LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Kurs *SPSS*
Somersemester 2009

Kapitel
Clusteranalyse und Faktorenanalyse
Zwei multivariate Verfahren

Susanne Heim und Christian Heumann

Inhalt dieses Abschnitts

Zwei multivariate Verfahren

Ideen

- Clusteranalyse
- Faktorenanalyse

Motivation

Beispiel

- Es liegen die Einzelergebnisse aller 10 Disziplinen von verschiedenen Zehnkämpfern vor.
- Eine Fragestellung könnte lauten: gibt es verschiedene Typen von Zehnkämpfern, die einander bezüglich der Ergebnisse sehr ähnlich sind? In diesem Fall wollen wir Gruppen ähnlicher Zehnkämpfer finden. Erfordert eine Operationalisierung von „Ähnlichkeit“ bzw. „Verschiedenheit“ von Objekten. Dies ist die typische Fragestellung der Clusteranalyse. Man spricht auch von *Klassifikation* („unsupervised“ im Gegensatz zur Diskriminanzanalyse, die „supervised“ ist).
- Eine weitere Fragestellung könnte lauten: gibt es latente (das heißt, nicht beobachtete) Faktoren, die die einzelnen Merkmale (10 Disziplinen) und deren Ausprägungen im Wesentlichen bedingen. Beispiele wären Schnelligkeit, Kraft, Ausdauer. Ziel ist also, die 10 Merkmale auf wenige Faktoren zu reduzieren, die hinreichend die Ergebnisse erklären können. Dies ist die typische Fragestellung der Faktorenanalyse. Man spricht auch von *Dimensionsreduzierung*.

Motivation

Abgrenzung Clusteranalyse und Faktorenanalyse

- Clusteranlayse klassifiziert Daten entlang der Objekte (Fälle, hier: Zehnkämpfer), Faktorenanalyse komprimiert entlang der Merkmale (hier: die 10 Einzeldisziplinen).
- Allerdings: Clustering auf Merkmalen ebenfalls möglich (und Bi-Clustering), es wird aber — im Gegensatz zur Faktorenanalyse — keine Reduktion auf weniger Variablen vorgenommen.
- Beide: explorative Verfahren (Ausnahme: konfirmatorische Faktorenanalyse, SPSS AMOS).

Clusteranalyse

Überblick

- Es gibt sehr viel verschiedene Verfahren. Es gibt keine „beste Methode“. Zum Beispiel gibt es agglomerative (d.h. jedes Objekt ist zunächst ein Cluster, sukzessive werden Cluster zusammengefasst) und divisive Verfahren (zunächst bilden alle Objekte zusammen einen Cluster, sukzessive wird in Teilmengen partitioniert), modellbasierte Verfahren (z.B. multivariate Normalverteilung als Annahme), etc.
- Wie werden die Abstände zwischen den Clustern definiert? Auch hier gibt es verschiedende Methoden.
- Wie werden die Abstände zwischen den Objekten definiert? Auch hier gibt es, je nach Skalenniveau der Merkmale, verschiedene Möglichkeiten.
- D.h. ein Clusterverfahren wird durch zahlreiche „Parameter“ gesteuert.
- SPSS bietet neuerdings ein innovatives Clusterverfahren an, das einige Probleme mit den älteren Verfahren überwinden soll (bedarf aber wohl noch der genauen Evaluation).

Clusteranalyse

Abstände zwischen Clustern

- Average Linkage (SPSS: Linkage zwischen den Gruppen). Berechnet durchschnittliche Distanz aller Elemente eines Clusters mit allen Elementen eines anderen Clusters
- Single Linkage (SPSS: nearest neigbor, nächstgelegener Nachbar). Distanz entspricht der geringsten Distanz zweier Objekte in verschiedenen Clustern.
- Complete Linkage (SPSS: furthest neigbor, entferntester Nachbar). Distanz entspricht der größten Distanz zweier Objekte in verschiedenen Clustern.
- Zentroid–Verfahren (Distanz der Clustermittelwerte). Cluster, deren euklidische Distanz der Mittelwerte am geringsten ist, werden zusammengefasst. Nur sinnvoll, wenn alle Merkmale metrisch sind.
- Methode nach Ward (beruht auf der Homogenität innerhalb der Cluster/Partitionen). Nur sinnvoll, wenn alle Merkmale metrisch sind.

