**Austrian Journal of Statistics**
**Volume 27 (1998), Number 3, 127-138**



# Technik und Weiterentwicklung der multiplen Korrespondenzanalyse

Bernhard Böhm
Institut für Statistik, Universität Innsbruck

**Zusammenfassung:** Nach der Darstellung der Technik der multiplen Korrespondenzanalyse wird gezeigt, daß die multiple Korrespondenzanalyse insbesondere bei der Analyse von qualitativen Merkmalen mit vielen Ausprägungen geringe prozentuale Erklärungsbeiträge der Faktorenachsen und damit geringe globale Erklärungsgüten niedrigdimensionaler Faktorenräume liefert. Die zur Erhöhung der relativen Erklärungsbeiträge der Faktorenachsen von Greenacre (1993) vorgeschlagene Methode der Achsen-Reskalierung wird dargestellt und numerisch beurteilt. Greenacres Joint Correspondence Analysis motiviert im Anschluß die Weiterentwicklung der Korrespondenzanalyse über direkte Kommunalitätenschätzer.

**Abstract:** After the presentation of Multiple Correspondence Analysis it is shown, that Multiple Correspondence Analysis has the disadvantage of low contributions to the total inertia along its principal axes, if qualitative variables with many values are analyzed. To improve the contributions of the principal axes, Greenacre (1993) develops the rescaling of the principal axes of a Multiple Correspondence Analysis. This technique is explained and tested numerically. In respect of Greenacre's Joint Correspondence Analysis, the author develops an approach to improve Multiple Correspondence Analysis by using direct estimators for the communalities of the variables.

**Schlüsselwörter**: Binäre und multiple Korrespondenzanalyse, Achsen-Reskalierung, Joint Correspondence Analysis, Kommunalität.

## 1 Multiple Korrespondenzanalyse

Die von Benzécri et al. (1973) entwickelte faktorenanalytische Technik der binären Korrespondenzanalyse dient der niedrigdimensionalen graphischen Darstellung der Zeilen- und Spaltenprofile einer Kontingenztafel **K** und führt auf die kanonische Analyse zweier qualitativer Merkmale. Dies erlaubt die direkte Verbindung der Korrespondenzanalyse mit der verallgemeinerten kanonischen Analyse nach Carroll (siehe Carroll, 1968, und Böhm, 1997). Im Unterschied zur verallgemeinerten kanonischen Analyse, die im Rahmen ihres Modells unmittelbar die Betrachtung beliebig vieler qualitativer Merkmale ermöglicht, schlägt Benzécri (1977) zur Analyse von mehr als zwei Merkmalen lediglich die direkte Anwendung der Technik der binären

Korrespondenzanalyse auf beliebig viele qualitative Merkmale vor. Benzécris Vorschlag zur multiplen Korrespondenzanalyse findet insbesondere durch Greenacre (1984) allgemeine Anerkennung in der englischsprachigen Literatur.

## 1.1 Multiple Korrespondenzanalyse der Indikatormatrix Z

Bei der Analyse von Objekten, die an Hand von mehr als zwei qualitativen Merkmalen $m_j$ beschrieben werden, erfolgt üblicherweise zunächst die Binärkodierung der Daten in Form einer disjunktiven, vollständigen Indikatormatrix $\mathbf{Z} = (\mathbf{Z}_1 | \mathbf{Z}_2 | ... | \mathbf{Z}_p)$.

Weist das Objekt $o_i$ (i = 1, ..., n) bezüglich des Merkmals $m_j$ (j = 1, ..., p) die Ausprägung k der Ausprägungsmenge $X_j$ auf, so wird an entsprechender Stelle $z_{ik}$ der Teilmatrix $\mathbf{Z}_j$ von $\mathbf{Z}$ eine 1 eingetragen.

Somit gilt für $\mathbf{Z}_j$ (j = 1, ..., p): $z_{ik} = 1$, falls $o_i$ Ausprägung $k \in X_j$ aufweist; $z_{ik} = 0$, sonst. Für die Anzahl $r_j$ der Spalten von $\mathbf{Z}_j$ gilt $r_j = |X_j|$. Dabei ist $|X_j|$ die Anzahl der Ausprägungen von $m_j$.

Wird auf die p-variate Indikatormatrix $\mathbf{Z} = (\mathbf{Z}_1 | \mathbf{Z}_2 | ... | \mathbf{Z}_p)$ mit $r = \sum_{j=1}^{p} r_j$ Spalten und n Zeilen unmittelbar die binäre Korrespondenzanalyse angewendet, so ergibt sich in der Notation der binären Korrespondenzanalyse die Gesamtsumme $f^Z = \sum_{i=1}^{n}\sum_{k=1}^{r} z_{ik} = p \cdot n$ der Indikatormatrix $\mathbf{Z}$. Dabei wird $\mathbf{Z}$ als Matrix der absoluten Häufigkeiten 0 oder 1 interpretiert. Es folgen relative Häufigkeiten $f_{ik}^Z = \frac{z_{ik}}{f^Z} = \frac{z_{ik}}{n \cdot p}$.

Wegen $\sum_{k=1}^{r} z_{ik} = p$ (i = 1, ..., n) haben alle Objekte $o_i$ dieselbe Masse $f_{i.}^Z = \sum_{k=1}^{r} f_{ik}^Z = \frac{1}{n}$.

Die Massen bzw. relativen Randhäufigkeiten der Spalten von $\mathbf{Z}$ beschreiben die Verteilung der Ausprägungen der Merkmale $m_1$ bis $m_p$ mit $f_{.k}^Z = \sum_{i=1}^{n} f_{ik}^Z = \frac{1}{n \cdot p}\sum_{i=1}^{n} z_{ik}$.

