UNIVERSITÄT POTSDAM
Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 5

Jörg Betzin

Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten

Potsdam 1996
ISSN 0949-068X

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 5

Jörg Betzin

Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten

Herausgeber: Lehrstuhl Statistik (Prof. Dr. Hans Gerhard Strohe)
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Universität Potsdam
Postfach 90 03 27
D-14439 Potsdam
Tel. (+49 331) 977-32 25
Fax. (+49 331) 977-32 10
1996
ISSN 0949-068X

Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten

Inhalt

1. Einführung 4

2. Die mathematische Beschreibung des Woldschen Pfadmodells 7

3. Der Begriff der Homogenität und seine Verwandtschaft zur PCA 9

4. Homogenitätsanalyse für das Pfadmodell 13

5. Der PLS-Basis-Algorithmus nach H. Wold 16

6. Kategoriale Daten im Pfadmodell 20

7. Kurzer Abriß zur Korrespondenzanalyse 23

8. Ein korrespondenzanalytischer Ansatz für das Pfadmodell 35

9. Beispiel 37

10. Schlußfolgerungen und Ausblick 44

Anhang 46
- Symboltabelle 46
- Abkürzungsverzeichnis 47
- Liste der im Beispiel verwendeten Variablen- und Kategorienbezeichnungen 48
- Literaturverzeichnis 49

1. Einführung

Pfadmodelle sind in der wirtschaftswissenschaftlichen Anwendung in jüngster Zeit häufiger anzutreffen. Als Pfadmodell wird im vorliegenden Diskussionsbeitrag sowohl ein graphisches Modell zur Darstellung theoretisch fundierter Zusammenhänge oder Hypothesen als auch ein mathematischer Methodenapparat zur Bewertung dieser Zusammenhänge betrachtet. Insbesondere werden hier Pfadmodelle betrachtet, welche latente, nicht direkt meßbare Größen / Merkmale einbeziehen. Diese latenten Größen, im weiteren als latente Variable (LV) bezeichnet, sollen durch manifeste Variable (MV), der Messung zugängliche Größen / Merkmale bestimmbar sein und untereinander kausale bzw. funktionale Beziehungen haben.

Zur Illustration soll ein Beispiel aus dem Marketing-Bereich angeführt werden:

Abb. 1.1: Anspruch an und Wahrnehmung von Standortfaktoren in Abhängigkeit von Unternehmenscharakteristika

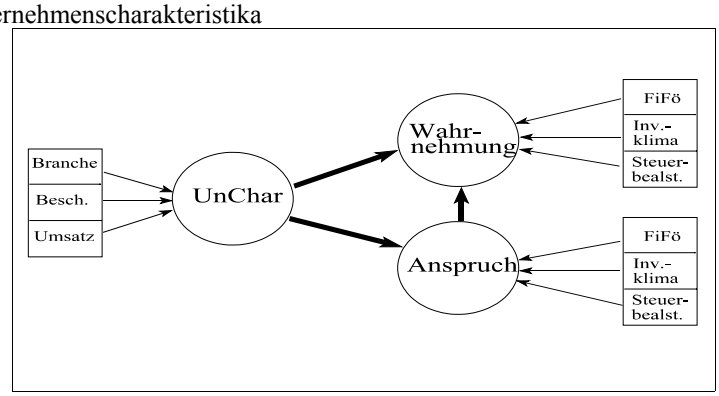
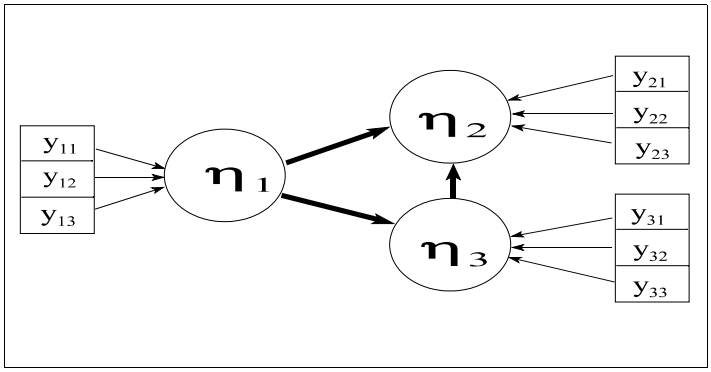


Abb. 1.1 ist ein theoretisches Modell, das den Einfluß bestimmter Unternehmenscharakteristika (UnChar) wie Umsatz, Beschäftigtenzahl und Branchenzugehörigkeit auf den Anspruch an und die Wahrnehmung von Standortfaktoren (SF) darstellen soll. Zusätzlich wird postuliert, daß der Anspruch an Standortfaktoren Einfluß auf deren Wahrnehmung nimmt. Als Standortfaktoren treten Größen wie Lohnkostenniveau, geringe Steuerbelastung, gutes Investitionsklima usw. auf. Eine ausführlichere Beschreibung des Datensatzes und der verwendeten Variablen findet sich in Kapitel 9.

Neben dem Vorteil, daß mit einem solchen Pfadmodell theoretische Konstrukte an empirischen Daten nachgeprüft werden können und sollen, bieten sie den nicht zu unterschätzenden Vorteil, daß die Ergebnisse visuell vermittelbar sind und der Anwender das Gefühl hat, die Struktur in den Daten sehen zu können.
Wenn eine visuelle Darstellung auch nicht das vorrangige Ziel statistischer Analyse ist, so ist der Anwender doch eher bereit, statistische Methoden anzunehmen, die er sehen kann. *Kaum ein Datenanalytiker wird bezweifeln, daß graphische Darstellungen berechneter statistischer Größen deren Interpretation oft erheblich erleichtern.* (Schnell, 1996, S. 1)

Um eine allgemeine mathematische Beschreibung des Pfadmodells zu ermöglichen wird das Modell der Abb. 1.1 in eine abstrakte Form überführt:

Abb. 1.2: Theoretisches Pfadmodell



Hierin lassen sich die in dem komplexen Gleichungsmodell betrachteten Strukturen wie folgt erklären:

1. Die empirischen Daten werden durch y-Variable dargestellt.
2. Es gibt verschiedene Blöcke empirischer Daten, die in einem fachwissenschaftlichen Zusammenhang stehen.
3. Den Datenblöcken werden sogenannte latente Variable η_j zugeordnet, die den zugrundeliegenden fachwissenschaftlichen Sachverhalt widerspiegeln.
4. Die latenten Variablen stehen in einem Zusammenhang, der theoretisch fundiert sein soll und dem eine kausale Richtung unterstellt wird.
5. Die eingetragenen Koeffizienten geben Auskunft über die Gewichtung, die die einzelnen Daten in dem latenten Konstrukt besitzen und über die Stärke des Zusammenhanges der latenten Variablen.

(Abb. 1.2 ist eine vereinfachte Darstellung, die keine Fehlerterme enthält.)

Diese Konstruktion ist wohl eine der komplexesten Möglichkeiten Zusammenhänge zu modellieren ohne an Anschaulichkeit zu verlieren.

Solche Pfadmodelle sind vor allem unter dem Stichwort LISREL[1] bekanntgeworden. LISREL bezeichnet dabei sowohl einen Methodenapparat zur Lösung solcher Gleichungssysteme als auch ein weit verbreitetes Programmpaket zur Berechnung. LISREL wurde zu Beginn der siebziger Jahre von Jöreskog und Sörböm entwickelt und erfreut sich seitdem, nicht zuletzt wegen seiner sehr guten programmtechnischen Umsetzung, einer großen Beliebtheit (s. u. a. Jöreskog / Sörbom, 1989).
Die oben benutzte Bezeichnungsweise (Abb. 1.1 und 1.2) weist auf einen alternativen Weg zur Berechnung pfadanalytischer Modelle hin, der unter dem Namen PLS[2] bekannt geworden ist und von H. Wold, ebenfalls zu Beginn der siebziger Jahre, entwickelt wurde (s. u. a. Wold, 1982).
Der PLS-Apparat beruht dabei, im Gegensatz zum LISREL-Modell, vollständig auf Kleinst-Quadrat-Schätzungen. Er ist daher wesentlich leichter einsetz- und handhabbar, wenn auch nicht so mächtig wie der LISREL-Apparat.

---

1 LISREL = **L**inear **S**tructural **REL**ationships

2 PLS = **P**artial **L**east **S**quare

Ein großer Vorteil ist dabei die starke Nähe von PLS zur Hauptkomponentenanalyse (PCA[3]). Deren Ergebnisse lassen sich, unter der Annahme eines linearen Modells, sehr gut interpretieren und auch graphisch veranschaulichen. Vor allem hat sie nicht mit Verteilungsvoraussetzungen zu kämpfen, die für wirtschaftswissenschaftliche Daten teilweise nur schwer verfügbar sind. Wenn im weiteren Verlauf auf die Behandlung kategorialer Daten eingegangen wird, wird das um so deutlicher.
Im Verfahren der Hauptkomponentenanalyse spielt der Begriff der Homogenität von Variablen eine zentrale Rolle. Die Homogenität von Variablen, als eine allen Variablen gemeinsame Größe *"Historically, the idea of homogeneity is closely related to the idea that different variables may measure 'the same thing.'* (Gifi, 1990, S. 81) wird im 3. Kapitel dargestellt und ihre Bedeutung innerhalb der PCA gezeigt. Unter der Überschrift "Homogenitätsanalyse für das Pfadmodell" wird die Anwendung des Homogenitätsbegriffes anschließend auf ein PLS-Pfadmodell erweitert.
Ein Grundkonzept des von H. Wold entwickelten PLS-Ansatzes ist die Konstruktion latenter Variablen. Der dafür verwendete Basis-Algorithmus wird in Kapitel 5 vorgestellt. Dabei steht, neben der mathematischen Beschreibung des Algorithmus, die Motivation der Vorgehensweise im Kontext des Pfadmodells im Vordergrund.
Während sich bis dahin das Vorgehen nur auf metrische Daten beschränkt, befassen sich die folgenden Kapitel mit dem Eingang von kategorialen Daten in das Modell. Zunächst wird kurz auf das Vorkommen und die Behandlung kategorialer Daten eingegangen. Im Kapitel 7 wird das Verfahren der Korrespondenzanalyse, im Sinne einer Hauptkomponentenanalyse für kategoriale Daten, vorgestellt. In Verbindung mit dem PLS-Basisalgorithmus wird dann im 8. Kapitel die Übertragung der Korrespondenzanalyse in ein Pfadmodell vorgenommen. Abschließend findet sich das in Abb. 1.1 erwähnte Beispiel mit der Bewertung und Interpretation der einzelnen Pfade wieder.

Es sei an dieser Stelle noch auf die verwendete Symbolik eingegangen (um den Leser nicht zu sehr abzuschrecken, wird eine Symboltabelle erst im Anhang dargeboten).
Im folgenden werden mit y stets die beteiligten manifesten Variablen (MV) bezeichnet, die mit einem Index j (j=1, ..., P) behaftet sein können, sobald mehrere MV auftreten. In diesem Falle bezeichnet P die Anzahl der beteiligten MV. Treten mehrere MV-Blöcke auf (s. Abb. 1.2), werden diese mit m=1, ..., M indiziert, Blockindizes erscheinen in Klammern.
Mit $\boldsymbol{\eta}$ bzw. $\boldsymbol{\eta}_{(m)}$ wird die einem MV-Block zugeordnete latente Variable bezeichnet.
Es wird durchweg davon ausgegangen, daß ein Datensatz für N Objekte vorliegt. Die entsprechenden Datenvektoren werden mit $\mathbf{y}_j$ ($\mathbf{y}_{(m)j}$) bzw. $\boldsymbol{\eta}$ ($\boldsymbol{\eta}_{(m)}$) bezeichnet[4]. Objektindizes werden durch den Index i dargestellt, y_{ij} ($y_{(m)ij}$) stellt dann die Ausprägung des i-ten Objektes in der j-ten Variable dar.
Die Matrix $\mathbf{Y}$ ist eine Matrix der Spaltenvektoren $\mathbf{y}_j$, wobei die Anzahl der Spaltenvektoren in der Regel aus dem Kontext ersichtlich wird.
Zur Vereinfachung der Schreibweise werden die y_j ($y_{(m)ij}$), sofern es sich bis Kapitel 5 um metrische Variablen handelt, als standardisert angenommen (d. h. die Vektoren haben einen Mittelwert von 0 und eine (euklidische) Länge von 1). Diese Standardisierung entspricht einer Normierung der verschiedenen Variablen, um Modelleinflüsse, die sich lediglich aus unterschiedlicher Skalierung der Merkmale ergeben, zu eliminieren.

---

3 PCA = **P**rincipal **C**omponent **A**nalysis

4 **fett** geschriebene Symbole bedeuten immer Matrizen bzw. Vektoren

2. Die mathematische Beschreibung des Woldschen Pfadmodells

Zur Beschreibung des mathematischen Konzeptes des PLS-Ansatzes in der Pfadanalyse gehen wir zunächst von metrischem Niveau der vorliegenden Daten aus. D. h. die Merkmalsausprägungen der Variablen sind jeweils Elemente aus dem Bereich der reellen Zahlen, sie besitzen eine Ordnung und die Abstände zwischen zwei Merkmalsausprägungen einer Variable sind interpretierbar.
Der PLS-Ansatz, ebenso wie der LISREL-Ansatz, unterscheidet im Pfadmodell ein Meßgleichungssystem (MGS) und ein Strukturgleichungssystem (SGS).
Im MGS werden die Beziehungen zwischen den Blöcken manifester Variabler und der zugehörigen latenten Variablen modelliert, während das SGS das Zusammenhangsgefüge der latenten Variablen untereinander ausführt. In beiden Gleichungssystemen werden zwischen den Variablen lineare Beziehungen angenommen. In diesem Sinne gehört auch der PLS-Ansatz zu den linearen Strukturgleichungsmodellen.
Im Grundmodell wird weiter angenommen, daß die manifesten Variablenblöcke durch jeweils eine latente Variable repräsentiert werden und es ergibt sich:

MGS: $$\boxed{y_{(m)j} = \eta_{(m)}\lambda_{(m)j} + \theta_{(m)j}} \qquad m = 1, ..., M \,/\, j = 1, ..., P_m^{5}. \tag{2.1}$$

Die manifesten Variablen werden dabei auch als Indikatoren der latenten Variablen, die Koeffizienten $\pi_{(m)j}$ als Ladungen bezeichnet. Die Vektoren $\varepsilon_{(m)j}$ sind die Meßfehler der j-ten Variablen im m-ten Block.

