Korrespondenzanalyse

Ziel der Korrespondenzanalyse

Zeilen und Spalten einer rechteckigen Matrix so analysieren, dass sie als Punkte in einem gering dimensionierten Raum (zwei oder drei Dimensionen) dargestellt werden können.

Anforderungen an die Daten (Stärke des Verfahrens)

Homogene Variablen müssen lediglich > 0 und in einer rechteckigen Matrix angeordnet sein. Es können **nicht-metrische** Daten verwendet werden. Keine Anforderungen an Stichprobenumfang und keine Verteilungsannahmen.

Einordnung in die multivariaten Verfahren

Exploratives, interdependentes Verfahren, mathematisches Verfahren „Singular Value Decomposition“

Normierung der Daten

Um die Daten der Matrix vergleichbar machen zu können, müssen diese normiert werden. Dazu werden relative Häufigkeiten in den Zeilen und Spalten gebildet.

Masse der Zeilen (Summe der Nennungen der Zeile geteilt durch Summe der Gesamtnennungen) sagt aus, wie sehr die angebotenen Merkmale auf das Objekt zutreffen. [$W_i = n_i / n$]

Masse der Spalten (analoge Berechnung) sagt aus, wie häufig ein Merkmal auf alle untersuchten Objekte zutrifft.

Festlegung des geometrischen Schwerpunktes

- Messung der Streuungsinformation, indem die Abweichungen der Marken- bzw. Merkmalsprofile von ihrem geometrischen Schwerpunkt berechnet werden.
- Dieser Schwerpunkt stellt das Durchschnittsprofil der Punktwolke dar und wird auch als Zentroid bezeichnet.

- Der Datenvektor des Schwerpunktes *q* lässt sich berechnen, indem man die I Zeilenprofile r_i mit ihrer Masse w_i gewichtet und anschließend aufsummiert:

  Formel: $q = \Sigma_i \; w_i r_i = w_1r_1 + w_2r_2 + w_3r_3 + w_4r_4 + w_5r_5$

- Die getrennte reihen- und spaltenweise Normierung der Matrix *N* führt dazu, dass die Koordinaten des geometrischen Schwerpunktes der Reihenprofile *q* exakt den Massen der Spaltenprofile q_j entsprechen.
- Es gilt auch der Umkehrschluss: Der Schwerpunkt der Spaltenprofile w stimmt mit den Massen w_i der Reihenprofile r_i überein.
- Abb. zeigt den Zentroiden q als geometrischen Schwerpunkt der Punktwolke aus konkurrierenden Biersorten.
- Der in der Gleichung skizzierte Algorithmus, mit dem der Zentroid der Reihenprofile berechnet wird, bewirkt, dass q in der räumlichen Darstellung tendenziell näher an Punkten mit hohen Massen (z.B. Biersorte 4 mit w4 = 0,324) als an Punkten mit niedrigen Massen (z.B. Biersorte 1 mit w1 = 0,045) liegt.

Berechnung der Distanzen zwischen den Punkten

- Abstände der Punkte zueinander bzw. zum Schwerpunkt geben Aufschluss über die Ähnlichkeiten bzw. Unterschiede ihrer Profile
- Punkte nah beieinander: Ähnlichkeit
- Punkte weiter entfernt: Unähnlichkeit
- Nah am Zentroid: Alternativen die dem Durchschnittsprofil ähneln
- Berechnung der Entfernung mit euklidischem Distanzmaß hier nicht günstig:
  - berücksichtigt nicht, dass die Dimensionen in Abhängigkeit von den Spaltenhäufigkeiten unterschiedliche Skalengrößen aufweisen
  - Entfernungen zwischen den Produktpunkten würden durch Merkmale, die den Alternativen insgesamt häufiger zugeordnet werden, sehr viel stärker beeinflusst als durch Merkmale die nur geringe Häufigkeiten aufweisen.
- Berechnung mit gewichteter euklidischer Distanz:
  - zusätzliche Gewichtung der Distanz mit dem inversen Element des Durchschnittsprofils der Spalte j
- Formel:

$$d^2_{(i,k)} = \sum_j \frac{(n_{ij}/n_i - n_{kj}/n_k)^2}{n_j/n} = \frac{(r_{ij} - r_{kj})^2}{q_j}$$

- Gewichtete euklidische Distanz ist proportional zur Chi²-Statistik, deshalb auch Chi²-Distanz genannt.
- Vorteil:
  Interpretation als euklidische Distanz. Durch die Gewichtung der Punktekoordinaten mit dem Faktor $1/q_j$ entspricht sie der Chi²-Distanz, so dass man die Verzerrungseffekte, die durch unterschiedliche Spaltenhäufigkeiten hervorgerufen werden, eliminiert.