Für jede Teilmatrix $\mathbf{Z}_j$ von $\mathbf{Z}$ gilt $\sum_{i=1}^{n}\sum_{k=1}^{r_j} z_{ik} = n$. Für die Summe der relativen Randhäufigkeiten der Spalten von $\mathbf{Z}_j$ und die Gewichtung jedes Merkmals $m_j$ folgt deswegen $\sum_{k=1}^{r_j} f_{.k}^Z = \frac{1}{p}$.

Die Gesamtvarianz der p-variaten Indikatormatrix $\mathbf{Z}$ hängt mit

$$\Phi^2(\mathbf{Z}) = \frac{r - p}{p} \qquad (1)$$

lediglich von der Anzahl p der Merkmale $m_j$ und der Anzahl r sämtlicher Merkmalsausprägungen ab. Die Varianz jeder Teilmatrix $\mathbf{Z}_j$ von $\mathbf{Z}$ ist $\Phi^2(\mathbf{Z}_j) = \frac{r_j - 1}{p}$ und steigt linear mit der Anzahl $r_j$ der Ausprägungen des Merkmals $m_j$ (siehe Greenacre, 1984).

Bildet man die Matrix $\mathbf{X}$ mit

$$x_{ik} = \frac{f_{ik}^{Z} - f_{i.}^{Z} f_{.k}^{Z}}{\sqrt{f_{i.}^{Z} f_{.j}^{Z}}},$$

so führt die Korrespondenzanalyse von $\mathbf{Z}$ auf die Diagonalisierung der Matrix $\mathbf{T} = \mathbf{X}^T\mathbf{X}$ und liefert $\rho = r - p$ Faktoren $\mathbf{u}_\alpha \in \Re^r$ und $\mathbf{v}_\alpha \in \Re^n$. Es gilt grundsätzlich $\lambda_\alpha \leq 1$ (siehe Jambu, 1992). Mit steigender Anzahl der Ausprägungen je Merkmal nimmt die Anzahl $\rho$ der Eigenwerte $\lambda_1$ bis $\lambda_\rho$ der Analyse von $\mathbf{Z}$ zu und die relativen Erklärungsbeiträge $\frac{p \cdot \lambda_\alpha}{r - p}$ der Faktorenachsen stark ab (siehe Tenenhaus und Young, 1985).

Für die Profile der Zeilen und Spalten von $\mathbf{Z}$ wird wie bei der Korrespondenzanalyse von $\mathbf{K}$ der $\chi^2$-Abstand definiert. Aufgrund des Prinzips der Verteilungsäquivalenz kann die im allgemeinen recht umfangreiche Indikatormatrix $\mathbf{Z}$ zu Beginn der Analyse komprimiert werden: zwei identische Zeilenvektoren $\mathbf{z}_{i1}$ und $\mathbf{z}_{i2}$ von $\mathbf{Z}$ werden durch einen einzigen Zeilenvektor $\mathbf{z}_i = \mathbf{z}_{i1} + \mathbf{z}_{i2}$ zusammengefaßt.

## 1.2 Analyse der Burt-Matrix $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$

Die binäre Korrespondenzanalyse kann auch auf die multiple Kontingenztafel oder Burt-Matrix $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$ (siehe Burt, 1950) angewandt werden.

$\mathbf{B}$ hat die symmetrische Blockstruktur

$$\mathbf{B} = \mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1^T\mathbf{Z}_1 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_p \\ \vdots & \ddots & \vdots \\ \mathbf{Z}_p^T\mathbf{Z}_1 & \cdots & \mathbf{Z}_p^T\mathbf{Z}_p \end{pmatrix}$$

und umfaßt die zweidimensionalen Kontingenztafeln $\mathbf{Z}_i^T\mathbf{Z}_j$ ($i = 1, ..., p$; $j = 1, ..., p$; $i \neq j$) sowie die Diagonalmatrizen $\mathbf{Z}_j^T\mathbf{Z}_j$ ($j = 1, ..., p$) entlang der Hauptdiagonalen.

Jede Diagonalmatrix $\mathbf{Z}_j^T\mathbf{Z}_j$ enthält als Spaltensummen der Matrix $\mathbf{Z}_j$ die Anzahl der einzelnen Ausprägungen des Merkmals $m_j$.

Da $\mathbf{B}$ positiv semidefinit und symmetrisch ist, liefert die multiple Korrespondenzanalyse von $\mathbf{B}$ zwei identische Koordinatensätze für Zeilen- und Spaltenprofile.

Im Unterschied zur Analyse der Indikatormatrix $\mathbf{Z}$ können die Beitrage einzelner Objekte zur Gesamtvarianz der Punktwolke im Rahmen der Analyse von $\mathbf{B}$ nicht mehr zurückverfolgt werden. Um die Objekte $o_i$ einer Objektmenge $O$ trotzdem im Faktorenraum der Achsen $\mathbf{v}_\alpha^B \in \Re^r$ der Analyse von $\mathbf{B}$ zu repräsentieren, können die Objekte nachträglich in Form der Zeilenvektoren $\mathbf{z}_i$ von $\mathbf{Z}$ in den Faktorenraum projiziert werden.

Die Faktoren $\mathbf{v}_\alpha{}^Z$ und $\mathbf{v}_\alpha{}^B$ der Analyse von **Z** und **B** sind identisch. Der Unterschied beider Analysen liegt jedoch in den Erklärungsbeiträgen der einzelnen Faktorenachsen. Zwischen den Eigenwerten $\lambda_\alpha{}^Z$ und $\lambda_\alpha{}^B$ der Analyse von **Z** bzw. **B** besteht der Zusammenhang $\lambda_\alpha{}^B = (\lambda_\alpha{}^Z)^2$.