Die Beziehungen im Strukturgleichungssystem werden als kausale Zusammenhänge interpretiert und wie folgt dargestellt:

SGS: $$\boxed{\eta_{(m)} = \sum_{m' \in C_m^{Pr}} \eta_{(m')}\gamma_{(m)(m')} + \epsilon_{(m)}} \qquad m = 1, ..., M. \tag{2.2}$$

Zur Vereinfachung der Schreibweise wird mit C_m die Indexmenge der mit $\eta_{(m)}$ direkt verbundenen LV bezeichnet. Indem wir weiter mit C_m^{Pr} bzw. C_m^{Su} die Indexmenge der "Vorgänger" (**Pr**edecessor) bzw. der "Nachfolger" (**Su**ccessor) von $\eta_{(m)}$ bezeichnen erhalten wir:

$$C_m = C_m^{Pr} \cup C_m^{Su}$$
$$\text{wobei: } C_m^{Pr} \cap C_m^{Su} = \varnothing .$$

(Die Notationen stimmen zum großen Teil mit der von Mathes (1993a) verwendeten Symbolik überein).
Die Koeffizienten $\gamma_{(m)(m')}$ werden als Pfadkoeffizienten bezeichnet, die $\epsilon_{(m)}$ stellen wiederum Fehlergrößen dar. Die latente Variable $\eta_{(m)}$ ist statistisch linear abhängig von ihren im Pfadmodell auftretenden latenten Vorgängervariablen $\eta_{(m')}$. Unter Vorgängervariablen bezogen auf $\eta_{(m)}$ betrachten wir dabei alle diejenigen latenten Variablen, die im Pfadmodell mit $\eta_{(m)}$

---

[5] P_m ist die Anzahl manifester Variabler im m-ten Block

verbunden sind und die entsprechend der Pfeilnotation als ihre Vorgänger auftreten.

Zu den Fehlergrößen in beiden Gleichungssystemen werden hier keine weiteren Angaben gemacht, da das vorliegende Diskussionspapier nicht vorrangig an der Schätzung des MGS bzw. SGS interessiert ist, sondern sein Hauptaugenmerk auf die Bestimmung der latenten Variablen richtet.
Die Bestimmung von Werten (Scores) für die LV ist ein wesentlicher Bestandteil des PLS-Modells: *"Ein wesentlicher Unterschied zwischen PLS und LISREL betrifft die Zusatzannahmen, die das Modell schätzbar machen. In PLS werden die latenten Variablen als gewichtete Aggregate geschätzt; das impliziert, daß die LV-Werte (Faktorwerte) zum konstituiven Modellbestandteil werden."* (Lohmöller, 1984, S. 45)
Mit gewichteten Aggregaten ist eine Aggregation der LV aus den zugehörigen MV gemeint und das konstituierende Gleichungssystem wird als Gewichtungsgleichungssystem (GGS) bezeichnet:

GGS:
$$\eta_{(m)} = \mathbf{Y}_{(m)}\omega_{(m)} \qquad m = 1, ..., M. \tag{2.3}$$

Das GGS ist kein stochastisches Modell im eigentlichen Sinne, sondern postuliert, daß sich die LV als gewichtete Summe der zugehörigen MV ergeben. Die Gewichtskoeffizienten $\omega_{(m)} = (\omega_{(m)1}, \omega_{(m)2}, \ldots, \omega_{(m)P_m})'$ sind zunächst unbekannt und werden in einem folgenden Iterationszyklus unter gewissen Optimierungskriterien bestimmt. P_m ist die Anzahl der MV im m-ten Block.
Das GGS erscheint vorderhand als ein "gewöhnliches" Problem der Hauptkomponentenanalyse und tatsächlich ist das PLS-Modell mit LV *"somit eine Generalisierung der Hauptkomponentenanalyse..."* (Lohmöller 1984, S. 45). Noch nicht berücksichtigt ist allerdings die innere Struktur des Pfadmodells, die Beziehungen der $\mathbf{0}_{(m)}$ untereinander. Sie findet in der besonderen Wahl der Gewichte $\mathbf{T}_{(m)}$ ihren Ausdruck.
Zu diesem Zweck werden sogenannte Umgebungsvariable definiert:

UV:
$$\eta^*_{(m)} = \sum_{m' \in C_m} \eta_{(m')} r_{(m)(m')} \qquad m = 1, \ldots, M. \tag{2.4}$$

Zu beachten ist, daß in die $\eta^*_{(m)}$ die Beziehungen zu allen mit $\eta_{(m)}$ benachbarten LV, also sowohl den Vorgängern als auch den Nachfolgern eingehen. Bei der Konstruktion der LV werden die kausalen Richtungen noch nicht berücksichtigt, dies erfolgt erst bei der Schätzung der Strukturkoeffizienten im inneren Modell (SGS).
Die Lösung des Pfadmodells erfolgt jetzt im PLS-Modell in 2 Stufen. In der ersten Stufe werden die latenten Variablen bzw. die dazu benötigten Gewichte $\mathbf{T}_{(m)}$ berechnet und in der 2. Stufe werden die Ladungs- und Pfadkoeffizienten des MGS und SGS mittels OLS-Methoden geschätzt.

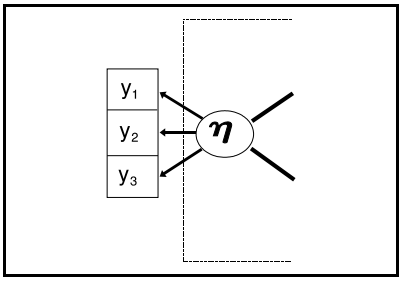
Die erste Stufe der Bestimmung der latenten Variablen wird in Kapitel 5 unter dem Stichwort PLS-Basis-Algorithmus dargestellt. Zunächst soll aber das im PLS-Ansatz enthaltene Modell der Hauptkomponentenanalyse näher untersucht und insbesondere mit dem Begriff der Homogenität von Variablen verknüpft werden.

3. Der Begriff der Homogenität und seine Verwandtschaft zur PCA

Der bereits in der Einführung zitierte Gifi schreibt in dem Kapitel zur Einführung der Homogenitätsanalyse: *"Since the early beginnings of quantitative social science there has been a lively interest in the problem of reduction of multivariate data to univariate scales by means of 'weighted averaging'."* (Gifi, 1990, S. 82)

Das ist die klassische Aufgabe der Faktorenanalyse und der Hauptkomponentenanalyse.
Betrachten wir zunächst ein spezielles Pfadmodell, mit nur einem Block manifester Variablen:

Abb. 3.1: 1-Block Pfadmodell



Definieren wir die LV η als gewichtete Linearkombination der Variablen $\mathbf{y}_j$:

$$\eta := \frac{1}{P}\sum_{j=1}^{P}\mathbf{y}_j\omega_j, \tag{3.1}$$

mit geeigneten (skalaren) Gewichten ω_j, so werden die $\mathbf{y}_j$ durch η umso besser repräsentiert, je "homogener" die Variablen y_j untereinander sind.
Die Homogenität zweier Vektoren $\mathbf{x}$ und $\mathbf{y}$ im $\mathbb{R}^N$ messen wir dabei als quadrierten euklidischen Abstand:

$$\|\mathbf{x}-\mathbf{y}\|^2 := \sum_{i=1}^{N}(x_i-y_i)^2. \tag{3.2}$$

Ist der Vektor $\mathbf{y}$ gegeben und gestatten wir ihm einen Gewichtskoeffizienten ω, so läßt sich eine Abstandsfunktion in Abhängigkeit von $\mathbf{x}$ und ω (bei gegebenem $\mathbf{y}$) definieren:

$$\sigma(\mathbf{x},\omega) := \|\mathbf{x}-\mathbf{y}\omega\|^2. \tag{3.3}$$

Im folgenden sei eine solche Abstandsfunktion $\sigma(...)$ als Verlustfunktion bezeichnet, interpretiert in der Form, daß eine Art Verlust betrachtet wird, der auftritt, wenn wir den Vektor $\mathbf{y}$ durch den Vektor $\mathbf{x}$ ersetzen (messen).
Eine Verlustfunktion für η bzgl. der $\mathbf{y}_j$ kann dann im mehrdimensionalen Fall definiert werden als:

$$\sigma(\eta,\omega) := \frac{1}{P}\sum_{j=1}^{P}\|\eta-\mathbf{y}_j\omega_j\|^2 = \frac{1}{P}\sum_{i=1}^{N}\sum_{j=1}^{P}(\eta_i-y_{ij}\omega_j)^2. \tag{3.4}$$

$\omega = (\omega_1, \omega_2, \ldots, \omega_P)'$ ist hier ein P-dimensionaler Gewichtsvektor, die Verlustfunktion (3.2) ist ein Spezialfall für den eindimensionalen Fall. Die Aufgabe Homogenitätsanalyse besteht darin, solche Gewichte $\mathbf{T}$ zu finden, die die Verlustfunktion (3.4) minimieren.
Offensichtlich erfüllt die Lösung:

$$\omega_j = 0 \qquad \text{für } j=1, \ldots, P$$

und damit

$$\eta_i = 0 \qquad \text{für } i=1, \ldots, N$$

diese Minimierungsaufgabe, vermittelt aber keinerlei Informationen über die Zusammenhänge in den Daten.
Um diese "Nulllösung" zu verhindern, werden i. a. Normierungsbedingungen an $\mathbf{T}$ bzw. $\mathbf{0}$ gestellt. Wir wollen hier fordern, daß die quadrierte Summe der Werte $\mathbf{0}_i$ gleich 1 sei[6], in Vektorschreibweise erhalten wir:

$$\eta'\eta = 1. \tag{3.5}$$

Die Lösung dieser Minimierungsaufgabe entspricht genau dem Anliegen der klassischen Hauptkomoponentenanalyse.

Als Lösung bietet sich ein Verfahren an, das unter dem Namen

Alternierende Kleinste Quadrate (ALS[7])

bekannt ist (s. u. a. Gifi, 1990, S. 88 ff.).

Damit ist ein Algorithmus gemeint, der alternierend Minima der Verlustfunktion (3.4) bzgl. $\mathbf{0}$ und $\mathbf{T}$ im Sinne kleinster Quadrate liefert.
Ausgehend von zunächst beliebigen Startvektoren $\mathbf{T}_0$ wird $\mathbf{0}_0$ als gewogenes arithmetisches Mittel der Variablen $\mathbf{y}_j$ gebildet[8]:

$$\tilde{\eta}_0 = \mathbf{Y}\omega_o$$

und normiert:

$$\eta_0 = \tilde{\eta}_0(\tilde{\eta}_0'\tilde{\eta}_0)^{-1/2}$$

Nach Bereitstellung dieser Startvektoren schließt sich folgender Iterationsalgorithmus an:

$$\text{i)} \quad \tilde{\eta}_{neu} = \mathbf{Y}\omega_{alt}$$

$$\text{ii)} \quad \eta_{neu} = \tilde{\eta}_{neu}(\tilde{\eta}_{neu}'\tilde{\eta}_{neu})^{-1/2}. \tag{3.6}$$

---

6 Das entspricht einer Normierung der latenten Variable auf die (euklidische) Länge 1

7 ALS = **A**lternating **L**east **S**qaures

8 Wir werden aus Gründen der Anschaulichkeit im weiteren Matrixschreibweise verwenden, die entsprechende Symbolik läßt sich aus der Symboltabelle im Anhang entnehmen.

Anschließend werden neue Gewichte $\mathbf{T}$ als gewogenes Mittel gebildet:

$$\text{iii)} \qquad \omega_{\mathbf{neu}} = \mathbf{Y}'\eta_{\mathbf{neu}}. \tag{3.7}$$

Während in i) die Verlustfunktion (3.4) bezüglich fester Werte $\mathbf{T}_{alt}$ minimiert wird, ergibt iii) ein Minimum in (3.4) für die gewählten $\mathbf{0}_{neu}$.
Auf diese Weise wird der Wert der Verlustfunktion schrittweise (alternierend) bzgl. $\mathbf{T}_{alt}$ und $\mathbf{0}_{neu}$ immer kleiner. Die Konvergenz des Algoritmus ist damit gesichert, da die Verlustfunktion nach unten durch Null beschränkt ist.

Zum späteren Gebrauch, insbesondere bei der Behandlung von Pfadmodellen im Zusammenhang mit der Homogenitätsanalyse benötigen wir noch eine "relative" Verlustfunktion, die wir im folgenden erläutern wollen.

Betrachten wir noch einmal die Verlustfunktion (3.4):

$$\sigma(\eta,\omega) := \frac{1}{P}\sum_{j=1}^{p}\|\eta - \mathbf{y}_j\omega_j\|^2,$$

so ergibt sich für die einzelnen Summanden in Matrixschreibweise:

$$\begin{aligned}\|\eta - \mathbf{y}_j\omega_j\|^2 &= (\eta - \mathbf{y}_j\omega_j)'(\eta - \mathbf{y}_j\omega_j)\\ &= \eta'\eta - \omega_j\mathbf{y}_j'\eta - \eta'\mathbf{y}_j\omega_j + \omega_j\mathbf{y}_j'\mathbf{y}_j\omega_j\end{aligned}.$$

Beachtet man jetzt die Summation in (3.4) und die Definition von $\mathbf{0}$ als

$$\eta := \sum_{j=1}^{p}\mathbf{y}_j\omega_j,$$

so ergibt sich:

$$\begin{aligned}\sigma(\eta,\omega) &= \frac{1}{P}(P\cdot\eta'\eta - \sum_{j=1}^{P}(\omega_j\mathbf{y}_j')\eta - \eta'\sum_{j=1}^{P}(\mathbf{y}_j\omega_j) + \sum_{j=1}^{P}(\omega_j\mathbf{y}_j'\mathbf{y}_j\omega_j))\\ &= \frac{1}{P}(P\cdot\eta'\eta - P\cdot\eta'\eta - P\cdot\eta'\eta + \sum_{j=1}^{P}(\omega_j\mathbf{y}_j'\mathbf{y}_j\omega_j))\\ &= \frac{1}{P}\sum_{j=1}^{P}(\omega_j\mathbf{y}_j'\mathbf{y}_j\omega_j) - \eta'\eta\end{aligned}.$$

Schreiben wir die obige Gleichung in Matrixform, ergibt sich:

$$\sigma(\eta,\omega) = \omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega - \omega'\mathbf{Y}'\mathbf{Y}\omega, \quad (!) \tag{3.8}$$

wobei wir mit $\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}$ die Diagonalmatrix mit der Hauptdiagonalen von $\mathbf{Y}'\mathbf{Y}$ bezeichnen:

$$\mathbf{Y}'\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1'\mathbf{y}_1 & \mathbf{y}_1'\mathbf{y}_2 & \cdots & \mathbf{y}_1'\mathbf{y}_P \\ \mathbf{y}_2'\mathbf{y}_1 & \mathbf{y}_2'\mathbf{y}_2 & \cdots & \mathbf{y}_2'\mathbf{y}_P \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_P'\mathbf{y}_1 & \mathbf{y}_P'\mathbf{y}_2 & \cdots & \mathbf{y}_P'\mathbf{y}_P \end{pmatrix}$$

$$\mathbf{D}_{\mathbf{Y}'\mathbf{Y}} = \begin{pmatrix} \mathbf{y}_1'\mathbf{y}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{y}_2'\mathbf{y}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{y}_P'\mathbf{y}_P \end{pmatrix}.$$

Beachtet man, daß die $\mathbf{y}_j$ standardisiert sind, so enthält die Matrix $\mathbf{Y}'\mathbf{Y}$ die empirischen Korrelationskoeffizienten zwischen den manifesten Variablen. Die Matrix $\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}$ enthält dann die empirischen Korrelationskoeffizienten der Variablen mit sich selbst. Diese sind natürlich an dieser Stelle gleich 1. Trotzdem wollen wir die Schreibweise $\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}$ beibehalten um die Analogie zum allgemeinen Fall mehrerer **Y**-Blöcke im folgenden Kapitel zu verdeutlichen.
Nach dieser etwas langwierigen Herleitung läßt sich jetzt zu der "absoluten" Verlustfunktion (3.4) eine "relative" Verlustfunktion[9]:

$$\sigma_{rel}(\eta,\omega) := \frac{\sigma(\eta,\omega)}{\omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega} = 1 - \frac{\omega'\mathbf{Y}'\mathbf{Y}\omega}{\omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega} \tag{3.9}$$

betrachten.
Der Minimierung von (3.9) entspricht dann eine Maximierung von:

$$\tilde{\sigma}(\omega) := \frac{\omega'\mathbf{Y}'\mathbf{Y}\omega}{\omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega}. \tag{3.10}$$

Die Lösung für (3.10) im Sinne einer Maximierungsaufgabe bzgl $\mathbf{T}$ entspricht dann den im ALS-Algorithmus gefunden Gewichtsvektoren.