Clusteranalyse

Abstände/Distanzen oder Ähnlichkeit zwischen Objekten

- Abhängig vom Skalenniveau der Merkmale
- Euklidische Distanz in der Regel nur sinnvoll, wenn alle Merkmale intervallskaliert oder alle Merkmale binär sind, nicht aber, wenn Mischung aus Skalenniveaus (welche Distanzmaße bei nominal, ordinal?)
- Intervallskaliert: Vergleichbarkeit der Skalen muss hergestellt werden (z.B. Standardisierung, SPSS: Z-Werte)
- Beispiel wenn alle p Merkmale binär: Ähnlichkeit ist dann die Anzahl der Merkmale, bei denen beide Objekte übereinstimmen (Übereinstimmung, wenn beide eine 1 bei einem bestimmten Merkmal haben oder eine 0) geteilt durch Anzahl der Merkmale (p). Alternativ: Übereinstimmung nur, wenn Eigenschaft vorhanden, also wenn beide eine 1 im betrachteten Merkmal haben (Jaccard–Koeffizient).
- Wahl des Maßes hängt von der Anwendung ab.

Clusteranalyse

SPSS

- Hierarchische Clusteranalyse (agglomeratives Verfahren)
- Two-Step Cluster (relativ neu), schnell für große Stichproben.
- K-Means Cluster (Quick Cluster), Variablen müssen vorher standardisiert werden, schnell für große Stichproben.

Faktorenanalyse

Überblick

- Motivation 1: reine Dimensionsreduktion.
- Motivation 2: Aufspüren weniger, wichtiger Faktoren, die auch inhaltlich interpretiert werden können.

Faktorenanalyse

Modell

$$\mathbf{y} - \mu = \mathbf{L}\mathbf{f} + \mathbf{e}$$

- $\mathbf{y}$: p beobachtete Variablen/Merkmale mit Erwartungswert μ.
- $\mathbf{f}$: k *gemeinsame* Faktorenvariablen ($k < p$).
- $\mathbf{L}$: Matrix der Faktorenladungen.
- $\mathbf{e}$: Spezifische Einzelfaktoren (die nur beim jeweiligen Merkmal auftreten) und Meßfehler
- Annahmen:

$$\begin{aligned} E(\mathbf{f}) &= 0 \\ E(\mathbf{e}) &= 0 \qquad cov(\mathbf{e}) = \mathbf{V} = diag(v_1^2, \ldots, v_p^2) \\ cov(\mathbf{f}, \mathbf{e}) &= 0 \end{aligned}$$

- Weitere Anfangsannahme: $cov(\mathbf{f}) = \mathbf{I}$ (Einheitsmatrix)

Faktorenanalyse

Verfahren

- Hauptkomponenten–Analyse (PCA), wenn Motivation 1 oder wenn andere Verfahren schief gehen („PCA geht immer“)
- Maximum–Likelihood (weniger robust), wenn Motivation 2.
- Hauptachsen–Faktorenanalyse (Hauptfaktorenanalyse), wenn Motivation 2.
- Weitere Verfahren (KQ, verallgemeinerte KQ, Image/Anti–Image, etc.).

Identifizierbarkeitsprobleme

- Rotationsproblem. Matrix der Faktorladungen, **L**, ist, grob gesagt, nicht eindeutig. Oft verwendet man die *Varimax*–Rotation, die versucht, Faktoren herzustellen, die in einigen Variablen hoch, in den anderen Variablen niedrig „laden“. Das verbessert die Interpretierbarkeit.
- Kommunalitätenproblem und Kommunalitätenschätzproblem. Kommunalitäten: Anteile der gemeinsamen Faktoren an der Varianz der beobachteten Merkmale. Diese sind im Wesentlichen quadrierte Korrelationskoeffizienten, die daher immmer zwischen 0 und 1 liegen müssen. Treten während des Verfahrens Schätzungen größer oder gleich 1 auf, spricht man von einem Heywood case. Größer 1 würde auch bedeuten, dass ein spezifischer Einzelfaktor bzw. der Messfehler eine Varianz kleiner als 0 hat, was natürlich Unsinn ist.

Ist sie für gegebene Daten sinnvoll?

- Beurteilung oft durch sog. Kaiser–Meyer–Olkin Maß. Sollte größer 0.5 sein (Faustregel)

Anzahl der Faktoren

- Meist wird auf die Eigenwerte geachtet. Anzahl der Faktoren wird oft gleich der Anzahl der Eigenwerte gesetzt, die größer als 1 sind.
- Grafische Möglichkeit: Scree–Plot