Abgesehen von Maßstabsänderungen entlang der Faktorenachsen sind somit die Analysen von **Z** und **B** praktisch äquivalent (siehe Greenacre, 1984).

Im Rahmen der Analyse von **B** ist die zugrunde gelegte Gesamtvarianz mit

$$\Phi^2(\mathbf{B}) = \frac{1}{p^2}\left(\sum_i \sum_{j \neq i} \Phi^2(\mathbf{Z}_i{}^T \mathbf{Z}_j) + (r - p)\right) \qquad (2)$$

geringer als die Varianz $\Phi^2(\mathbf{Z})$ von **Z** (siehe Gleichung 1).

Wegen der Reduktion der Gesamtvarianz $\Phi^2(\mathbf{Z})$ von **Z** auf $\Phi^2(\mathbf{B})$ von **B** liefert die Korrespondenzanalyse der Burt-Matrix **B** höhere prozentuale Varianzanteile der Faktorenachsen als die Analyse von **Z**. Betrachtet man die $\chi^2$-Unabhängigkeitstest-Prüfgrößen

$$\chi^2(\mathbf{Z}_i{}^T \mathbf{Z}_j) = f \sum_{s=1}^{r_i} f_{s.} \sum_{t=1}^{r_j} \frac{\left(\frac{f_{st}}{f_{s.}} - f_{.t}\right)^2}{f_{.t}}$$

(siehe Greenacre, 1993) der Kontingenztafeln $\mathbf{Z}_i{}^T \mathbf{Z}_j$ mit der Gesamtsumme f, den relativen Häufigkeiten $f_{st}$ sowie den relativen Randhäufigkeiten $f_{s.}$ und $f_{.t}$ ($s = 1, ..., r_i$; $t = 1, ..., r_j$) und berechnet $\chi^2(\mathbf{B})$ entsprechend, so gilt

$$\Phi^2(\mathbf{B}) = \frac{\chi^2(\mathbf{B})}{n \cdot p^2} = \frac{\sum_i \sum_{j \neq i} \chi^2(\mathbf{Z}_i{}^T \mathbf{Z}_j) + n \cdot (r - p)}{n \cdot p^2}.$$

Der dominante Term $n \cdot (r - p)$ von $\chi^2(\mathbf{B})$ ergibt sich aus $n \cdot (r - p) = \sum_j \chi^2(\mathbf{Z}_j{}^T \mathbf{Z}_j)$ als Summe der $\chi^2$-Unabhängigkeitstest-Prüfgrößen $\chi^2(\mathbf{Z}_j{}^T \mathbf{Z}_j)$ der Diagonalmatrizen $\mathbf{Z}_j{}^T \mathbf{Z}_j$ von **B**.

Sowohl bei der Analyse von **Z** als auch bei der Analyse von **B** steigt damit die Gesamtvarianz der Punktwolke mit steigender Anzahl von Ausprägungen je Merkmal und führt schließlich zu sinkenden Erklärungsbeiträgen $\lambda_\alpha{}^Z$ und $\lambda_\alpha{}^B$ der Faktorenachsen (siehe Greenacre, 1988).

# 2 Weiterentwicklung der Korrespondenzanalyse

## 2.1 Achsen-Reskalierung einer Korrespondenzanalyse

Die Gesamtvarianz $\Phi^2(\mathbf{B})$ der Burt-Matrix **B** ist mit der durchschnittlichen Varianz sämtlicher Kontingenztafeln $\mathbf{Z}_i{}^T \mathbf{Z}_j$ ($i \neq j$),

$$\overline{\Phi}^2 = \frac{1}{p \cdot (p-1)} \sum_i \sum_{j \neq i} \Phi^2 (\mathbf{Z}_i{}^T \mathbf{Z}_j),$$

laut Gleichung 2 gleich

$$\Phi^2(\mathbf{B}) = \frac{p-1}{p} \cdot \overline{\Phi}^2 + \frac{r-p}{p^2} \text{ (siehe Greenacre, 1993).}$$

Um im Rahmen der Analyse den Kontingenztafeln $\mathbf{Z}_i{}^T\mathbf{Z}_j$ $(i \neq j)$ auf Kosten der Diagonalmatrizen $\mathbf{Z}_j{}^T\mathbf{Z}_j$ mehr Gewicht zu verleihen, definiert Greenacre (1990) anstelle von $\Phi^2(\mathbf{B})$ die durchschnittliche Varianz $\overline{\Phi}^2$ der $p \cdot (p-1)$ Kontingenztafeln $\mathbf{Z}_i{}^T\mathbf{Z}_j$ $(i \neq j)$ als zu erklärende Varianz der Punktwolke.

In Anlehnung an Benzécri (1979) wird von Greenacre (1990) für $\lambda_\alpha{}^Z > \frac{1}{p}$ die Berechnung der positiven Eigenwerte $\rho(\lambda_\alpha{}^Z)$ der Analyse einer modifizierten Burt-Matrix **B'** vorgeschlagen, bei der die Diagonalelemente der Matrizen $\mathbf{Z}_j{}^T\mathbf{Z}_j$ gleich Null gesetzt werden. Es gilt

$$\rho(\lambda_\alpha{}^Z) = \left(\frac{p}{p-1}\right)^2 \cdot \left(\lambda_\alpha{}^Z - \frac{1}{p}\right)^2.$$

Als prozentuale Varianzanteile ergeben sich die Werte $\frac{\rho(\lambda_\alpha{}^Z)}{\overline{\Phi}^2}$.

Die Achsen $\mathbf{v}_\alpha$ der Analyse von **B** werden anschließend mit $\rho(\lambda_\alpha{}^Z)$ reskaliert: die mittels Division mit $\lambda_\alpha{}^Z$ standardisierten Faktorenkoordinaten der Merkmalsausprägungen werden mit $\rho(\lambda_\alpha{}^Z)$ multipliziert.