Die Formulierung der Optimierungsaufgabe für die Verlustfunktion (3.4) als Maximumproblem bzgl. (3.10) erleichtert uns im folgenden den Übergang zum Problem der Homogenitätsanalyse

---

[9] Analog zum Varianzzerlegungssatz in der Varianzanalyse läßt sich $\mathbf{T} = \omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega$ als totale Varianz der gewichteten Y-Variablen definieren und $\mathbf{B} = \omega'\mathbf{Y}'\mathbf{Y}\omega$ als Varianzanteil, der zwischen den Variablen auftritt (B=between).

für mehrere Variablenblöcke.

4. Homogenitätsanalyse für das Pfadmodell

Geht man vom Problem nur eines manifesten Variablenblocks (M=1) über zu dem Problem mehrerer MV-Blöcke (M>1), so lassen sich innerhalb der MV-Blöcke wiederum "optimale" $\mathbf{0}_{(m)}$ (m=1, ..., M) finden, die, jede innerhalb ihres MV-Blockes, die Verlustfunktion (3.4) minimieren.
Sind wir aber an den Strukturbeziehungen zwischen den MV-Blöcken interessiert, sollten solche $\eta_{(m)} := \mathbf{Y}_{(m)}\omega_{(m)}$ konstruiert werden, die untereinander möglichst homogen sind.
Es sei bemerkt, daß hier unter dem Stichwort Homogenitätsanalyse keine gerichteteten Zusammenhänge betrachten werden. Das klassische Beispiel für den Fall zweier MV-Blöcke ist die Methode der kanonischen Korrelationsanalyse.
Analog zu den obigen Ausführungen für das 1-Block-Modell erhält man das folgende Minimumproblem:

$$\sigma(\omega^*) := \underset{\omega}{\text{Min}} \sum_{m=1}^{M} \|\tilde{\eta} - \eta_{(m)}\|^2 \tag{4.1}$$

mit: $\eta_{(m)} = \mathbf{Y}_{(m)}\omega_{(m)}$

und $\eta'_{(m)}\eta_{(m)} = 1$ (Normierungsbedingung),

wobei: $\omega = \begin{pmatrix} \omega_{(1)} \\ \omega_{(2)} \\ \vdots \\ \omega_{(M)} \end{pmatrix}$ $\omega_{(m)} = \begin{pmatrix} \omega_{(m)1} \\ \omega_{(m)2} \\ \vdots \\ \omega_{(m)P_m} \end{pmatrix}$.

$\tilde{\eta}$ wird dabei als Mittelwert der LVdefiniert:

$$\tilde{\eta} := \sum_{m=1}^{M} \eta_{(m)},$$

"um" den die $\mathbf{0}_{(m)}$ möglichst wenig schwanken sollen.

Definieren wir jetzt

$$\mathbf{Y} := (\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots, \mathbf{Y}_{(M)})$$

als Matrix der manifesten Variablen in M Blöcken, so erhalten wir

$$\mathbf{Y}'\mathbf{Y} := \begin{pmatrix} \mathbf{Y}_{(1)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(1)}'\mathbf{Y}_{(2)} & \cdots & \mathbf{Y}_{(1)}'\mathbf{Y}_{(M)} \\ \mathbf{Y}_{(2)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(2)}'\mathbf{Y}_{(2)} & \cdots & \mathbf{Y}_{(2)}'\mathbf{Y}_{(M)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Y}_{(M)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(M)}'\mathbf{Y}_{(2)} & \cdots & \mathbf{Y}_{(M)}'\mathbf{Y}_{(M)} \end{pmatrix}. \tag{4.2a}$$

Dabei sind $\mathbf{Y}_{(m)}'\mathbf{Y}_{(m)}$ Blockmatrizen der Korrelationskoeffizienten im m-ten MV-Block. Indem wir jetzt die Blockdiagonalmatrix $\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}$ definieren als

$$\mathbf{D}_{\mathbf{Y}'\mathbf{Y}} := \begin{pmatrix} \mathbf{Y}_{(1)}'\mathbf{Y}_{(1)} & 0 & \cdots & 0 \\ 0 & \mathbf{Y}_{(2)}'\mathbf{Y}_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Y}_{(M)}'\mathbf{Y}_{(M)} \end{pmatrix}, \tag{4.2b}$$

ergibt sich, in analoger Weise zum 1-Block-Problem, folgendes Maximierungsproblem[10]:

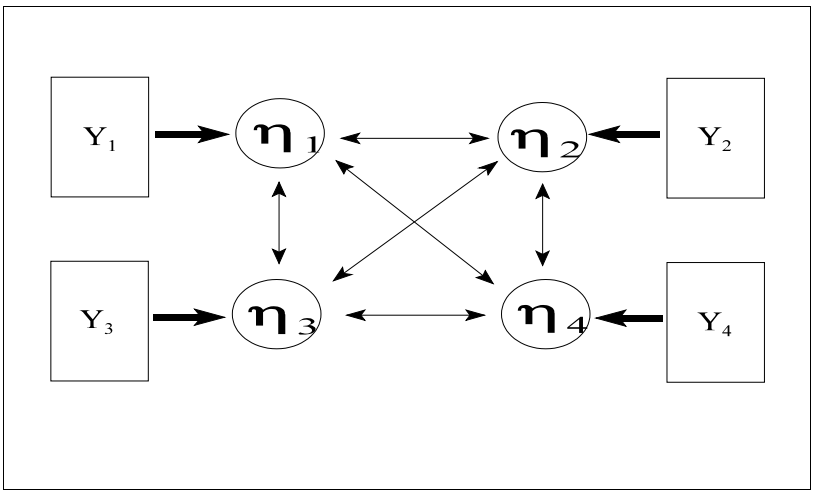
$$\sigma_M(\omega^*) = \underset{\omega}{\text{Max}} \left(\frac{\omega'\mathbf{Y}'\mathbf{Y}\omega}{\omega'\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}\omega} \right). \tag{4.3}$$

Das Maximierungsproblem in (4.3) entspricht dann der Homogenitätsanalyse für mehrere MV-Blöcke mit vollständigem Design. Dabei verstehen wir unter einem vollständigen Design ein Strukturmodell zwischen den LV, in dem jede LV von jeder anderen LV beeinflußt wird. Abb. 4.1 gibt das Beispiel eines vollständigen Designs für M=4 MV-Blöcke:

---

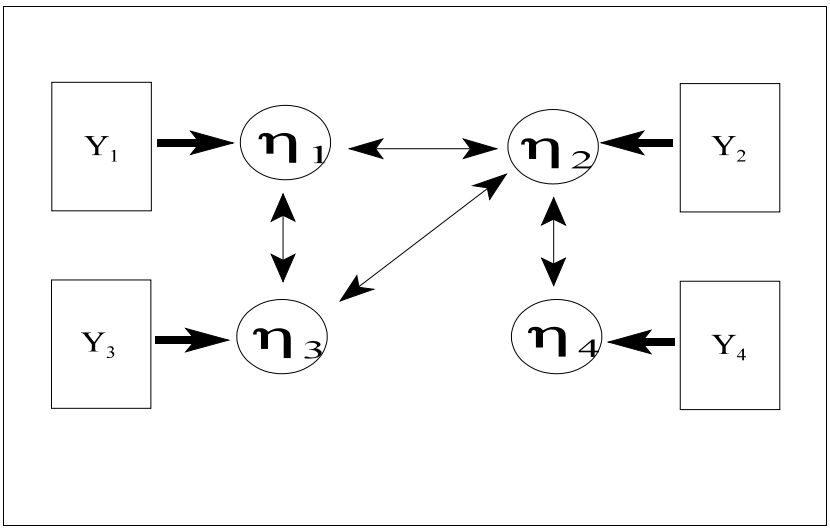
[10] aus Gründen der Übersichtlichkeit wird hier kein Beweis für diese Behauptung angegeben, er kann nachgelesen werden in (Gifi, 1990, S. 100 ff.)

Abb. 4.1: vollständiges Design für M=4 MV-Blöcke



Dieses Modell entspricht der Fragestellung der verallgemeinerten Kanonischen Analyse und kann als Spezialfall eines Pfadmodells angesehen werden.
Im allgemeinen werden im Pfadmodell aber unvollständige Designs betrachtet, d. h. solche Strukturbeziehungen zwischen den LV, in der es LV gibt, die keine direkte Beziehung zueinander haben, s. z. B. Abb. 4.2:

Abb. 4.2: unvollständiges Design für M=4 MV-Blöcke



Die Behandlung unvollständiger Designs in der Homogenitätsanalyse erfolgt durch eine entsprechende Modifizierung der Matrix $\mathbf{Y}'\mathbf{Y}$. Diejenigen Blöcke $\mathbf{Y}_{(m)}'\mathbf{Y}_{(m')}$ werden Null gesetzt, für die keine direkte Beziehung zwischen $\mathbf{0}_{(m)}$ und $\mathbf{0}_{(m')}$ postuliert wird. Für das Beispiel aus Abb. 4.2 ergibt sich:

$$\mathbf{Y}'\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{(1)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(1)}\mathbf{Y}_{(2)} & \mathbf{Y}_{(1)}'\mathbf{Y}_{(3)} & 0 \\ \mathbf{Y}_{(2)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(2)}'\mathbf{Y}_{(2)} & \mathbf{Y}_{(2)}'\mathbf{Y}_{(3)} & \mathbf{Y}_{(2)}'\mathbf{Y}_{(4)} \\ \mathbf{Y}_{(3)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(3)}'\mathbf{Y}_{(2)} & \mathbf{Y}_{(3)}'\mathbf{Y}_{(3)} & 0 \\ 0 & \mathbf{Y}_{(4)}'\mathbf{Y}_{(2)} & 0 & \mathbf{Y}_{(4)}'\mathbf{Y}_{(4)} \end{pmatrix}$$

Das Maximierungsproblem für das M-Block-Problem mit unvollständigem Design entspricht dann (4.3) mit entsprechend modifizierter Matrix $\mathbf{Y}'\mathbf{Y}$.

Die Ausführungen in diesem Kapitel wurden bewußt ohne Beweise und sehr allgemein gehalten, um erstens die enge Beziehung des Verfahrens der Hauptkomponentenanalyse und dem Mehrblockproblem zu verdeutlichen und zweitens, den Leser nicht durch eine Fülle wenig interessanter Herleitungen zu verwirren und vom Hauptgedanken abzulenken. Eine ausführliche Darstellung der Problematik findet sich in (Gifi, 1990).

Im folgenden Kapitel wird der von H. Wold entwickelte PLS-Basis-Algorithmus zur Bestimmung der latenten Variable im Pfadmodell angegeben und im Kontext der bisherigen Erkenntnisse betrachtet. Dabei wird einerseits eine Beziehung zum Alternierende-Kleinst-Quadrat-Algorithmus hergestellt und andererseits die Verwandtschaft zum Problem der Homogenitätsanalyse für das Pfadmodell gezeigt.

5. Der PLS-Basis-Algorithmus nach H. Wold

Zur Lösung von Pfadmodellen mit latenten Variablen entwickelte H. Wold in den siebziger Jahren ein Gleichungsmodell, welches basierend auf Kleinst-Quadrat-Berechnungen, zunächst latente Variable innerhalb des Modells berechnet und anschließend, auf deren Grundlage die Modellparameter schätzt. Im Unterschied zu den, ebenfalls in den siebziger Jahren entwickelten LISREL-Modellen, sind für das PLS-Modell keine Verteilungsvoraussetzungen notwendig und es traten keine Identifikationsprobleme auf.
Zur Berechnung der LV benutzt Wold einen Algorithmus, welcher sich an den ALS-Algorithmus anlehnt, indem wechselseitig Lösungen für $\mathbf{0}$ (s. u. Schritt 4) und $\mathbf{T}$ (s. u. Schritt 3) mittels Kleinster-Quadrate berechnet werden. Zusätzlich gehen in den Algorithmus Beziehungen zwischen den verschiedenen MV-Blöcken als sogenannte Umgebungsvariable ein. Den Beinamen partiell verdient der Algorithmus aus der Tatsache, daß die Iterationszyklen für jeden Block (partiell) durchgeführt werden, unter der Annahme, daß die LV und Parameter für die anderen Blöcke unverändert bleiben.