Berechnung der Gesamtstreuung in den Daten

- Berechnung der räumlichen Abstände eines Punktes vom geometrischen Schwerpunkt mit Hilfe der Chi² - Distanz:

$$d^2_{(i,q)} = \sum_j \frac{(n_{ij}/n_i - n_j/n)^2}{n_j/n} = \frac{(r_{ij} - q_j)^2}{q_j}$$

- Berechnung der Gesamtstreuung in (I) „total inertia“ um zu verdeutlichen, wie stark die einzelnen gewichteten Markenprofile um den Schwerpunkt streuen.

$$\sum_i w_i d^2_i = \sum_{ij} w_i \frac{(r_{ij} - q_j)}{q_j} = Chi^2/n = in(I)$$

→ Die in den Daten enthaltene Gesamtvariation in (i) ist direkt proportional zur

Chi²- Statistik, die prüft, ob die beobachteten Häufigkeiten statistisch signifikant von den erwarteten Häufigkeiten abweichen.

Dimensionsreduktion

- Ziel: Bestimmung der optimalen Unterräume bei möglichst geringem Informationsverlust

  Die graphische Interpretation einer Punktwolke ist nur in einem 2-3- dimensionalen Raum möglich.

- Vorgehensweise:
  - Bestimmung des Durchschnitts über die kürzeste Entfernung aller Punkte der betrachteten mehrdimensionalen Punktwolke zum Unterraum. (Methode: „singular value decomposition“ (SVD))
  - auch hier werden die Dimensionen gewichtet
  - Berechnung der Distanz zwischen r_i und $\underline{r}_j$ im zweidimensionalen Raum:

    $d^2_i = \Sigma (r_i - \underline{r}_i)^2 / q_j$

    r_i = beliebiger Profilpunkt
    $\underline{r}_i$ = Punkt im Unterraum mit kürzester Entfernung zu r_i
    q_j= Masse der Spaltenprofile
- Die Unterräume werden so bestimmt, dass der optimale eindimensionale Unterraum, d.h. die erste Achse, den größten Beitrag zur Erklärung der Gesamtstreuung leistet, die zweite Achse den zweitgrößten usw.

- Die Dimensionen der Unterräume werden als „principal axes“ (Hauptachsen), der Anteil an der Gesamtstreuung, den eine Dimension erklärt, als „principal inertia“ bezeichnet.
- Der Zentroid ist Mittelpunkt bzw. Ursprung aller Unterräume Optimierungsproblem: Je nachdem wie der Unterraum gewählt wird variiert dieser.

Geometrie der Spaltenprofile

Bisherige Überlegungen:

- Untersuchung der Zeilen- bzw. Markenprofile
- Die Analyseschritte dienen aber auch zur graphischen Darstellung der Spaltenprofile
- Die Merkmalsprofile werden in einem *fünfdimensionalen* Raum positioniert, den die *fünf Biersorten* aufspannen.

Symmetrie:

- Die Koordinaten des Durchschnittsprofils der Merkmale entsprechen exakt den Massen der Reihenprofile der fünf Biere.

  $W = \sum_j q_j c_j = q_1c_1 + q_2c_2 + q_3c_3$

- Äußert sich darin, dass die Koordinaten des geometrischen Schwerpunktes der Spaltenprofile exakt den Massen der Reihenprofile entsprechen und umgekehrt.