## 2.2 Numerische Beurteilung der Achsen-Reskalierung

Zur numerischen Beurteilung der Achsen-Reskalierung werden Daten von 98 amerikanischen High School Abgängern analysiert. Die Daten sind van Buuren und Heiser (1989) entnommen und sind ursprünglich Teil eines umfangreicheren Datensatzes in Fienberg (1980). Der Datensatz umfaßt die nominalen Merkmale

- der Abgänger möchte am College studieren - ColP: ja - Yes, nein - No
- Ausmaß der elterlichen Unterstützung des Abgängers - PEnc: niedrig - Low, hoch - High

sowie die ordinalen Merkmale

- sozio-ökonomischer Status - SES: niedrig - L, unterdurchschnittlich - LM, überdurchschnittlich - UM, hoch -H
- Intelligenz - IQ: niedrig - L, unterdurchschnittlich - LM, überdurchschnittlich - UM, hoch - H.

Sämtliche Auswertungen erfolgen mit dem Programmpaket BMDP, Release 7 (siehe BMDP Statistical Software Inc., 1992).

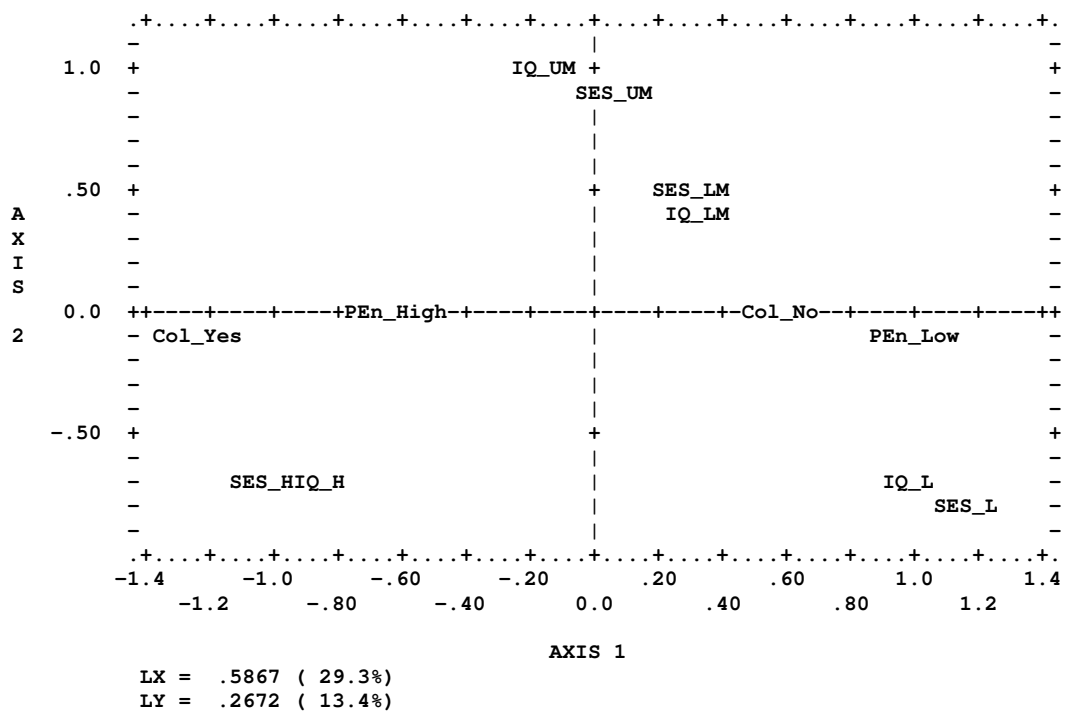


Abbildung 1: Repräsentation der Ausprägungen von IQ, ColP, PEnc, SES in der Ebene von $\mathbf{v}_1$ und $\mathbf{v}_2$

Abbildung 1 zeigt die Repräsentation der Ausprägungen der vier qualitativen Merkmale in der Ebene der Faktorenachsen $\mathbf{v}_1$ und $\mathbf{v}_2$ im Rahmen der multiplen Korrespondenzanalyse der Indikatormatrix **Z**. Die Achse $\mathbf{v}_1$ zeigt deutlich den Gegensatz der Ausprägungen Yes und No bzw. High und Low der nominalen Merkmale ColP und PEnc. Darüber hinaus trennt sie die Ausprägungen H und L der ordinalen Merkmale IQ und SES. $\mathbf{v}_2$ trennt die Ausprägungen UM und LM der ordinalen Merkmale IQ und SES von ihren extremen Ausprägungen H und L. Die Ausprägungen H, UM, LM und L der beiden ordinalen Merkmale IQ und SES sind in der Ebene der ersten beiden Achsen $\mathbf{v}_1$ und $\mathbf{v}_2$ entlang einer nach unten geöffneten Parabel angeordnet. Dieser, mit Horseshoe bezeichneter Effect ist typisch für die Repräsentation ordinaler Merkmale mit Hilfe der Korrespondenzanalyse und offenbart nicht-lineare Abhängigkeiten zwischen den einzelnen Faktorenachsen bei der Abbildung von Rangordnungen (siehe van Rijckevorsel, 1986).

Die Erklärungsbeiträge der Achsen sind $\lambda_1^Z = 0.587$ und $\lambda_2^Z = 0.267$. Die erste Achse erklärt 29.3% der Gesamtvarianz. Beide Achsen erklären zusammen 42.7% der Varianz von **Z**. Wird **B** anstelle von **Z** analysiert, so gilt für die Erklärungsbeiträge der Achsen $\lambda_1^B = 0.344$ und $\lambda_2^B = 0.071$. Die erste Faktorenachse erklärt nun 52.5% der betrachteten Gesamtvarianz. $\mathbf{v}_1$ und $\mathbf{v}_2$ erklären zusammen 63.4% von $\Phi^2(\mathbf{B})$.