PLS-Basis-Algorithmus

Schritt 0: Bereitstellung Startgewichte $\omega_{(m)\langle 0\rangle}$ und $\eta_{(m)\langle 0\rangle}$

$$\eta_{(m)\langle 0\rangle} = \mathbf{Y}_{(m)}\omega_{(m)\langle 0\rangle}f_{(m)\langle 0\rangle}$$

Schritt 1: $$\rho_{(mm')\langle k\rangle} = \text{sign}\left(\text{Corr}[\eta_{(m)\langle k\rangle},\eta_{(m')\langle k\rangle}] \quad \text{für } m' \in C_m\right)$$

Schritt 2: $$\eta^*_{(m)\langle k\rangle} = \sum_{m'\in C_m}\eta_{(m')\langle k\rangle}\rho_{(mm')\langle k\rangle}$$

Schritt 3: $$\mathbf{Y}_{(m)} = \eta^*_{(m)\langle k\rangle}\omega'_{(m)\langle k+1\rangle} + \mathbf{U}_{(m)} \qquad \text{(Modus A)}$$

oder

$$\eta^*_{(m)\langle k\rangle} = \mathbf{Y}_{(m)}\omega_{(m)\langle k+1\rangle} + \mathbf{u}_{(m)} \qquad \text{(Modus B)}$$

Schritt 4: $$\eta_{(m)\langle k+1\rangle} = \mathbf{Y}_{(m)}\omega_{(m)\langle k+1\rangle}f_{(m)\langle k+1\rangle} \qquad f_{(m)\langle k+1\rangle} = \frac{1}{\omega'_{(m)\langle k+1\rangle}\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)}\omega_{(m)\langle k+1\rangle}}$$

Stabilitätstest $$\sum_{m=1}^{M}\|\eta_{(m)\langle k+1\rangle} - \eta_{(m)\langle k\rangle}\|^2 < \epsilon \quad ?$$

(* Der Index +k, ist ein Iterationszähler *)

Ebenso wie im ALS-Algorithmus in der Homogenitätsanalyse werden zunächst in Schritt 0 beliebige Startgewichte benutzt und mit deren Hilfe latente Variable berechnet.
Anschließend werden die Zusammenhänge zwischen den miteinander verbundenen latenten Variablen bewertet. In Schritt 1 sind auch andere Koeffizienten üblich, wie z. B. Korrelationskoeffizienten zwischen den LV u. a.. Es soll hier aber keine Diskussion dieses Basisalgorithmus und seiner Modifikationen erfolgen. Wir betrachten hier die spezielle (einfachste) Form, mit Vorzeichengewichtung (sign(...)).
Im Schritt 2 werden die als Umgebungsvariablen bezeichneten $\eta^*_{(m)}$ berechnet, die Informationen über den Zusammenhang zwischen $\mathbf{0}_{(m)}$ und den verbundenen MV-Blöcken tragen.
In einem 3. Schritt werden neue Gewichtskoeffizienten $\mathbf{T}_{(m)}$ geschätzt. Es wird aus Gründen der Anschaulichkeit nur der Modus B kurz beschrieben[11].
Der Modus B zur Bestimmung der Gewichtsvektoren $\mathbf{T}_{(m)}$ im +k+1,-ten Iterationsschritt ist angegeben als:

---

[11] Ausführliche Informationen zum PLS-Iterationsalgorithmus finden sich u. a. in Lohmöller (1989) und Mathes (1993a).

$$\eta^{*}_{(m)\langle k\rangle} = \mathbf{Y}_{(m)}\omega_{(m)\langle k+1\rangle} + \mathbf{u}_{(m)} \qquad m=1, ..., M \qquad (5.1)$$

und im Rahmen einer OLS-Schätzung[12] für die einzelnen Blöcke ergeben sich folgende Lösungen:

$$\omega_{(m)\langle k+1\rangle} = (\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)})^{-1}\mathbf{Y}'_{(m)}\eta^{*}_{(m)\langle k\rangle}. \qquad (5.2)$$

Nutzt man zunächst die spezielle Gestalt der $\eta^{*}_{(m)}$, so ergibt sich:

$$\begin{aligned}
\omega_{(m)\langle k+1\rangle} &= (\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)})^{-1}\mathbf{Y}'_{(m)}\eta^{*}_{(m)} \\
&= (\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)})^{-1}\mathbf{Y}'_{(m)}\sum_{m'\in C_m}\eta_{(m')}\rho_{(mm')\langle k\rangle} \\
&= (\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)})^{-1}\sum_{m'\in C_m}\mathbf{Y}'_{(m)}\eta_{(m')}\rho_{(mm')\langle k\rangle} \\
&= (\mathbf{Y}'_{(m)}\mathbf{Y}_{(m)})^{-1}\sum_{m'\in C_m}\mathbf{Y}'_{(m)}\mathbf{Y}_{(m')}\omega_{(m')\langle k\rangle}\rho_{(mm')\langle k\rangle}
\end{aligned} \qquad (5.3)$$

[12] OLS - steht für **O**rdinary-**L**east-**S**quares, also Kleinste-Quadrat-Methode

Fassen wir anschließend die $\mathbf{T}_{(m)}$ zu einem Gesamtgewichtsvektor $\mathbf{T}$ zusammen:

$$\omega = \begin{pmatrix} \omega_{(1)} \\ \omega_{(2)} \\ \vdots \\ \omega_{(M)} \end{pmatrix}$$

so läßt sich (s. Anhang) eine simultane Lösung für alle MV-Blöcke angeben:

$$\omega_{\langle k+1 \rangle} = (\mathbf{D}_{\mathbf{Y}'\mathbf{Y}})^{-1}((\mathbf{Y}'\mathbf{Y}) * \mathbf{P})\omega_{\langle k \rangle}. \tag{5.4}$$

$\mathbf{C}$ ist dabei die Matrix der aus Schritt 1 im Basisalgorithmus resultierenden Vorzeichengewichtungen zwischen den einzelnen Blöcken und mit " $*$ " sei die elementweise Multiplikation von Matrizen bezeichnet.
$\mathbf{C}$ ist dabei dort mit Nullen belegt, wo im Pfadmodell zwischen den einzelnen Blöcken keine Beziehungen bestehen. Ohne Beweis sei zusätzlich angemerkt, daß geeignete Umskalierungen (Multiplikation mit "-1") der y-Variablen dazu führen, daß alle Vorzeichen in Schritt 1 positiv sind[13].
$\mathbf{C}$ erklärt sich somit als eine Matrix mit "Nullen" und "Einsen", die Teilen von $\mathbf{Y}'\mathbf{Y}$ "Null" zuweist, an denen im Pfadmodell keine Verbindung besteht[14].

In Marketing-Beispiel der Abb. 1.1 mit geeigneter Umskalierung ergibt sich:

$$\mathbf{Y}'\mathbf{Y} * \mathbf{P} = \begin{pmatrix} 0 & \mathbf{Y}_{(1)}'\mathbf{Y}_{(2)} & \mathbf{Y}_{(1)}'\mathbf{Y}_{(3)} \\ \mathbf{Y}_{(2)}'\mathbf{Y}_{(1)} & 0 & \mathbf{Y}_{(2)}'\mathbf{Y}_{(3)} \\ \mathbf{Y}_{(3)}'\mathbf{Y}_{(1)} & \mathbf{Y}_{(3)}'\mathbf{Y}_{(2)} & 0 \end{pmatrix}.$$

Bezeichnen wir mit

$$\mathbf{Y}'\mathbf{Y}_{\mathbf{P}} := \mathbf{Y}'\mathbf{Y} * \mathbf{P},$$

so erhalten wir, im Fall der Konvergenz des Basis-Algorithmus, durch Grenzübergang folgende Gleichung:

---

[13] Durch diese Umskalierungen wird höchstens die Richtung des Zusammenhanges zu anderen Variablen, nicht aber seine Stärke geändert.

[14] $(\mathbf{Y}'\mathbf{Y}) * \mathbf{P}$ ergibt dann, bei der hier gewählten Vorzeichengewichtung, die Matrix $\mathbf{Y}'\mathbf{Y}$ für das unvollständige Design aus Kapitel 4.

$$\omega = (\mathbf{D}_{\mathbf{Y}}{}'_{\mathbf{Y}})^{-1}(\mathbf{Y}'\mathbf{Y}_{\mathbf{P}})\omega\,. \tag{5.5}$$

Das entspricht einem Eigenwertproblem für $\mathbf{T}$ bzgl. $(\mathbf{D}_{\mathbf{Y}}{}'_{\mathbf{Y}})^{-1}(\mathbf{Y}'\mathbf{Y}_{\mathbf{P}})$, deren Lösungen stationäre Punkte des Rayleigh Quotienten:

$$\frac{\omega'\mathbf{Y}'\mathbf{Y}_{\mathbf{P}}\omega}{\omega'\mathbf{D}_{\mathbf{Y}}{}'_{\mathbf{Y}}\omega}$$

sind (s. Heuser, 1986, S. 212).

Die in (3.9) vorgestellte relative Verlustfunktion, deren Lösung durch folgendes Maximumproblem gegeben war (vgl. (3.10)):

$$\tilde{\sigma}(\omega^*) := \underset{\omega}{\mathrm{Max}}\ \frac{\omega'\mathbf{Y}'\mathbf{Y}\omega}{\omega'\mathbf{D}_{\mathbf{Y}}{}'_{\mathbf{Y}}\omega}\,.$$

entspricht ebenfalls einem stationären Punkt in einem Rayleigh Quotienten mit der Matrix $\mathbf{Y}'\mathbf{Y}$ im Nenner. Insofern ist das PLS-Verfahren dem Anliegen der Homogenitätsanalyse verwandt.

Die Frage der Konvergenz des PLS-Basis-Algorithmus wird hier nicht weiter ausgeführt. Ihr ist ein späterer Beitrag gewidmet. Bisher ist das Konvergenzproblem noch nicht befriedigend gelöst. Alle bisher betrachteten praktischen PLS-Probleme scheinen Konvergenz aufzuzeigen. Über die obige Verwandtschaft zur Homogenitätsanalyse läßt sich der PLS-Algorithmus als ein Eigenwertproblem darstellen. Die Äquivalenz zum ALS-Algorithmus wie auch zu einem Eigenwertproblem wiederum legen Konvergenz nahe. Definitiv bewiesen ist sie bisher noch nicht. Jüngere Arbeiten von Glang (1988) oder Mathes (1993b) beweisen lediglich Optimierungseigenschaften bzw. Äquivalenz zu Eigenwertproblemen bei eintretender Konvergenz.

6. Kategoriale Daten im Pfadmodell

Ein Großteil der in wirtschaftswissenschaftlichen Untersuchungen anfallenden Daten sind kategorialer Natur, z. B. Branchenzugehörigkeit, regionale Kategorien, Produktpaletten u. v. a. m. Mit der zunehmenden Verarbeitung verhaltenswissenschaftlicher Aspekte, insbesondere in Gebieten wie Marketing oder Personalwesen, gelangen auch häufiger soziale und Verhaltenskategorien in wirtschaftswissenschaftlich relevante Fragestellungen.
Die Behandlung kategorialer Daten in Pfadmodellen ist dagegen nicht neu. Schon relativ früh wurden sowohl in LISREL- wie auch in PLS-Modellen kategoriale Variablen behandelt. Während LISREL-Modelle von sogenannten Schwellenwertmodellen ausgehen (z. B. Kukuk, 1991; Kühnel, 1994) werden in den bekannten PLS-Umsetzungen (s. u. a. Bertholet/Wold, 1984; Lohmöller, 1989) die Daten als empirische Häufigkeitsverteilungen in den einzelnen Kategorien abgebildet.

Eine allgemeinere Vorgehensweise, wie sie u. a. von Gifi vorgeschlagen wird, ist die Möglichkeit spezieller nichtlinearer Transformationen in der Verlustfunktion (3.4).
Zunächst ist die Verlustfunktion (3.4) für kategoriale Ausgangsdaten nicht geeignet, da Differenzen

$$(\eta - \mathbf{y}_j\omega_j)$$

in diesem Fall wenig Aussagewert besitzen.
Um dennoch diese Daten in der Homogenitätsanalyse behandeln zu können, werden einerseits die Variablen $\mathbf{y}_j$ (j=1, ..., P) in sogenannte Indikatormatrizen überführt und andererseits nichtlineare Transformationen in die Verlustfunktion aufgenommen.

i) Überführung der Ausgangsdaten in eine Indikatormatrix.

Eine Indikatormatrix ist eine Matrix, die lediglich aus Nullen und Einsen besteht. Dabei soll die Eins das Vorhandensein, die Null das Nichtvorhandensein eines bestimmten Wertes an einer bestimmten Stelle ausdrücken. Die zu einer kategorialen Variablen $\mathbf{y}_j$ gehörende Indikatormatrix $\mathbf{G}_j$ hat ebensoviele Zeilen wie Objekte für $\mathbf{y}_j$ gegeben sind und so viele Spalten wie unterschiedliche Kategorien für das Merkmal $\mathbf{y}_j$ möglich sind[15]. Die einzelnen Elemente in $\mathbf{G}_j$ werden wie folgt gebildet:

$$g_{ij_k} = \begin{pmatrix} 1 & \text{falls } y_{ij} \text{ die k-te Kategorie ausweist} \\ 0 & \text{sonst} \end{pmatrix}, \; i=1, ..., N \; ; \; k=1, ..., K_j,$$

d. h., in jeder Zeile von $\mathbf{G}_j$ steht genau eine “1" (falls keine Missing-Werte in den Daten auftreten), die restlichen Spalten der Zeile sind alle “0" und die “1" steht genau in der Spalte, die der Ausprägung des zugehörigen y_j-Objektes entspricht.

Für die Variable der Umsatzkategorien (s. Kapitel 9) sind die ersten 10 Objekte aus unserem Beispiel und die zugehörigen Zeilen der Indikatormatrix wie folgt anzugeben:

i	y_j	g_{ij_1}	g_{ij_2}	g_{ij_3}
1	KB	1	0	0
2	GB	0	0	1
3	GB	0	0	1
4	GB	0	0	1
5	KB	1	0	0
6	GB	0	0	1
7	MB	0	1	0
8	MB	0	1	0
9	KB	1	0	0
10	MB	0	1	0
!	!	!	!	!

[15] K_j bezeichnet im folgenden die Anzahl unterschiedlicher Kategorien des Merkmals y_j

Für diese Indikatormatrix ist die Kodierung der einzelnen Kategorienausprägung nicht mehr von Bedeutung und man kann die Elemente der Matrix als Häufigkeiten auffassen, mit der ein Objekt eine bestimmte Kategorienausprägung trägt. Nähere Erläuterungen und Interpretationen finden sich in den beiden folgenden Kapiteln.
Durch Zusammenfügung der $\mathbf{G}_j$ zu einer Gesamtmatrix

$$\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_P)$$

läßt sich dann einer Datenmatrix

$$\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_P)$$

eine Indikatormatrix $\mathbf{G}$ zuweisen.

ii) Nichtlineare Transformationen in der Verlustfunktion.

Die ursprüngliche Verlustfunktion in der Homogenitätsanalyse lautet:

$$\sigma(\eta,\omega) = \sum_{j=1}^{P} \|\eta - \mathbf{y}_j\omega_j\|^2,$$

d. h. zu den Ausgangsvariablen $\mathbf{y}_j$ sind geeignete lineare Transformationen zugelassen, mit dem Ziel, diese Verlustfunktion zu minimieren.

