalen Raum wird man entsprechend den Eigenvektor zum drittgrößten Eigenwert nehmen, usw.

Werden alle r zu $\lambda_i > 0$ gehörigen Eigenvektoren auf diese Weise verwendet, spricht man von einer kompletten Lösung, werden nur $k < r$ Eigenvektoren verwendet, von

einer verkürzten Lösung. Die Güte einer verkürzten Lösung gegenüber der kompletten wird angegeben durch

$$(3.17)\ \Gamma = \frac{\lambda_1^2 + \ldots + \lambda_k^2}{\lambda_1^2 + \ldots + \lambda_r^2}.$$

Ergibt sich für Γ ein ubefriedigend kleiner Wert, so ist die Analyse praktisch mißglückt.

4. Zusammenfassung

(a) Bestimmung der Zeilensummen, der Spaltensummen und der Gesamtsumme der in $\underline{N}$ enthaltenen Objekte

(b) Bildung der Diagonalmatrizen $\underline{D}_z$ und $\underline{D}_s$

(c) Bildung der Matrix $\underline{A} = \underline{D}_z^{-1/2}\underline{N}\underline{D}_s^{-1/2}$

(d) Eigenwertzerlegung der Matrix $\underline{A}\underline{A}'$

(e) Ermittlung des zum größten Eigenwert $\lambda_1 < 1$ gehörigen Eigenvektors $\underline{\xi}_1$

(f) Bildung der gesuchten Koordinaten $\underline{x}_1 = \sqrt{n}\ \underline{D}_z^{-1/2}\underline{\xi}_1$

(g) Bildung der gesuchten Koordinaten $\underline{y}_1 = \frac{1}{\lambda_1}\underline{D}_s^{-1}\underline{N}'\underline{x}_1$

(h) Ggf. Wiederholung der Schritte (e) – (g) für weitere k Koordinaten, wobei in (e) der zum k.-größten Eigenwert λ_k gehörige Eigenvektors $\underline{\xi}_k$ zu wählen ist

(i) Bestimmung der Güte der Anpassung $\Gamma = \frac{\lambda_1^2 + \ldots + \lambda_k^2}{\lambda_1^2 + \ldots + \lambda_r^2}$

5. Beispiel

Gegeben ist folgende Ausgangsdaten („Farbpyramidentest“, fiktives Beispiel nach Lienert):

	I – ausgewogen	**II – sensitiv**	**III – impulsiv**	**IV - beherrscht**
gelb	48	18	10	8
grün	21	31	10	9
rot	6	8	14	6
blau	14	9	11	27

N =

48	18	10	8
21	31	10	9
6	8	14	6
14	9	11	27

Bestimmung von $\underline{D}_z$ und $\underline{D}_s$:

Dz =

84	0	0	0
0	71	0	0
0	0	34	0
0	0	0	61

Ds =

89	0	0	0
0	66	0	0
0	0	45	0
0	0	0	50

Berechnung von $\underline{A} = \underline{D}_z^{-1/2}\underline{N}\underline{D}_s^{-1/2}$ nach (3.4)

A =

0.5551	0.2417	0.1627	0.1234
0.2642	0.4529	0.1769	0.1511
0.1091	0.1689	0.3579	0.1455
0.1900	0.1418	0.2100	0.4889

Bildung von $\underline{A}\underline{A}'$

A*A´ =

0.4083	0.3036	0.1776	0.2343
0.3036	0.3290	0.1906	0.2254
0.1776	0.1906	0.1897	0.1910
0.2343	0.2254	0.1910	0.3393

Eigenwertzerlegung von $\underline{AA}'$

```
E =

    0.3074   -0.4834   -0.5796    0.5797
   -0.6019    0.5442   -0.2400    0.5329
    0.6992    0.5398    0.2894    0.3688
   -0.2333   -0.4229    0.7230    0.4940

L =

    0.0400         0         0         0
         0    0.0732         0         0
         0         0    0.1531         0
         0         0         0    1.0000
```

Bestimmung von $\underline{x}_1$ und $\underline{x}_2$ nach (3.13)

```
x1 =

   -0.9998
   -0.4504
    0.7848
    1.4637

x2 =

   -0.8339
    1.0211
    1.4638
   -0.8561
```

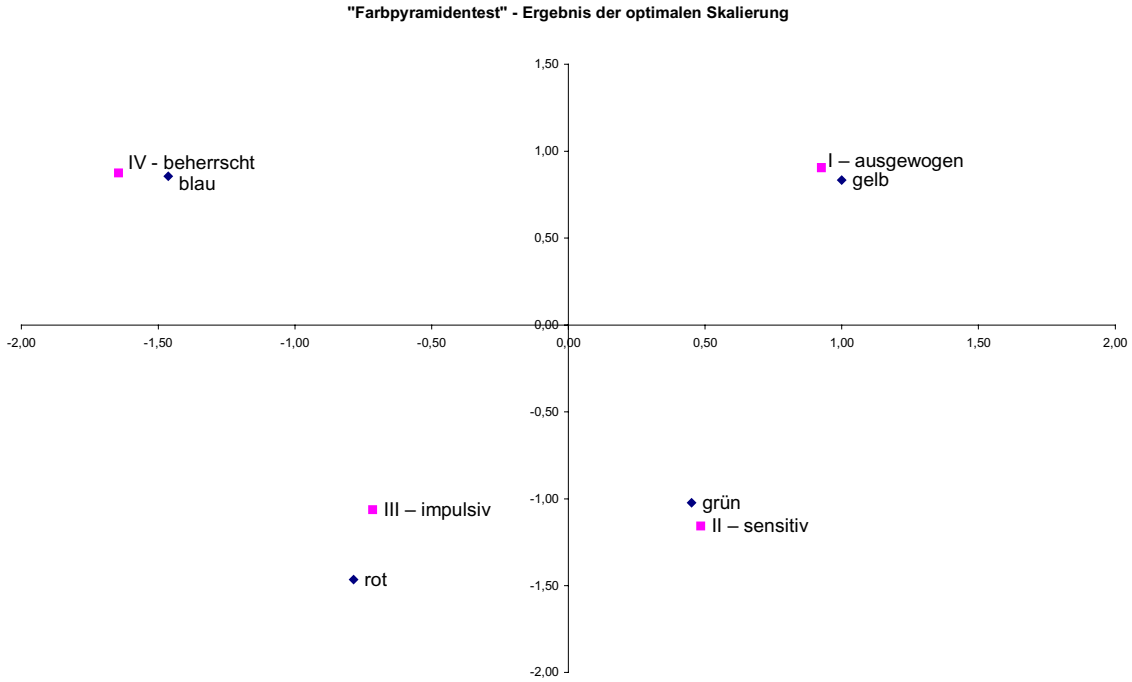
Bestimmung von $\underline{y}_1$ und $\underline{y}_2$ nach (3.15)

```
y1 =

   -0.9260
   -0.4843
    0.7147
    1.6445

y2 =

   -0.9045
    1.1561
    1.0632
   -0.8730
```

Berechnung der Güte der Anpassung (Gesamt, 1. Achse, 2. Achse)

```
G =

    0.8500

G1 =

    0.5750

G2 =

    0.2750
```

Graphische Darstellung



Hinweis: Die Berechnung der Koordinaten und die graphische Ausgabe wurden mit unterschiedlichen Programmen durchgeführt. Aus diesem Grund kommt es zu einer „Klappung“ der Achsen in der graphischen Ausgabe gegenüber der Berechnung der Koordinaten.