Zur Erhöhung der Erklärungsbeiträge wird nun die Achsen-Reskalierung herangezogen:

$\rho(\lambda_1^Z) = 0.201$ und $\rho(\lambda_2^Z) = 0.000514$ können direkt aus $\lambda_1^Z$ und $\lambda_2^Z$ berechnet werden. Mit $\Phi^2(\mathbf{B}) = 0.6553$ ergibt sich $\overline{\Phi}^2 = 0.207$ sowie ein unrealistisch hoher relativer Varianzanteil der ersten Faktorenachse von 97% an $\overline{\Phi}^2$. Der entsprechend berechnete relative Varianzanteil der Achse $\mathbf{v}_2$ an $\overline{\Phi}^2$ ist verschwindend gering.

**TOTAL INERTIA = SUM OF EIGENVALUES = 0.6094**

| AXIS | EIGENVALUE | % OF INERTIA | CUM % | HISTOGRAM |
|---|---|---|---|---|
| 1 | 0.201 | 33.1 | 33.1 | ************************ |
| 2 | 0.111 | 18.2 | 51.3 | ************* |
| 3 | 0.111 | 18.2 | 69.5 | ************* |
| 4 | 0.111 | 18.2 | 87.8 | ************* |
| 5 | 0.037 | 6.1 | 93.9 | **** |
| 6 | 0.028 | 4.5 | 98.4 | *** |
| 7 | 0.009 | 1.4 | 99.8 | * |
| 8 | 0.001 | 0.1 | 99.9 | |
| 9 | 0.000 | 0.1 | 100.0 | |
| 10 | 0.000 | 0.0 | 100.0 | |
| 11 | 0.000 | 0.0 | 100.0 | |

Abbildung 2: Prozentuale Verteilung der Erklärungsbeiträge der Korrespondenzanalyse von **B**'

Vergleicht man das Spektrum von **B**' mit den unmittelbar aus $\lambda_\alpha^Z$ berechneten Werten $\rho(\lambda_\alpha^Z)$, so gilt $\lambda_1^{B'} = \rho(\lambda_1^Z) = 0.201$ und $\lambda_8^{B'} = \rho(\lambda_2^Z) = 0.000514$ (siehe Abbildung 2).

Die drei Erklärungsbeiträge $\lambda_2^{B'}$, $\lambda_3^{B'}$ und $\lambda_4^{B'}$ sind mathematische Artefakte und entsprechen drei bzw. allgemein p - 1 Eigenwerten der Korrespondenzanalyse der Burt-Matrix **B** mit $\lambda_\alpha^B = 0$. Es gilt $\lambda_2^{B'} = \lambda_3^{B'} = \lambda_4^{B'} = \frac{1}{(p-1)^2} = 0.111$ (siehe Greenacre, 1984).

Die Eigenwerte $\lambda_5^{B'}$, $\lambda_6^{B'}$ und $\lambda_7^{B'}$ sind negativ. Für die dazugehörigen Erklärungsbeiträge der Analyse von **Z** bzw. **B** gilt $\lambda_\alpha^Z < \frac{1}{p}$ bzw. $0 < \lambda_\alpha^B < \frac{1}{p^2}$. Dies erklärt den erheblichen Größenunterschied zwischen den beiden Reskalierungswerten $\rho(\lambda_1^Z)$ und $\rho(\lambda_2^Z)$.

Die Eigenwerte $\lambda_9^{B'}$, $\lambda_{10}^{B'}$ und $\lambda_{11}^{B'}$ haben so geringe Beträge, daß sie vernachlässigt werden können.

## 2.3 Joint Correspondence Analysis

Greenacre (1988, 1989) schlägt als Maßzahl im Rahmen der Joint Correspondence Analysis

$$S = \sum_i \sum_{j>i} \chi^2(\mathbf{Z}_i{}^T\mathbf{Z}_j) \text{ bzw. } S' = \sum_i \sum_{j>i} \Phi^2(\mathbf{Z}_i{}^T\mathbf{Z}_j)$$

anstelle von $\Phi^2(\mathbf{B})$ für die zu erklärende Varianz einer Burt-Matrix **B** mit p Merkmalen vor. Im Gegensatz zur multiplen Korrespondenzanalyse, die die Punktwolke der vollständigen Matrix **B** einschließlich der p Diagonalmatrizen $\mathbf{Z}_j{}^T\mathbf{Z}_j$ in einem Teilraum des $\Re^r$ möglichst kleiner Dimension annähern möchte, verfolgt die von Greenacre (1988) entwickelte Joint Correspondence Analysis lediglich die Anpassung der Punktwolke der oberen Dreiecksmatrix der Kontingenztafeln $\mathbf{Z}_i{}^T\mathbf{Z}_j$ (i < j). Die Anpassung erfolgt iterativ nach dem Kriterium gewichteter kleinster Quadrate.

Die Joint Correspondence Analysis liefert als Lösung eine modifizierte Burtmatrix $\mathbf{B}^*_{(JCA)}$ und eine quadratische Diagonalmatrix $\mathbf{C}_{(JCA)}$. $\mathbf{B}^*_{(JCA)}$ rekonstruiert die Matrizen $\mathbf{Z}_i{}^T\mathbf{Z}_j$ vollständig. Die Elemente von $\mathbf{C}_{(JCA)}$ sind Residuen, die die Differenz zwischen den positiven Elementen der Diagonalmatrizen $\mathbf{Z}_j{}^T\mathbf{Z}_j$ von **B** und den Diagonalelementen von $\mathbf{B}^*_{(JCA)}$ ausgleichen. Es gilt somit $\mathbf{B} = \mathbf{B}^*_{(JCA)} + \mathbf{C}_{(JCA}$.