Betrachtet man dort nicht nur lineare Gewichte $\mathbf{T}_j$, sondern beliebige Transformationen $\mathbf{N}(\mathbf{y}_j)$ der Ausgangsdaten, ergibt sich:

$$\sigma(\eta,\phi) = \sum_{j=1}^{P} \|\eta - \phi(\mathbf{y}_j)\|_2.$$

Eine Möglichkeit solcher Transformation wäre die Einführung von Indikatormatrizen $\mathbf{G}_j$ für die kategorialen Variablen $\mathbf{y}_j$ (j=1, ... P). Mit entsprechenden Gewichtskoeffizienten $\mathbf{<}_j$ versehen läßt sich dann $\mathbf{N}$ definieren als:

$$\phi(\mathbf{y}_j) := \mathbf{G}_j\mathbf{n}\mathbf{y}_j = \mathbf{G}(\mathbf{y}_j)\mathbf{n}\mathbf{y}_j$$

und wir erhalten als Verlustfunktion

$$\sigma(\eta,\phi) = \sum_{j=1}^{p} \|\eta - \mathbf{G}_j\mathbf{n}\mathbf{y}_j\|_2. \qquad (6.1)$$

Die Gewichte $\mathbf{<}_j$ sind dann allerdings Vektoren der Länge K_j und nicht wie im Falle der $\mathbf{T}_j$ skalare Faktoren.

Ohne an dieser Stelle weiter ins Detail zu gehen (s. dazu, insbesondere die Interpretation der Vektoren $\mathbf{G_j n y_j}$, Kapitel 7 und 8) haben wir jetzt eine verwandte Situation zur bisher betrachteten Analyse mit metrischen Daten, wenn wir $\mathbf{Y}_j$ durch $\mathbf{G}_j$ und $\mathbf{T}_j$ durch $\mathbf{<}_j$ ersetzen[16].

7. Kurzer Abriß zur Korrespondenzanalyse

Für das Anliegen der Korrespondenzanalyse gibt es zwei unterschiedliche Auffassungsweisen. Während die einen Autoren sie als Methode zur graphischen Analyse von Häufigkeitsdaten (oder allgemeiner von nichtnegativen Daten) ansehen, *"The primary goal of correspondence analysis is to transform a table of numerical information into a graphic display, facilitating the interpretation of this information"* (Greenacre, 1994, S. 1), wird sie von anderen Autoren als nichtlineare Verallgemeinerung der Homogenitätsanalyse betrachtet, *"We rather adopt another definition of a model* (gemeint ist hier u. a. ein Modell für die Korrespondenzanalyse, Anm. d. A.)*, namely that a model is a nonlinear projection of the data on a (usually low-dimensional) parameter space."* (van der Heijden u. a., 1994, S. 79).
Beide Interpretationen sind durchaus nicht konträr und sie sind für die Einarbeitung in ein Pfadmodell von Bedeutung. Zunächst wird deshalb kurz die graphische Umsetzung der Korrespondenzanalyse vorgestellt und anschließend ihre Beziehung zur Homogenitätsanalyse erläutert.
Betrachten wir aus dem Beispiel in Kapitel 1 den MV-Block der Unternehmenscharakteristika. Zunächst werden daraus nur die beiden Variablen Anzahl Beschäftigte (Betriebsgröße) und Umsatz einer näheren Untersuchung unterzogen. Beide Variable haben jeweils drei Ausprägungen (s. Kapitel 9), so daß wir sie als kategoriale Variable behandeln können. Sie sind in folgenden Häufigkeiten vertreten:

Tabelle 7.1: Häufigkeiten der Variablen Beschäftigte und Umsatz

Beschäftigte	absolute Häufigkeit	relative Häufigkeit (%)	**Umsatz**	absolute Häufigkeit	relative Häufigkeit (%)
Kleinbetriebe (bis 50 B.)	147	39,6	unter 5 Mio. DM	124	33,4
Mittelbetriebe (51-500 B.)	95	25,6	5 - 100 Mio. DM	98	26,4
Großbetriebe (über 500 B.)	129	34,8	über 100 Mio. DM	149	40,2
Gesamt:	371	100	Gesamt:	371	100

Im weiteren Verlauf dieses Kapitels bezeichnen Häufigkeiten immer relative Häufigkeiten. Unter der Annahme der Gleichverteilung der Variablen, würden sich in jeder Variable und jeder

---

[16] Um eine einheitliche Schreibweise zu gewährleisten, werden in den folgenden Kapiteln die Gewichte bzw. Gewichtsvektoren durchgängig mit "$\mathbf{T}$" bezeichnet.

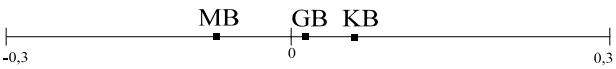
Ausprägung 0,33 als relative Häufigkeiten ergeben. Zieht man diese von den tatsächlichen Häufigkeiten ab und bezeichnet das Ergebnis als G-Residuen (unter der Annahme der Gleichverteilung der Kategorieausprägungen innerhalb der einzelnen Variablen), so ergibt sich folgendes Bild:

Tabelle 7.2: Residuen gegenüber Gleichverteilungannahme

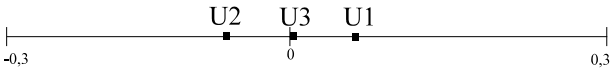
G-Residuen:	Beschäftigte r_i	Umsatz c_j
Ausprägung 1	0,063	0,001
Ausprägung 2	-0,078	-0,069
Ausprägung 3	0,015	0,069

(Aus später ersichtlichen Gründen werden die Häufigkeiten der Beschäftigten mit r_i und die der Umsätze mit c_j bezeichnet. Die Indizes i und j laufen dabei in unserem Beispiel von 1 bis 3, der Anzahl der unterschiedlichen Kategorieausprägungen. r steht dabei für *row* und c für *column*; in der zu behandelnden Kontingenztafel steht die Beschäftigtenzahl als Zeilenvariable und die Umsätze als Spaltenvariable.)
Diese G-Residuen lassen sich interpretieren als Abweichung von der Gleichverteilung. Die Abweichungen selbst können theoretisch jeden Wert zwischen -1 und +1 annehmen, wobei jeweils 2 Werte immer den dritten Wert bestimmen, da sich die Summe der Abweichungen zu 0 addiert. Somit lassen sich diese Abweichungen als stetige Größen interpretieren und auf einem Zahlenstrahl mit Nullpunkt plazieren. Links vom Nullpunkt erscheinende Werte entsprechen dann (im Sinne einer Gleichverteilung) unterrepräsentierten Werten, rechts vom Nullpunkt erscheinende Werte sind überrepräsentiert.

Residuenskala der Beschäftigten



Residuenskala der Umsatzkategorien



(KB = Kleinbetriebe / MB = Mittelbetriebe / GB = Großbetriebe)

(Umsätze: U1 = unter 5 Mio. DM / U2 = 5 bis 100 Mio. DM / U3 = über 100 Mio. DM)

Damit ist noch nicht viel gewonnen. Zu sehen ist lediglich, daß mehr Kleinbetriebe als die anderen Betriebsgrößenarten vorhanden sind. Andererseits ist ersichtlichtlich, daß ein höherer Anteil an Betrieben mit über 100 Mio. DM Umsatz vorhanden ist. Daraus allerdings den Schluß zu ziehen, daß Kleinbetriebe eher die großen Umsätze machen wiederspräche jeder Theorie und ist, wie wir gleich sehen werden auch voreilig.
Um die Beziehungen zwischen den beiden Variablen darstellen zu können bedarf es einer anderen Überlegung.
Betrachtet man zunächst die bedingten Häufigkeiten der Beschäftigtenkategorien unter der Bedingung Umsatz, ergibt sich folgendes Bild:

Tabelle 7.3: Spaltenprofile

$\mathbf{Q}^{(C)}$-Häufigkeiten*:	Beschäftigte unter Umsatz=1	Beschäftigte unter Umsatz=2	Beschäftigte unter Umsatz=3
Beschäftigte=1	0,952	0,275	0,013
Beschäftigte=2	0,048	0,684	0,148
Beschäftigte=3	0	0,041	0,839
Summe**:	1,0	1,0	1,0

(* Mit **Q** werden wir allgemein Tabellen oder Matrizen bezeichnen, die Häufigkeiten enthalten. Der Index C bezeugt in dem Falle, daß es sich um "Spaltenhäufigkeiten" handelt, also bedingten Häufigkeiten, wobei die Spalten der Tabelle/Matrix die Bedingung stellen. Die Spaltenhäufigkeiten werden im Kontext der Korrespondenzanalyse i. a. als "Spaltenprofile" benannt.
** Rundungsfehler werden in dieser und den folgenden Tabellen nicht gesondert ausgewiesen.)

Aus dieser Tabelle erhalten wir die erste echte Information zum Zusammenhang zwischen Beschäftigten und Umsatz. Würde man Unabhängigkeit der Zeilen unterstellen, müßten die Häufigkeiten in jeder Spalte die gleiche Verteilung aufweisen wie im Fall der oben betrachteten Randhäufigkeiten der Beschäftigtenkategorien. Ziehen wir diese "Randhäufigkeiten" r_i jetzt in den einzelnen Spalten ab

$$U_{ij} = Q_{ij}^{(C)} - r_i,$$

so müßten sich, im Fall der Unabhängigkeit, in den Zellen jeweils Nullen ergeben:

Tabelle 7.4: Residuen der Spaltenprofile

U-Residuen*:	Beschäftigte unter Umsatz=1	Beschäftigte unter Umsatz=2	Beschäftigte unter Umsatz=3	$\tilde{r}_i^{(C)}$	$r_i^{(C)}$
Beschäftigte=1	0,556	-0,121	-0,383	1,185	0,043
Beschäftigte=2	-0,208	0,428	-0,108	0,928	-0,241
Beschäftigte=3	-0,348	-0,307	0,491	1,313	0,171
Summe:	0	0	0	3,426	

(* Residuen gegenüber der Annahme der Unabhängigkeit zwischen den Spalten.)

Das ist hier ganz offensichtlich nicht der Fall. In der Spalte Umsatz=1 gibt es "zu viele" Kleinbetriebe, in der Spalte Umsatz=2 zu viele Mittelbetriebe und in der Spalte Umsatz=3 zu viele Großebtriebe, während die jeweils anderen Betriebsgrößenklassen stark unterrepräsentiert sind. Das stimmt, für unser einfaches Beispiel, sehr stark mit der Vorstellung überein, daß Kleinbetriebe natürlicherweise im allgemeinen weniger Umsatz zu verzeichnen haben, als Mittel- und Großbetriebe usw. Es gibt Ausnahmen, in der Studie erzielten immerhin 2 Unternehmen mit höchstens 50 Beschäftigten einen Umsatz von über 100 Mio. DM, aber der allgemeine Zusammenhang ist klar.

Wie lassen sich jetzt aber die Spaltenprofile mit der Randhäufigkeit r bzw. dem dortigen Spaltenresiduenvektor vergleichen?
Unter der Annahme der Gleichverteilung müßten, wie bereits erwähnt, die Komponenten der einzelnen Spaltenprofile alle Null sein. Demzufolge auch die Summe der einzelnen Komponenten. Diese kann aber auch Null sein, wenn keine Gleichverteilung vorliegt und sich Abweichungen in den Spalten gegenseitig kompensieren. Deshalb werden zunächst die Quadrate der einzelnen Vektorkomponenten aufsummiert (ein "Mittelwertvektor" funktioniert nicht, da er, wie auch die Summe, Nullkomponenten enthalten kann, die sich aus positiven und negativen Abweichungen ergeben):

$$\tilde{r}_i^{(C)} := \sum_{j=1}^{3} (Q_{ij}^{(C)} - r_i)^2 .$$

Dieser Vektor wird noch etwas verändert, um die Gegebenheiten in der Randverteilung in der Stichprobe miteinfließen zu lassen. Die einzelnen Komponenten von $\tilde{r}_i^{(C)}$ werden mit dem entsprechenden Wert r_i der Randverteilung relativiert:

$$\tilde{r}_i^{(C)} := \sum_{j=1}^{3} \frac{(Q_{ij}^{(C)} - r_i)^2}{r_i} .$$

Abweichungen einer Kategorie vom Unabhängigkeitsfall werden somit höher gewichtet, falls die Kategorie insgesamt einen niedrigen Anteil hat und umgekehrt[17]. Verschiebt man dann den so erhaltenen Vektor noch in der Weise, daß die neuen Komponenten um den Nullpunkt variieren:

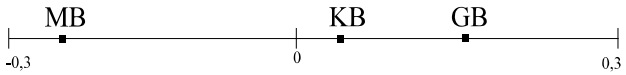
$$r_i^{(C)} := \tilde{r}_i^{(C)} - \overline{\tilde{r}_i^{(C)}},$$

dann ergibt sich ein Vektor, der mit den Spaltenresiduen von r_i vergleichbar ist und zusätzlich die Verhältnisse in Abhängigkeit von der Umsatzausprägung berücksichtigt:

---

[17] Dies ist der Grund, warum sich, wie in der Literatur angeführt, die Methode des *reciprocal averaging* als ein zur Korrespondenzanalyse äquivalentes Verfahren ergibt. Differenzen von 999-1000=-1 und 9-10=-1 haben relativ betrachtet eine völlig unterschiedliche Bedeutung.

Residuenskala der Spaltenprofile



Wie im Falle des Spaltenresiduenvektors sind die Mittelbetriebe unterrepräsentiert, während Klein- und Großbetriebe überrepräsentiert sind. Während aber bei den Randhäufigkeiten noch die Kleinbetriebe die höchste Ausprägung haben (0,063; dies entspricht der höchsten Randhäufigkeit von 0,396) sieht man hier, daß jetzt die Großbetriebe den höchsten Stellenwert besitzen. Dies ist von Bedeutung für die weitere Untersuchung und führt uns von dem oben angesprochenen Trugschluß fort, daß Kleinbetriebe mit hohen Umsätzen verbunden sind. Offensichtlich sind in diese Betrachtung bereits Abhängigkeiten innerhalb der Kontingenztafel mit eingeflossen.

Ein ganz ähnliches Bild ergibt sich für die bedingten Häufigkeiten der Variable Umsatz:

Tabelle 7.5: Zeilenprofile

$\mathbf{Q}^{(R)}$-Häufigkeiten*:	Umsatz=1	Umsatz=2	Umsatz=3	Summe:
Umsatz unter Beschäftigte=1	0,803	0,184	0,014	1,0
Umsatz unter Beschäftigte=2	0,063	0,705	0,232	1,0
Umsatz unter Beschäftigte=3	0	0,031	0,969	1,0

(*s. Bem. in Tabelle 7.3 der Spaltenprofile oben; die Zeilen werden hier als “Zeilenprofile” angesprochen)

Wieder unter der Prämisse der Unabhängigkeit müßten sich, nach Abzug der Randhäufigkeiten Nullen in den einzelnen Zellen der obigen Tabelle ergeben:

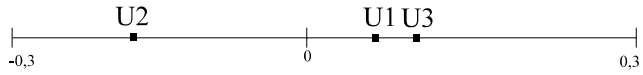
$$\mathbf{U}_{ij} = \mathbf{Q}_{ij}^{(R)} - \mathbf{c}_j .$$

Tabelle 7.6: Residuen der Zeilenprofile

U-Residuen*:	Umsatz=1	Umsatz=2	Umsatz=3	Summe:
Umsatz unter Beschäftigte=1	0,469	-0,080	-0,388	0
Umsatz unter Beschäftigte=2	-0,271	0,441	-0,170	0
Umsatz unter Beschäftigte=3	-0,334	-0,233	0,567	0
$\tilde{c}_j^{(R)} := \sum_{i=1}^{3} \frac{(Q_{ij}^{(R)} - c_j)^2}{c_j}$	1,211	0,967	1,248	3,426
$c_j^{(R)} := \tilde{c}_j^{(R)} - \overline{\tilde{c}_j^{(R)}}$	0,069	-0,175	0,106	

(* Residuen gegenüber der Annahme der Unabhängigkeit zwischen den Zeilen.)