- *Berechnung der Distanz des Spaltenpunktes zum Schwerpunkt*:

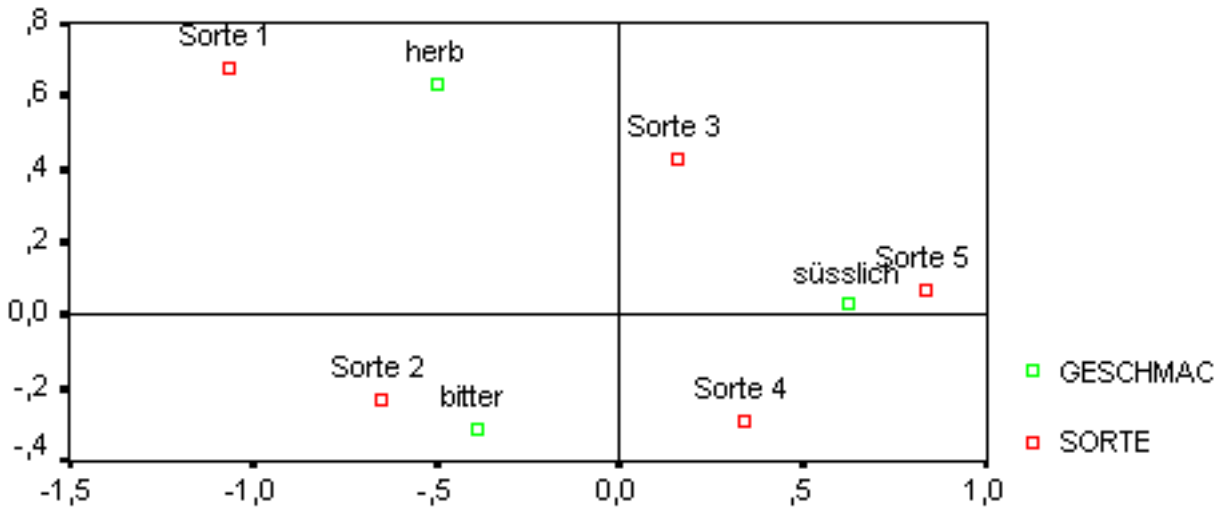
  $d^2 (_{j,q}) = \sum_i (c_{ij} - w_i)^2 / w_i$

- Die Gesamtstreuung in den Daten und die Erklärungsbeiträge der Hauptachsen für die Analyse der Reihen- und Spaltenprofile identisch sind!

  $(\ln(I) = \sum_i w_i d^2_i = \sum_j q_j d^2_j)$

Simultane graphische Darstellung der Zeilen- und Spaltenprofile

- „singular values“ (singuläre Werte) =„principal interia“ (Anteil der Gesamtstreuung, den eine Dimension erklärt).



Interpretation:
Die räumliche Position einer Biermarke liegt tendenziell immer dann in der Nähe eines Merkmals, wenn die Alternative bezüglich dieses Merkmals im Vergleich zu anderen Merkmalen eine hohe bedingte relative Häufigkeit aufweist.
→ Es werden nicht nur Ähnlichkeiten und Unterschiede innerhalb einer Punktwolke, sondern auch die Beziehungen zwischen den beiden Wolken visualisiert.
Anmerkungen:

Wichtig!!!
Die räumliche Entfernung zwischen einem Punkt und einem Merkmal darf **nicht direkt** interpretiert werden, da eine Distanz zwischen Punkten unterschiedlicher Wolken nicht explizit definiert ist. Die Distanzen zwischen den Punkten einer Wolke können mit Hilfe der Chi²- Distanz ermittelt werden.

Ablauf und Auswertung in SPSS

- /Analysieren/Dimensionsreduktion/Korrespondenzanalyse
- Zeile → Sorte (Bereich definieren: Anzahl der Sorten)
- Spalte → Merkmale (Bereich definieren: Anzahl der Merkmale)
- Modelle → Distanzmaß: Chi² → Normailisierungsmethode: symmetrisch
- Statistik → Korrespondenztabelle → Übersicht Zeilenpunkte und Spaltenpunkte → Zeilen- und Spaltenprofile
- Diagramme → Biplot → Zeilenpunkte und Spaltenpunkte

Auswertung

Dimension	Singulärwert	Auswertung für Trägheit	Chi²	Sig.	Anteil der Trägheit		Singulärwert für Konfidenz	
					Bedingen	Kumuliert	Standardabweichung	Korrelation 2
1	,265	,070			,845	,845	,051	-,002
2	,114	,013			,155	1,000	,059	
Auswertung		,083	25,977	,001[a]	1,000	1,000		

a. 8 Freiheitsgrade

Dimensionen

Singulärwert

Auswertung der Trägheit

Chi-Quadrat

Signifikanz

Anteil der Trägheit

Standardabweichung