Die Repräsentation der Merkmalsausprägungen sowie die Bestimmung der Erklärungsbeiträge der Achsen erfolgt anschließend über die multiple Korrespondenzanalyse von $\mathbf{B}^*_{(JCA)}$ (siehe Greenacre, 1988).

Greenacre (1993) kündigt die Erweiterung des von ihm entwickelten Programms zur Korrespondenzanalyse SimCA 2 (siehe Greenacre Research, 1990) um ein Modul zur Joint Correspondence Analysis an. Da das Programm SimCA 2 zur Durchführung dieser Arbeit nicht verfügbar war, steht ein Verfahrenstest der Joint Correspondence Analysis mit den herangezogenen Daten noch aus.

## 2.4 Burt-Matrix-Modifikation über direkte Kommunalitätenschätzer

Greenacre (1994) interpretiert die Diagonalelemente von $\mathbf{B}^*_{(JCA)}$ und $\mathbf{C}_{(JCA)}$ als Schätzwerte für die gemeinsame und spezifische Varianz der p Merkmale. Die Beziehung zwischen der multiplen Korrespondenzanalyse und der Joint Correspondence Analysis wird daher wie das Verhältnis von Hauptkomponenten- und Hauptachsenanalyse zur Repräsentation quantitativer Daten betrachtet. Zur Darstellung der Hauptkomponenten- und Hauptachsenanalyse sei beispielsweise auf Marinell (1995) verwiesen.

Neben iterativen Verfahren zur Kommunalitätenschätzung erfolgt die Bestimmung der gemeinsamen Varianzanteile der betrachteten Merkmale bei der Hauptachsenanalyse oftmals direkt durch Schätzer wie den betragsmäßig größten Korrelationskoeffizienten bzw. das Quadrat der multiplen Korrelation der Merkmale (siehe Bortz, 1993).

Der folgende Vorschlag des Verfassers möchte diese Vorgehensweise durch Verwendung geeigneter Schätzer für die Kommunalität der p qualitativen Merkmale auf die multiple Korrespondenzanalyse übertragen.

Die binäre Korrespondenzanalyse der Kontingenztafel $\mathbf{K} = \mathbf{Z}_i{}^T\mathbf{Z}_j$ ($i \neq j$; fest) stimmt mit der kanonischen Analyse der Indikatormatrix $\mathbf{Z} = (\mathbf{Z}_i | \mathbf{Z}_j)$ überein und die Erklärungsbeiträge $\lambda_\alpha$ der binären Korrespondenzanalyse sind mit den quadrierten

kanonischen Korrelationskoeffizienten $\gamma_\alpha^2$ der kanonischen Variablen identisch (siehe Böhm, 1997).

Die unmittelbare Übertragung des Formalismus der binären Korrespondenzanalyse auf die p-variate Indikatormatrix $\mathbf{Z} = (\mathbf{Z}_1|...|\mathbf{Z}_p)$ im Rahmen der multiplen Korrespondenzanalyse französicher Prägung erlaubt grundsätzlich nicht die exakte Interpretation der Erklärungsbeiträge $\lambda_\alpha^Z$ als quadrierte multiple Korrelationskoeffizienten. Trotzdem kann durch einen großen und dominanten, nicht trivialen Eigenwert $\lambda_1^Z$ mit $0 \leq \lambda_\alpha^Z \leq 1$ auf einen hohen Anteil der gemeinsamen Varianz an der Gesamtvarianz der p Merkmale geschlossen werden.

Es werden deshalb folgende Annahmen getroffen:

- Die im bivariaten Fall korrekte Interpretation der Erklärungsbeiträge $\lambda_\alpha$ als quadrierte kanonische Korrelationskoeffizienten $\gamma_\alpha^2$ kann zur Kommunalitätenschätzung näherungsweise auf den p-variaten Fall übertragen werden.
- $\sqrt{\lambda_1^Z}$ bzw. $\lambda_1^Z$ können als einheitliche Schätzer für die Kommunalität der p qualitativen Merkmale einer Indikatormatrix $\mathbf{Z}$ verwendet werden. $\sqrt{\lambda_1^Z}$ entspricht dann einem Korrelationskoeffizienten; $\lambda_1^Z$ einem Bestimmtheitsmaß.

Um die positiven Elemente der Diagonalmatrizen $\mathbf{Z}_j^T\mathbf{Z}_j$ der ursprünglichen Burt-Matrix $\mathbf{B}$ besser an die Kontingenztafeln $\mathbf{Z}_i^T\mathbf{Z}_j$ $(i \neq j)$ anzupassen, werden die Diagonalmatrizen $\mathbf{Z}_j^T\mathbf{Z}_j$ von $\mathbf{B}$ mit dem Schätzwert $\sqrt{\lambda_1^Z}$ oder $\lambda_1^Z$ für den Anteil der gemeinsamen Varianz an der Gesamtvarianz der p Merkmale multipliziert.

Wird beispielsweise $\sqrt{\lambda_1^Z}$ als Kommunalitätenschätzer verwendet, so wird $\mathbf{B}$ durch

$$\mathbf{B}^{mod} = \begin{pmatrix} \sqrt{\lambda_1^Z} \cdot \mathbf{Z}_1^T\mathbf{Z}_1 & \cdots & \mathbf{Z}_1^T\mathbf{Z}_p \\ \vdots & \ddots & \vdots \\ \mathbf{Z}_p^T\mathbf{Z}_1 & \cdots & \sqrt{\lambda_1^Z} \cdot \mathbf{Z}_p^T\mathbf{Z}_p \end{pmatrix}$$

ersetzt. Anschließend erfolgt die multiple Korrespondenzanalyse von $\mathbf{B}^{mod}$.