Residuenskala der Zeilenprofile



Die Residuenskala kann analog zu der der Spaltenprofile interpretiert werden.

Mit diesem Vorwissen läßt sich jetzt die Frage angehen, ob vielleicht für beide Variable eine gemeinsame Skala gefunden werden kann, auf denen Spalten- und Zeilenprofile verglichen werden können und die gleichzeitig Ausdruck über die Zusammenhänge liefert.
Dazu wird die Tabelle der unbedingten Häufigkeiten betrachtet, die man, in Analogie zur obigen Bezeichnung Spalten- und Zeilenprofile, als Gesamtprofil bezeichnen könnte:

Tabelle 7.7: Gesamtprofil

Q-Häufigkeiten*:	Umsatz=1	Umsatz=2	Umsatz=3	Summe (r_i):
Beschäftigte=1	0,318	0,073	0,005	0,396
Beschäftigte=2	0,016	0,181	0,059	0,256
Beschäftigte=3	0	0,011	0,337	0,348
Summe (c_j):	0,334	0,265	0,401	1

(*s. Bem. in der Tabelle 7.3 der Spaltenprofile)

Als “Ränder” ergeben sich die bereits oben angegebenen Randhäufigkeiten r_i und c_j.
Dann läßt sich wiederum die Matrix der Abweichungen von der Unabhängigkeit angeben:

$$U_{ij} = Q_{ij} - r_i \cdot c_j .$$

Tabelle 7.8: Residuen der Gesamtprofile

U-Residuen*:	Umsatz=1	Umsatz=2	Umsatz=3	$\tilde{r}_i^{(G)}$		$r_i^{(G)}$
Beschäftigte=1	0,186	-0,032	-0,154	0,418		0,036
Beschäftigte=2	-0,069	0,113	-0,044	0,263		-0,119
Beschäftigte=3	-0,116	-0,081	0,197	0,466		0,084
$\tilde{c}_j^{(G)}$	0,433	0,27	0,446	MW:	0,383	
				0,383		
$c_j^{(G)}$	0,05	-0,113	0,063			

(* Residuen gegenüber der Annahme der Unabhängigkeit Spalten und Zeilen.)

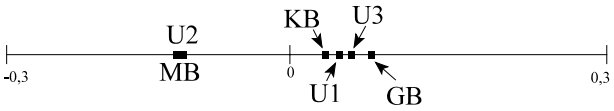
Ebenso wie bei den Spalten- und Zeilenprofilen ergeben sich die Abweichungsskalen $r_i^{(G)}$ und $c_j^{(G)}$ als gewichtete Summe der jeweiligen quadrierten Abweichungen, zentriert um ihren Mittelwert:

$$\tilde{r}_i^{(G)} := \sum_{j=1}^{3} \frac{(Q_{ij} - r_i c_j)^2}{r_i c_j} \qquad r_i^{(G)} := \tilde{r}_i^{(G)} - \overline{\tilde{r}_i^{(G)}}$$

und analog:

$$\tilde{c}_j^{(G)} := \sum_{i=1}^{3} \frac{(Q_{ij} - r_i c_j)^2}{r_i c_j} \qquad c_j^{(G)} := \tilde{c}_j^{(G)} - \overline{\tilde{c}_j^{(G)}} .$$

Residuenskalen der Gesamtprofile



Dies sind jetzt zwei Skalen, die dieselbe Gewichtung in den einzelnen Komponenten erfahren, und wir können Sie miteinander vergleichen.

In der obigen Abbildung läßt sich jetzt der Zusammenhang der Beschäftigtenkategorien mit den Umsatzkategorien erkennen. Erkennbar ist, daß die Größenklassen der Beschäftigten zu den entsprechenden Umsatzklassen am besten “passen”.

Es muß hier allerdings bemerkt werden, daß für eine genauerer Analyse der Bilder von Korrespondenzanalysen sehr vorsichtig vorgegangen werden muß.

Die soeben skizzierte Vorgehensweise entspricht dem Vorgehen in der Korrespondenzanalyse mit zwei kategorialen Variablen. Die Ergebnisse sind, bis auf gewisse Normierungsbedingungen, auf die weiter unten kurz eingegangen werden soll, äquivalent.
Eine Beziehung zur Homogenitätsanalyse ergibt sich, wenn nicht die Kontingenztafel der zwei Variablen analysiert wird, sondern die Matrix der Ausgangsvariablen in eine Indikatormatrix transformiert und diese im Sinne einer Hauptkomponentenanalyse betrachtet wird. Die Vorgehensweise wurde im Kapitel 6 erläutert. Es soll jetzt anschließend die mathematische Umsetzung gezeigt werden, deren Eigenschaften wesentlich für die Anwendung der Korrespondenzanalyse im Pfadmodell sind. Es wird dabei von P Variablen y_j ausgegangen, der Fall für P=2, wie er oben am Beispiel für Umsatz und Beschäftigte betrachtet wurde ist ein Spezialfall.

Wir bezeichnen mit[18]

$\underset{[N,P]}{\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_P})}$ die Datenmatrix,

k_j die Anzahl der Merkmalsausprägungen der Variable yj (j=1, ..., P),

$\mathbf{K} := \sum_{j=1}^{P} \mathbf{k_j}$ die Summe aller Merkmalsausprägungen aller Variablen

und

$\underset{[N,K]}{\mathbf{G} = (\mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_P})}$ die zugehörige Indikatormatrix, wie Sie in Kapitel 6 definiert wurde.

Die Methodik der Korrespondenzanalyse bezieht sich, wie oben gezeigt, auf die relativen Häufigkeiten der einzelnen Merkmalskombinationen. Zu diesem Zweck wird die Matrix G in eine Matrix **Q** von relativen Häufigkeiten überführt. Diese Matrix werden wir im weiteren Korrespondenzmatrix nennen, sie definiert sich wie folgt:

$$\mathbf{Q} := \left(\frac{1}{\mathbf{P \cdot N}}\right)\mathbf{G}.$$

Die Bezeichnung der Komponenten von **Q** als relative Häufigkeiten rechtfertigt sich über die Interpretation der Matrix **G** als Matrix absoluter Häufigkeiten im Sinne einer Kontingenztafel. Dabei ist die "Spaltenvariable" das entsprechende Objekt selbst, die Zeilenvariable sind die verschiedenen Kategorien über alle Variablen. Jedes Objekt erscheint dann mit der Häufigkeit

---

[18] Die an einigen Stellen angegebenen Werte in eckigen Klammern, [...], geben die Dimensionen der darüberstehenden Matrix an. Sie sind hier zur Erleichterung für den Leser angebracht.

1 in je einer Kategorie einer jeden Variablen, während es in den anderen Kategorien mit der Häufigkeit 0 ausgestattet ist. Die Randhäufigkeiten für die Objekte beträgt dann gerade P, die Anzahl der Variablen. Die Randhäufigkeiten der Kategorien entsprechen den üblichen Häufigkeiten der jeweiligen Kategorie in dem Datensatz.
Durch die Gewichtung von **G** mit dem Faktor $\frac{1}{P \cdot N}$ wird durch die "Gesamthäufigkeit" dividiert und man erhält in den Zellen von **Q** relative Häufigkeiten. Die Summe über alle Elemente von **Q** ist dann gerade 1. Die Zeilensummen sind identisch 1/n und die Spaltensummen sind das $\frac{1}{P}$-fache der Randverteilungen der Kategorieausprägungen der jeweiligen Variablen.

Für die weiteren Berechnungen werden wir die Zeilen- bzw. Spaltensummen algebraisch angeben:

$$\underset{[N,1]}{\mathbf{r}} := \mathbf{Q}\underline{1} = (1/N)\mathbf{G}\underline{1} \qquad \text{Zeilensummen .}$$

($\underline{1}$ - bezeichnet einen Vektor mit lauter Einsen, die Dimension des Vektors ergibt sich aus dem Kontext.)

Die Zeilensummen sind also für jede Zeile gleich, was auf der Tatsache beruht, daß jedes Objekt genau einmal in der Häufigkeitsmatrix auftaucht.
Die Spaltensummen sind:

$$\underset{[K,1]}{\mathbf{c}} := \mathbf{Q}^{t}\underline{1} = (1/PN)\mathbf{G}^{t}\underline{1} \qquad \text{Spaltensummen.}$$

Wie wir später sehen werden, bilden die Zeilen- bzw. die Spaltensummen eine Art durchschnittliches Profil für die prozentuale Verteilung in den einzelnen Spalten bzw. Zeilen.
Zur Vereinfachung der Schreibweise werden wir zwei Hilfsmatrizen definieren. Dies sind zwei Diagonalmatrizen, welche auf der Hauptdiagonale die Zeilensummen r bzw. die Spaltensummen c enthalten und sonst mit 0 gefüllt sind:

$$\underset{[N,N]}{\mathbf{D}_r} := \mathrm{diag}(\mathbf{r}) \qquad \text{Diagonalmatrix der Zeilensummen}$$

$$\underset{[K,K]}{\mathbf{D}_c} := \mathrm{diag}(\mathbf{c}) \qquad \text{Diagonalmatrix der Spaltensummen.}$$

Mit Hilfe dieser Diagonalmatrizen konstruieren wir sogenannte Profilmatrizen. Die Profile sind der eigentliche Ausgangspunkt der Korrespondenzanalyse und ein Profil wird definiert als eine bedingte relative Häufigkeitsverteilung (s. auch Tab. 7.3 und 7.5).

Zunächst definieren wir Zeilenprofile als

$$\tilde{r}_i^{|} := \frac{1}{r_i} \cdot \left(Q_{i1}, Q_{i2}, \ldots, Q_{iK}\right) \text{ für i=1, ... N.}$$

D. h. als Zeilenprofil betrachten wir für jedes Objekt (Zeile) die relativen Häufigkeiten gewichtet mit dem Reziproken der entsprechenden Zeilensumme. Auf diese Weise sind die Zeilensummen die durch Objekt i bedingten relativen Häufigkeiten.

Analog definieren wir Spaltenprofile als die durch die Merkmalsausprägung (Spalte) k bedingte relative Häufigkeit:

$$\tilde{c}_k^{|} := \frac{1}{c_k} \cdot \left(Q_{1k}, Q_{2k}, \ldots, Q_{Nk}\right) \qquad \text{für k=1, ... K.}$$

Zeilen und Spaltenprofile werden in den Matrizen R und C geeignet zusammengefaßt:

$$\mathbf{R} = \begin{pmatrix} \tilde{\mathbf{r}}_1' \\ \tilde{\mathbf{r}}_2' \\ \vdots \\ \tilde{\mathbf{r}}_N' \end{pmatrix} = \mathbf{D}_r^{-1}\mathbf{Q}$$

[N,K]

und

$$\mathbf{C} = (\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \ldots, \tilde{\mathbf{c}}_K) = \mathbf{Q}\mathbf{D}_c^{-1}$$

[K,N]

Diese beiden Matrizen **R** und **C** spielen eine Schlüsselrolle bei der Adaption der Korrespondenzanalyse im PLS-Algorithmus zur Bestimmung der latenten Variablen. Sie dienen als Analogon zu der Verwendung der Datenmatrix **Y**und $\mathbf{Y}'$ im metrischen Fall.
In der Korrespondenzanalyse werden jetzt die Zeilenprofile als N Punkte im K-dimensionalen Raum der reellen Zahlen ($\mathbb{R}^K$) und die Spaltenprofile als K Punkte im N-dimensionalen Raum $\mathbb{R}^N$ betrachtet.
Die Fragestellung für die Korrespondenzanalyse lautet, ähnlich wie in der Hauptkomponentenanalyse, läßt sich ein q-dimensionaler Unterraum finden, so daß $q \ll \min\{N, K\}$ und daß auch die N Zeilenprofile bzw. K Spaltenprofile so dargestellt werden können, daß "möglichst wenig Informationsverlust" eintritt. Die Korrespondenzanalyse behandelt vorrangig den Fall q=2, um eine graphische Darstellung der Ergebnisse zu ermöglichen. Für die Verwendung im Pfadmodell wird lediglich die eindimensionale Lösung (q=1) benötigt.
Der Informationsverlust wird über eine χ^2-Metrik definiert, auf die hier nicht näher eingegangen werden soll (s. z. B. Lebart u. a., 1984; Greenacre, 1984). Für 2-dimensionale Kontigenztafeln, P=2, ergibt er sich als das übliche χ^2-Kontingenzmaß. Betrachtet man die im ersten Teil dieses Kapitels eher verbal erfolgte Ableitung der Korrespondenzanalyse für eine zweidimensionale Kontingenztafel, so erkennt man, daß die Summe der Komponenten der letzten Skalen $\tilde{\mathbf{r}}^{(G)}$ und $\tilde{\mathbf{c}}^{(G)}$ genau der Definition des χ^2-Kontingenzmaß enstpricht.
Ähnlich wie in der Hauptkomponentenanalyse wird versucht für die Zeilen- bzw. Spaltenprofile, als Punkten im $\mathbb{R}^K$ bzw. $\mathbb{R}^N$, einen gemeinsamen Unterraum zu komponieren. Zur Korrespondenzmatrix **Q** werden Hauptachsen berechnet. Genauer gesagt werden für die Matrix:

$$\tilde{\mathbf{Q}} = \mathbf{Q} - \mathbf{r}\mathbf{c}'$$

Hauptkomponenten gesucht. Die Matrix

$$\mathbf{r}\mathbf{c}'$$

entspricht dabei, wie im ersten Teil des Kapitels, der Häufigkeitsverteilung unter der Annahme, daß Zeilen und Spalten unabhängig sind. Die Elemente von $\mathbf{rc}^{\prime}$ werden in der Korrespondenzanalyse auch als Schwerpunkte bezeichnet. Durch die Operation $\mathbf{Q} - \mathbf{rc}^{\prime}$ zentrieren wir die Korrespondenzmatrix und betrachten für die Zeilen- und Spaltenprofile nur noch die Abweichungen von ihren Schwerpunkten (s. auch Tab. 7.8).

Die Vorgehensweise zur Berechnung der Hauptkomponenten für $\tilde{\mathbf{P}}$ wird hier nicht ausgeführt. Sie erfolgt, wie im metrischen Fall, über einen alternierenden Kleinst-Quadrat-Algorithmus (s. Kapitel 3) oder über Singulärwertzerlegung (s. z. B. Greenacre, 1984, S. 37 ff.).

Bedeutsam für die weitere Anwendung sind die Ergebnisse.

Ist $\mathbf{0}$ die erste Hauptkomponente zu $\tilde{\mathbf{Q}}$ und $\mathbf{T}$ der entsprechende Gewichtsvektor(s. Kapitel 6), so ergibt sich, unter Zuhilfenahme der Zeilen- und Spaltenprofilmatrizen, folgende nützliche Umrechnung:

$$\eta = \mathbf{R}\omega \cdot c_{\eta}$$

und (7.1)

$$\omega = \mathbf{C}\eta \cdot c_{\omega}.$$

Dies ist ein Gleichungssystem welches der Bildung der latenten Variablen im metrischen Fall adäquat ist. Die Faktoren c sind Skalare, die eine gewisse Normierung bedeuten, sie fallen im Zuge der späteren Verwendung im Pfadmodell weg und werden deshalb hier nicht weiter diskutiert. Der Gewichtsvektor $\mathbf{T}$ läßt sich in Gewichtsvektoren $\mathbf{T}_j$ (j=1, ..., P) für die einzelnen Kategorien der Variable y_j zerlegen:

$$\omega = (\omega_1, \omega_2, \ldots, \omega_P).$$

Für den Fall zweier Variabler, P=2, entsprechen die einzelnen Gewichtsvektoren den im ersten Teil dieses Kapitels betrachteten Skalen $\mathbf{r}^{(G)}$ und $\mathbf{c}^{(G)}$ für die Kategorien der beiden Variablen. In diesem Sinne lassen sich die Komponenten in $<$ als Gewichte für die einzelnen Kategorienausprägungen betrachten, die insbesondere auch die Zusammenhänge zwischen Kategorien unterschiedlicher y_j widerspiegeln. Ähnliche Gewichte belegen einen Zusammenhang zwischen den betreffenden Kategorien, stark unterschiedliche Gewichte einen Gegensatz.

Die bisher skizzierte Methodik der Korrespondenzanalyse bezog sich auf nur einen Block von P manifesten Variablen.

Unter Ausnutzung der Umrechnungsgleichungen (7.1) wird im folgenden Kapitel die Korrespondenzanalyse in den PLS-Algorithmus zur Berechnung latenter Variabler eingebaut.

8. Ein korrespondenzanalytischer Ansatz für das Pfadmodell

Für den Fall, daß die in das Pfadmodell eingehenden Blöcke manifester Variabler nicht mehr stetige, sondern kategoriale Daten enthalten, ist das Konzept des Meßgleichungssystems (s. Kapitel 2) nicht mehr in dieser Form anwendbar und damit muß auch der PLS-Basis-Algorithmus zur Berechnung der LV an diese Datensituation angepaßt werden.
Im vorigen Abschnitt wurde erläutert, wie mit Hilfe der Korrespondenzanalyse aus Blöcken manifester kategorialer Daten latente Variable bestimmt werden können, die erstens die bivariaten Zusammenhänge in den Blöcken widerspiegeln, zweitens in der Konstruktion den latenten Variablen aus der Homogenitätsanalyse entsprechen und drittens als stetige Variable behandelt werden können.
Diese Eigenschaften sollen ausgenutzt werden, um im folgenden einen Iterationsalgorithmus vorzustellen, der stetige LV für das Pfadmodell erzeugt.
Die einzelnen Iterationsschritte werden wie in Kapitel 5 mit Schritt 0 bis Schritt 4 bezeichnet und sind mit diesen auch vergleichbar.

Schritt 0: Als Ausgangsgewichte für die zu bildenden LV wird eine Korrespondenzanalyse innerhalb eines jeden MV-Blockes berechnet. Mit den dort ermittelten Gewichten $\omega_{(m)\langle 0\rangle}$ werden die LV als gewichtete Summe der "Ausgangsmatrix" R der Zeilenprofile (s. Kapitel 7) gebildet und normiert[19].

$$\eta_{(m)\langle 0\rangle} = \mathbf{R}\omega_{(m)\langle 0\rangle} \cdot f_{(m)\langle 0\rangle}$$

Schritt 1: Die aus Schritt 0 resultieredenden LV $\eta_{(m)\langle 0\rangle}$ werden jetzt als stetige Repräsentanten der Objektprofile angesehen. Die zwischen Ihnen bestehenden Zusammenhänge werden, wie in der metrischen PLS als Vorzeichenkoeffizienten der bivariaten Korrelationen der verschiedenen LV gebildet:

$$\rho_{(mm')\langle k\rangle} := \mathrm{sign}[\eta^{I}_{(m)\langle k\rangle}\eta_{(m')\langle k\rangle}] \qquad m=1, ..., M ; \; m' \in C_m .$$

Schritt 2: Jetzt werden ebenfalls Umgebungsvariable gebildet, die eine Linearkombination der betreffenden LV, gewichtet mit ihren Zusammenhangskoeffizienten darstellen:

$$\eta^{*}_{(m)\langle k\rangle} := \sum_{m' \in C_m} \rho_{(mm')\langle k\rangle}\eta_{(m')\langle k\rangle} .$$

Schritt 3: Im PLS-Basis-Algorithmus werden die Umgebungsvariable als diejenigen LV angesehen, die die Information der Beziehungen zu den benachbarten MV-Blöcken tragen. Analog dazu werden Sie hier als Objektprofile des m-ten Blockes

---

[19] s. Umrechnungsformeln (7.1) in Kapitel 7 / der Index +k, ist wiederum ein Iterationszähler und der Faktor $f_{(m)}$ ein Normierungsfaktor

betrachtet, die die Informationen der Nachbarblöcke tragen.
Die in Kapitel 7 hervorgehobenen Umrechnungsformeln (7.1) für Zeilen- in Spaltenprofile erlauben es nun, mit diesen Umgebungsvariablen (Objekt-/Zeilenprofilen) neue Kategoriengewichte (Spaltenprofile) zu berechnen:

$$\omega_{(m)_{\langle k+1\rangle}} := C\eta^{*}_{(m)_{\langle k\rangle}} \cdot c.$$

Genau wie im PLS-Basis-Algorithmus erhalten wir damit neue Gewichte, die sich aus den Beziehungen zu den Nachbarblöcken ergeben.

Schritt 4: Im abschließenden Iterationsschritt werden dann zu den unter Schritt 3 berechneten neuen Gewichten die neuen LV gebildet.

$$\eta_{(m)_{\langle k+1\rangle}} := R\omega_{(m)_{\langle k+1\rangle}} \cdot f_{(m)_{\langle k+1\rangle}}$$

Sollten sich die neuen LV von den alten nicht mehr wesentlich unterscheiden, wird der Iterationsalgorithmus abgebrochen, im anderen Fall wird ein neuer Iterationszyklus durchlaufen.

Ein mögliches Abbruchkriterium wäre:

$$\sum_{m=1}^{M} \| \eta_{(m)_{\langle k+1\rangle}} - \eta_{(m)_{\langle k\rangle}} \|^2 < \epsilon,$$

mit vorgegebenem kleinem ϵ.

Die sich ergebenden $\eta_{(m)}$ lassen sich dann, genau wie im zweidimensionalen Fall der Korrespondenzanalyse als Bewertung der einzelnen Kategorienkombinationen in den einzelnen Blöcken ansehen.
Die anschließende Berechnung der inneren Modellstruktur verläuft nun ähnlich wie im Fall der metrischen PLS-Methode.
Die latenten Variablen werden als stetige Repräsentanten der Häufigkeitsverteilungen der MV-Blöcke angesehen und ihre innere Zusammenhangsstruktur wird mit einem Regressionsmodell abgebildet, welches nach der Methode der kleinsten Quadrate gelöst wird.
Das äußere Modell entspricht der Definition von Gewichten aus den neuen latenten Variablen, gemäß dem oben beschriebenen Iterationsschritt Schritt 3.

Der Vorteil der PLS-Methode mit Hilfe einer Korrespondenzanalyse besteht darin, daß erstens, sofern das Modell nicht zu umfangreich ist, sich das innere Modell graphisch darstellen läßt und die Strukturen des Zusammenhangs verschiedener Kategoriekombinationen *ersichtlich* werden. Zweitens ist eine Klassifikation der Kategoriekombinationen in Abhängigkeit von Kategoriekombinationen anderer MV-Blöcke möglich.

9. Beispiel

Das bereits in der Einführung genannte Beispiel zu **"Wahrnehmung von und Anspruch an Standortfaktoren in Abhängigkeit von Unternehmenscharakteristika"** soll hier mit Hilfe der im vorigen Kapitel ausgeführten Theorie analysiert werden.

Die untersuchten Daten sind Teil eines empirischen Forschungsprojektes unter dem Thema "Entwicklung eines effizienten Marketingkonzeptes zur wirtschaftlichen Förderung Brandenburgs" (Balderjahn/Aleff, 1996), das am Lehrstuhl für Betriebswirtschaftslehre, Schwerpunkt Marketing durchgeführt wurde.
Im Rahmen dieses Forschungsvorhabens wurden "*Standortanforderungen seitens privater Unternehmungen identifiziert und in ihrer Bedeutung beurteilt...*" (Balderjahn/Aleff, 1996, S. 23). Dazu wurde eine Batterie von Standortanforderungen entwickelt und die Unternehmen aufgefordert auf einer Viererskala (sehr wichtig/wichtig/weniger wichtig/unwichtig) ihre jeweilige Einschätzung der Wichtigkeit dieser Standortanforderungen für ihr Unternehmen anzugeben. Dieselbe Anforderungsbatterie wurde verwendet, um die Unternehmen, ebenfalls auf einer Viererskala (sehr stark/stark/schwach/sehr schwach), nach ihrer Einschätzung der Ausprägung der entsprechenden Standortanforderung in Brandenburg zu befragen.
Die befragten Unternehmen wurden in *in Brandenburg ansässige Unternehmen (Investoren)* und in *nicht in Brandenburg ansässige Unternehmen (Nichtinvestoren)* unterteilt, um "*ein zielgruppenorientiertes Standortmarketing zu ermöglichen*" (Balderjahn/Aleff, 1996, S. 24) und auch nach unternehmensspezifischen Merkmalen, wie Zahl der Beschäftigten und Umsatz, befragt.

Das hier vorgestellte Modell beschränkt sich lediglich auf die in Brandenburg ansässigen Unternehmen (im weiteren Investoren genannt) und auf eine Teilauswahl der Standortfaktoren, um ein überschaubares und inhaltlich interpretierbares Modell zu erhalten.
Für die Investoren wurde eine Vollerhebung in Brandenburg angestrebt und aus den verfügbaren Quellen 857 Brandenburger Unternehmen identifiziert (s. Balderjahn/Aleff, 1996, S. 28 ff.). Es wurde die Form einer postalischen Befragung gewählt und eine Rücklaufquote von 47,6 % erzielt, so daß 408 auswertbare Fragebögen vorlagen.

Das hier vorgestellte PLS-Modell soll Argumentationen bezüglich folgender Annahmen liefern:

A1: Die Unternehmen lassen sich nach unternehmensspezifischen Merkmalen (Unternehmenscharakteristika) klassifizieren.

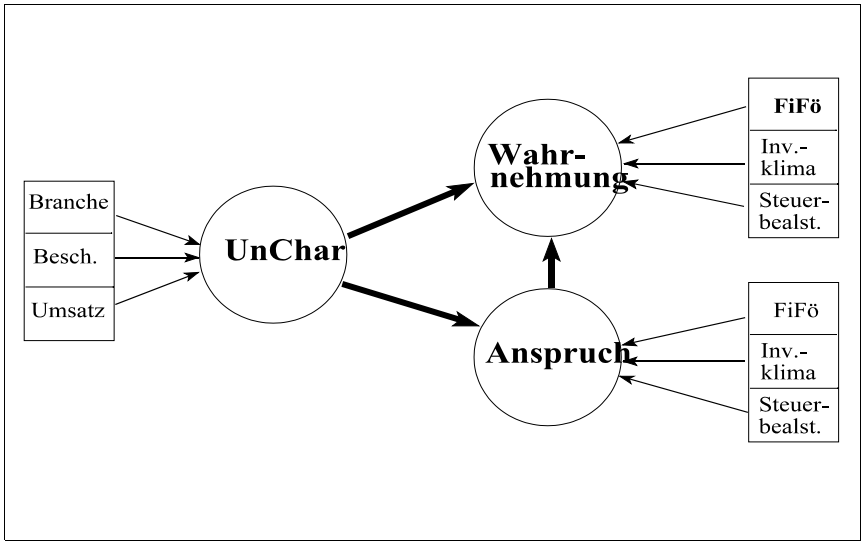
A2: Es existieren Gruppen von Standortfaktoren, welche inhaltlich verwandt sind und innerhalb der Ansprüche und Wahrnehmungen gemeinsame Sachverhalte messen.

A3: Die Unternehmenscharakteristika wirken sowohl auf die Ansprüche als auch auf die Wahrnehmungen von Standortfaktoren.

A4: Die Ansprüche an bestimmte Standortfaktoren beeinflussen die Wahrnehmung ihrer Ausprägung in Brandenburg.