Die Anpassung der Diagonalelemente von $\mathbf{B}$ erfolgt hierbei undifferenziert über den einheitlichen Schätzer $\sqrt{\lambda_1^Z}$.

Bei der Analyse des bekannten Datensatzes wird aus der Burt-Matrix $\mathbf{B}$ unter Verwendung des Kommunalitätenschätzers $\sqrt{\lambda_1^Z} \approx 0.7662$ die Matrix $\mathbf{B}^{mod}$ mit den an die Kontingenztafeln $\mathbf{Z}_i^T\mathbf{Z}_j$ $(i \neq j)$ angepaßten Elementen in der Hauptdiagonalen berechnet. Die multiple Korrespondenzanalyse von $\mathbf{B}^{mod}$ liefert für die ersten beiden Faktorenachsen $\mathbf{v}_1^{B^{mod}}$ und $\mathbf{v}_2^{B^{mod}}$ die Erklärungsbeiträge $\lambda_1^{B^{mod}} = 0.314$ und $\lambda_2^{B^{mod}} = 0.048$. Mit $\Phi^2(\mathbf{B}^{mod}) = 0.5153$ hat die erste Faktorenachse einen prozentualen Erklärungsbeitrag von 61%. Die Summe der relativen Erklärungsbeiträge von $\mathbf{v}_1^{B^{mod}}$ und $\mathbf{v}_2^{B^{mod}}$ ist 70.4%.

Der Vergleich von Abbildung 3 mit Abbildung 1 zeigt, daß die Repräsentation der Merkmalsausprägungen in den Ebenen der Achsen $\mathbf{v}_1^{Bmod}$ und $\mathbf{v}_2^{Bmod}$ gegenüber ihrer ursprünglichen Repräsentation im Rahmen der Analyse von $\mathbf{Z}$ bzw. $\mathbf{B}$ nicht verzerrt ist.

Treten wie im vorliegenden Fall bei der Analyse von $\mathbf{B}^{mod}$ keine Verzerrungen auf, so kann die Erklärungsgüte und Aussagekraft der Korrespondenzanalyse von **B** oder **Z** auf Basis der höheren prozentualen Erklärungsbeiträge $\frac{\lambda_\alpha{}^{B^{mod}}}{\Phi^2(B^{mod})}$ realistischer beurteilt werden.

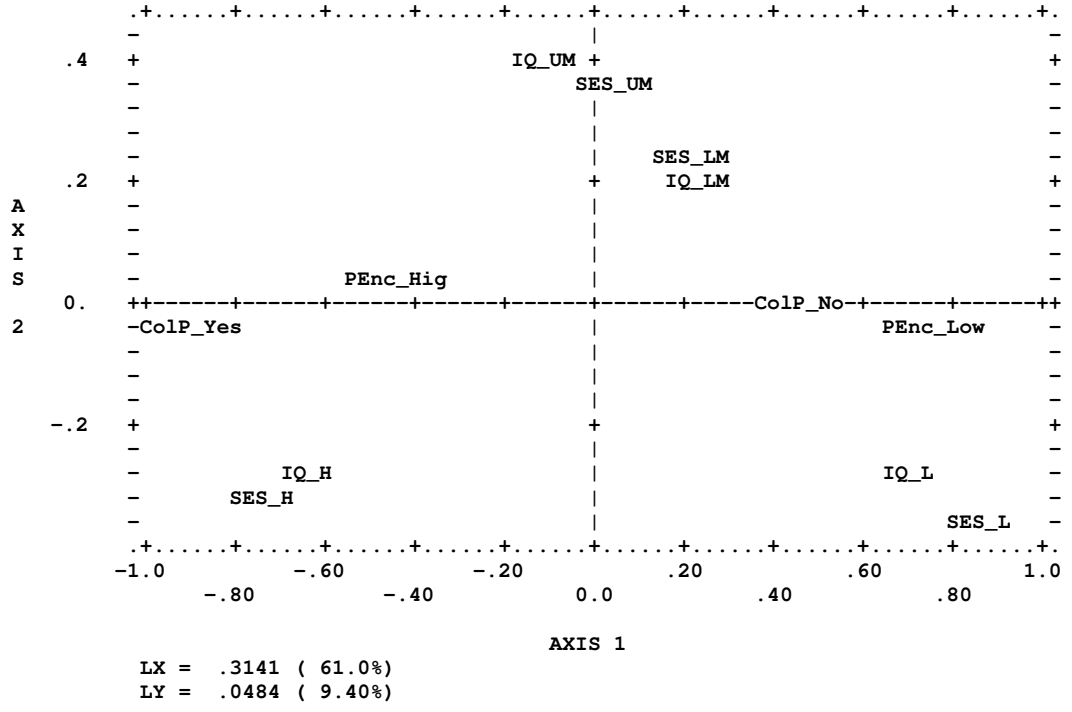


Abbildung 3: Repräsentation der Ausprägungen von IQ, ColP, PEnc, SES bezüglich $\mathbf{v}_1{}^{B^{mod}}$ und $\mathbf{v}_2{}^{B^{mod}}$

# 3 Zusammenfassung und Ausblick

Im Unterschied zur Achsen-Reskalierung und der Joint Correspondence Analysis verzichtet der vorgestellte Ansatz zur Erhöhung der prozentualen Erklärungsbeiträge einer multiplen Korrespondenzanalyse auf die definitorische Reduzierung der zu erfassenden Varianz der betrachteten Punktwolke. In Anlehnung an die Joint Correspondence Analysis, die iterativ die gemeinsame und spezifische Varianz der Merkmale zur differenzierten Anpassung der Diagonalelemente der Burt-Matrix an die Teilmatrizen ihrer oberen Dreiecksmatrix ermittelt, werden mit Hilfe direkter Schätzer für die gemeinsame Varianz der Merkmale eine einfache Modifikation der Burt-Matrix ermöglicht (Multiplikation der Diagonalelemente mit dem Schätzer). Als Kommunalitätenschätzer wird die Quadratwurzel des größten Eigenwerts der Korrespondenzanalyse der der Burt-Matrix zugrunde liegenden Indikatormatrix verwendet. Im

numerischen Test weist die Korrespondenzanalyse der so modifizierten Burt-Matrix bei höheren prozentualen Erklärungsbeiträge mit der ursprünglichen Korrespondenzanalyse vergleichbare Repräsentationsergebnisse auf.

Ob die vorgeschlagene Vorgehensweise in anderen Anwendungsfällen zu unverzerrten Ergebnissen führt, wird jedoch jeweils zu prüfen sein.

Das Verfahren erfordert keine besondere Software, kann ohne größeren Aufwand durchgeführt werden und erhöht nach Meinung des Verfassers grundsätzlich das Verständnis über die zu analysierenden Daten.

# Danksagung

Die Arbeit ist Teil einer 1997 am Institut für Statistik der Universität Innsbruck in Zusammenarbeit mit dem Institut für Statistik und Mathematische Wirtschaftstheorie der Universität Karlsruhe angefertigten Dissertation. Gegenstand dieser Dissertation ist die Anwendung und Weiterentwicklung der Korrespondenzanalyse zur Faktoren-, Cluster- und Diskriminanzanalyse qualitativer Daten. Es wird aufgezeigt, daß die Korrespondenzanalyse als universelles und flexibles Werkzeug der qualitativen Datenanalyse den Zugang des Anwenders zu den grundlegenden taxonomischen Aufgabenstellungen der Repräsentation, Klassifikation und Identifikation ermöglicht (siehe Opitz, 1980).

Abschließend danke ich Herrn Prof. Dr. Gerhard Marinell und Herrn Prof. Dr. Kuno Egle für die Unterstützung bei meiner Dissertation.

# Literatur

J.-P. Benzécri et al.. *L'Analyse des Données. Tome 1: La Taxonomie. Tome 2: L'Analyse des Correspondances*. Dunod, Paris, 1973.

J.-P. Benzécri. Sur l'Analyse des Tableaux Binaires Associés à une Correspondance Multiple. *Les Cahiers de l'Analyse des Données*, 2: 55-71, 1977.

J.-P. Benzécri. Sur le Calcul des Taux d'Inertie dans l'Analyse d'un Questionnaire. *Les Cahiers de l'Analyse des Données*, 4: 377-378, 1979.

BMDP Statistical Software Inc.. *BMDP Statistical Software Manual*. University of California Press, Berkeley, 1992.

B. Böhm. *Anwendung und Weiterentwicklung der Korrespondenzanalyse zur Faktoren-, Cluster- und Diskriminanzanalyse qualitativer Daten*. Dissertation am Institut für Statistik, LEOPOLD-FRANZENS-Universität Innsbruck, 1997.

J. Bortz. *Statistik für Sozialwissenschaftler*. Springer, Berlin, 4. Auflage, 1993.

C. Burt. The Factorial Analysis of Qualitative Data. *Br. J. Psychol. (Statistical Section)*, 3: 166-185, 1950.

J. D. Carroll. Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. In *Proceedings of 76th Annual Convention of the American Psychological Association*, 3: 227-228, 1968.

S. Fienberg. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Camebridge, 2nd edition, 1980.

M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

M. Greenacre. Correspondence Analysis of Multivariate Categorical Data by Weighted Least - Squares. *Biometrika*, 75: 457-467, 1988.

M. Greenacre. Measuring Total Variation and its Components in Multiple Correspondence Analysis. *Supplement to the conference „Journées Inernationales de l'Analyse des Données et Informatique"*: 13-21, 1989.

M. Greenacre. Some Limitations of Multiple Correspondence Analysis. *Computational Statistics Quaterly*, 3: 249-256, 1990.

M. Greenacre. *Correspondence Analysis in Practice*. Academic Press, London, 1993.

M. Greenacre. Multiple and Joint Correspondence Analysis. In M. Greenacre, J. Blasius, editors, *Correspondence Analysis in the Social Sciences*, pages 141-161. Academic Press, London, 1994.

Greenacre Research. *SimCA Version 2 User's Manual*. Irene, 1990.

M. Jambu. *Explorative Datenanalyse*. G. Fischer, Stuttgart. 1992.

G. Marinell. *Multivariate Verfahren*. Oldenbourg, München, Wien, 4. Auflage, 1995.

O. Opitz. *Numerische Taxonomie*. G. Fischer, Stuttgart, 1980.

M. Tenenhaus, F. Young. An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, 50 (1): 91-119, 1985.

S. van Buuren, W. Heiser. Clustering N Objects into K Groups under Optimal Scaling of Variables. *Psychometrika*, 54 (4): 699-706, 1989.

J. Rijckevorsel. About Horeshoes in Multiple Correspondence Analysis. In W. Gaul, M. Schader, editors, *Classification as a Tool of Research*, pages 377-388. North-Holland, Amsterdam, 1986.

Adresse des Autors:

Dipl.-Wirtschaftsing. Dr. Bernhard Böhm
Ob den Gärten 21
D-76228 Karlsruhe
e-Mail: b.boehm@t-online.de