Zur Abbildung dieser Annahmen wird das folgende Pfadmodell verwendet:

Abb. 9.1 Anspruch an und Wahrnehmung von Standortfaktoren in Abhängigkeit von Unternehmenscharakteristika



Zur Charakterisierung der Unternehmen sollen hier die Merkmale Branchenzugehörigkeit, Zahl der Beschäftigten im Unternehmen und der Jahresumsatz des Unternehmens dienen, die in folgender Ausprägung und Häufigkeit vorliegen:

Branchen-zugehörigkeit:	verarbeitendes Gewerbe	Baugewerbe	Handel/Banken/ Versicherungen	sonstige Dienstleistungen	gesamt:
abs. Häufigkeit	71	73	67	70	281
rel. Häufigkeit	25,30%	26,00%	23,80%	24,90%	100,00%

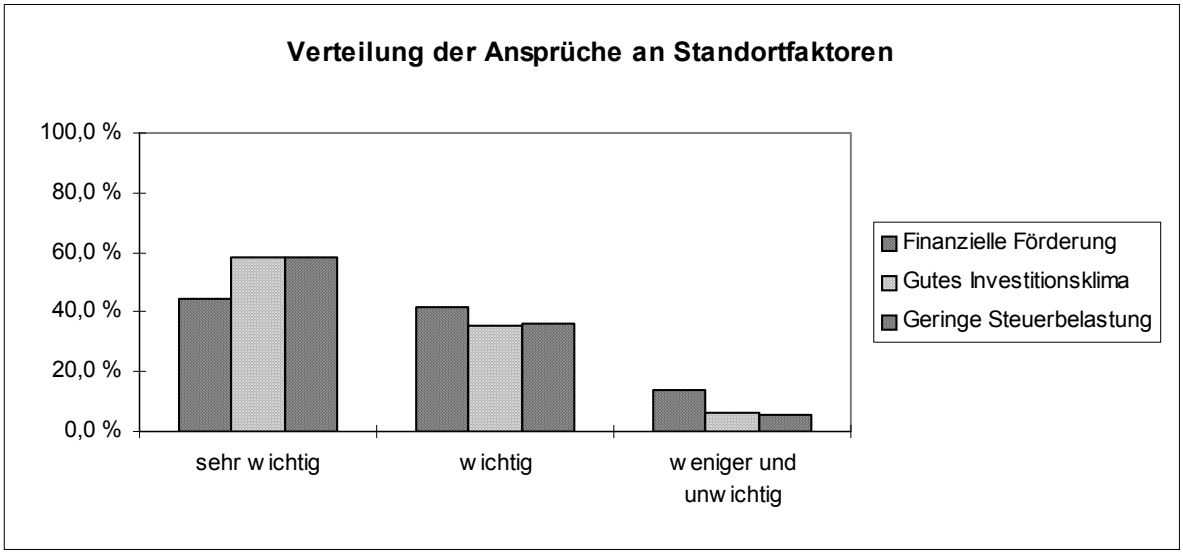
Beschäftigte im Unternehmen:	bis 50	51 - 500	über 500	gesamt:
abs. Häufigkeit	122	69	90	281
rel. Häufigkeit	43,40%	24,60%	32,00%	100,00%

Umsatz im Unternehmen:	unter 5 Mio. DM	5-100 Mio. DM	über 100 Mio. DM	gesamt:
abs. Häufigkeit	104	77	100	281
rel. Häufigkeit	37,00%	27,40%	35,60%	100,00%

Die Gesamtzahl von 281 Unternehmen in der obigen Darstellung ergab sich aufgrund teilweise fehlender Angaben in den Unternehmensdaten als auch den folgenden Standortfaktoren. Um durch eine Diskussion der Auswirkungen von Missing-Daten im PLS-Modell das Anliegen des Diskussionsbeitrages nicht zu überlagern wurden die Datensätze mit fehlenden Werten vorerst herausgenommen.

Von den im Fragebogen auftretenden 42 Standortfaktoren wurden drei herausgenommen, Finanzielle Förderung durch Land/Kommune, Gutes Investitionsklima und Geringe Steuerbelastung. Diese bilden sowohl eine inhaltliche Einheit und sind auch im Anspruch für die meisten Unternehmen relativ wichtig.
Unterschieden nach Ansprüchen und Wahrnehmungen ergibt sich dabei folgendes Bild:

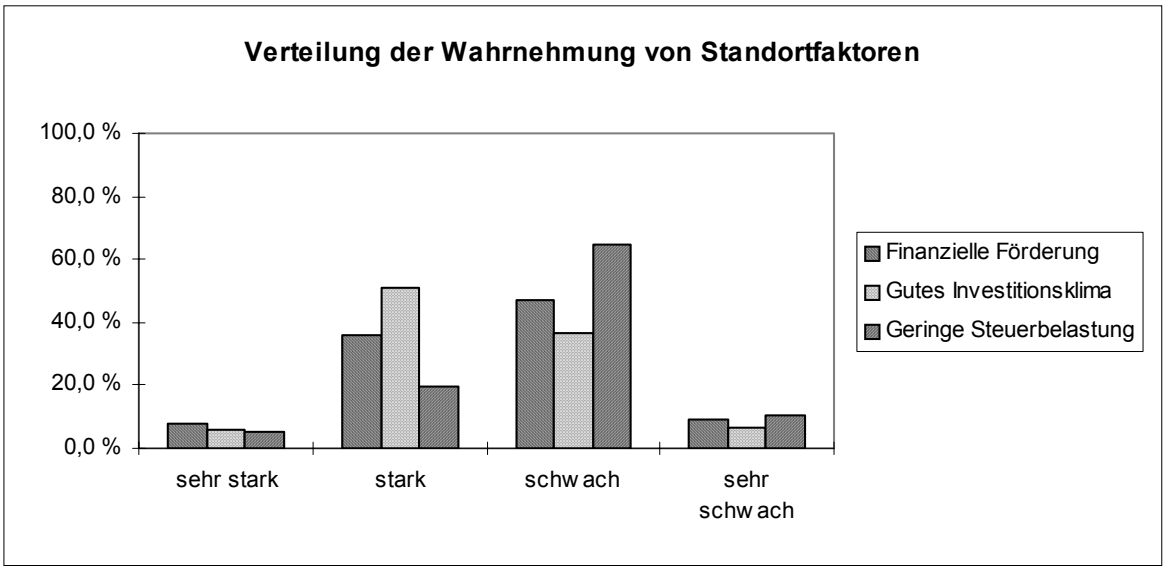
Abb. 9.2



Offensichtlich haben die Ansprüche für alle drei betrachteten Faktoren ein ähnliches Verteilungsbild. Die ursprünglich extra geführte Kategorie “unwichtig”, wurde hier mit “weniger

wichtig" zusammengefügt, da nur äußerst wenige Unternehmen, einen dieser drei Standortfaktoren als "unwichtig" empfanden. Bei den Wahrnehmungen sind allerdings alle vier Kategorien erhalten geblieben:

Abb. 9.3



Die Verteilungen der drei Standortfaktoren weisen bei den Wahrnehmungen nicht mehr die großen Ähnlichkeiten auf, wie bei der Formulierung der Ansprüche. Es wird ein eher heterogenes Bild ersichtlich, wobei aber nur relativ wenige Unternehmen die Extrempositionen bei der Wahrnehmung angeben. Offensichtlich zeichnet sich aber bezüglich der Finanziellen Förderung durch Land und Kommune, ebenso wie bei der Geringen Steuerbelastung ein für Brandenburg eher negatives Bild (s. a. Tab. 9.1 und 9.2). Besonders deutlich wird das bei der Einschätzung der Geringen Steuerbelastung, die 3/4 der 281 Unternehmen als schwach bis sehr schwach einschätzen. Die Frage stellt sich, inwieweit die Wahrnehmung z. B. steurlicher Vorteile durch die Unternehmenscharakteristika und oder die Ansprüche eventuell beeinflußt werden.
Während in der Unterteilung nach Branchen evtl. eine Differenzierung der Steuerbelastung erkennbar ist (sie ist nicht signifikant; s. Tab. 9.3), gibt es für die beiden anderen unternehmensspezifischen Variablen keinerlei Differenzierung.

Tab. 9.1 Wahrnehmungsunterschiede der "Geringen Steuerbelastung" innerhalb der Branchenzugehörigkeit

	Wahrnehmung: Geringe Steuerbelastung	
Branche	sehr stark / stark	sehr schwach / schwach
verarb. Gewerbe	25,4 %	74,6 %
Baugewerbe	17,8 %	82,2 %
Handel/Banken/Versicherungen	31,0 %	69,0 %
sonstige Dienstleistungen	25,7 %	74,3 %
gesamt:	24,9 %	75,1 %

Ein etwas klareres Bild ergibt sich bei Betrachtung der Wahrnehmung der Finanziellen Förderung:

Tab. 9.2 Wahrnehmungsunterschiede der "Finanziellen Förderung" innerhalb der Branchenzugehörigkeit

	Wahrnehmung: Finanzielle Förderung	
Branche	sehr stark / stark	sehr schwach / schwach
verarb. Gewerbe	63,4 %	36,6 %
Baugewerbe	31,5 %	68,5 %
Handel/Banken/Versicherungen	32,9 %	67,1 %
sonstige Dienstleistungen	47,1 %	52,9 %
gesamt:	43,8 %	56,2 %

Hier nehmen offensichtlich die Unternehmen des verarbeitenden Gewerbes die Finanzielle Förderung als stärker ausgebaut wahr, als die anderen Branchen[20]. Ein P^2-Unabhängigkeitstest für die entsprechende Kontingenztafel mit einer Irrtumswahrscheinlichkeit von 0,01 ergibt ein signifikantes Ergebnis.
Führt man für alle bivariaten Kontingenztafeln der beteiligten Variablen einen solchen P^2-Unabhängigkeitstest (zum Signifikanzniveau 0,01) durch ergibt sich folgende Tabelle:

[20] Es sollen hier keine Gründe für die jeweiligen Erscheinungen diskutiert werden. Z. B. kann die Ursache unterschiedlicher Wahrnehmung in den unterschiedlichen Finanzierungsleistungen für unterschiedlich strukturierte Unternehmen liegen.

Tab. 9.3 signifikante Ergebnisse für P^2-Unabhängigkeitstest[21]

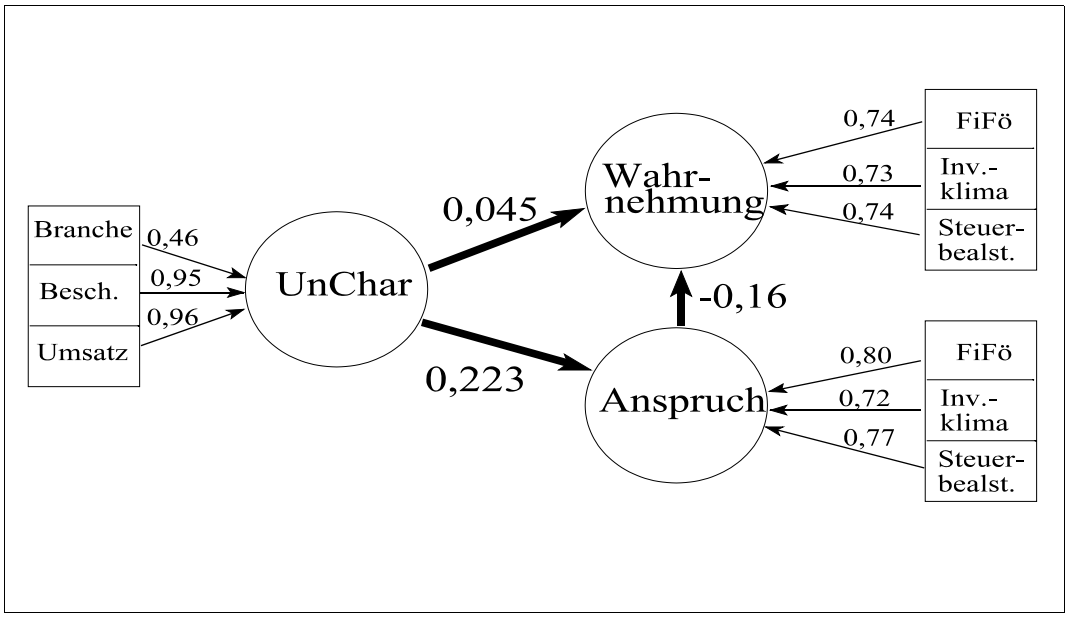
	BR	BG	UM	FF-A	GI-A	GS-A	FF-W	GI-W	GS-W
BR		**	**				**		
BG			**	**		**			
UM				**		**			
FF-A					**	**	**		
GI-A						**			
GS-A									**
FF-W								**	**
GI-W									**
GS-W									

(** kennzeichnet signifikante Ergebnisse für den entsprechenden P^2-Test)

Die in der obigen Tabelle stark umrandeten Felder heben die Ergebnisse innerhalb derselben Variablengruppe hervor. Die Werte für den P^2-Unabhängigkeitstest sind dort immer signifikant und können als Indiz dafür gelten, daß die Annahmen A1 und A2 nicht ganz unbegründet sind. Die Annahmen A3 und A4 sind aber nicht mehr so klar intuitiv nachvollziehbar.

Mit den in Kapitel 7 und 8 angegebenen Methoden ergeben sich für das Pfadmodell folgende Bewertungen:

Abb. 9.2: Anspruch an und Wahrnehmung von Standortfaktoren in Abhängigkeit von Unternehmenscharakteristika



Dabei sind die Koeffizienten die aus dem Struktur- und Meßgleichungsmodell (s. Kapitel 3)

---

21 Im weiteren werden für die Variablen und Kategorien zum Teil selbsterklärende Abkürzungen verwendet; eine Liste findet sich im Anhang.

berechneten Regressionskoeffizienten.
Im Sinne von Kapitel 7 werden die latenten Variablen dabei als stetige Realisation der bivariaten Häufigkeitsverhältnisse in dem zugehörigen Variablenblock betrachtet.
Für die kategorialen Einzelvariablen der MV-Blöcke ergeben sich durch die Gewichte **T** sogenannte optimale Skalierungen. Das sind stetige Realisationen der Häufigkeitsverteilung innerhalb der Variablen im Zusammenhang zu den anderen Variablen des MV-Blockes.

Als Bestimmtheitsmaße[22] für die Regressionen der MV auf die LV ergeben sich folgende Varianzaufklärungen durch die jeweilige Variable:

Tab. 9.4: Varianzerklärungsanteile der manifesten Variable

Variable	Varianzaufklärung (in %)
Branche	20
Beschäftigte	90
Umsatz	91
FF-Anspruch	63
GI-Anspruch	52
GS-Anspruch	59
FF-Wahrnehmung	54
GI-Wahrnehmung	70
GS-Wahrnehmung	67

Für die latente Variable Unternehmenscharakteristik bedeutet dies, daß Sie vorwiegend durch die Beschäftigten- und Umsatzkategorien bestimmt ist. Die Branche liefert keine gute Erklärung in diesem Konstrukt und mißt einen anderen Sachverhalt.
Die latenten Variablen Anspruch und Wahrnehmung als Konstrukt aus den jeweiligen Indikatorvariablen sind, auch im Kontext des Pfadmodells durchaus erklärungskräftig. Sie werden durch die Indikatoren mit jeweils über 50 % erklärt.
Damit lassen sich die Annahmen A1 und A2, wie oben bereits angedeutet, auch aus dem Pfadmodell heraus belegen. Insbesondere für die Annahme A1 sind aber evtl. Überlegungen zu machen, die Unternehmenscharakteristik anders zu definieren, um das Pfadmodell besser anzupassen[23].
Die Annahme A3, daß die Unternehmenscharakteristik sowohl auf Ansprüche als auch auf Wahrnehmungen Einfluß ausübt, kann durch das Modell nicht bestätigt werden.

---

[22] Es werden einfache OLS-Regressionen der MV auf die LV entsprechend der Pfeilnotation berechnet. Als Bestimmtheitsmaß wird der Anteil der "erklärten" Varianz (Varianz der in der Regression geschätzten $\hat{\eta}$) an der Gesamtvarianz der **0** berechnet.

[23] Rechnet man für den MV-Block der so wie bisher definierten Unternehmenscharakteristik eine Korrespondenzanalyse, erhält man eine Varianzaufklärung der Branche von annähernd 60 %. Dies ist hier nicht weiter ausgeführt, belegt aber die Einflüsse des Pfadmodells auf die Bildung der Gewichte innerhalb der MV-Blöcke.

Zwar ist der Regressionskoeffizient der Unternehmenscharakteristik auf die Ansprüche mit 0,223 von Null verschieden (signifikanter T-Test auf 1%-Niveau[24]), aber die Erklärungskraft ist mit knapp 5 % nicht von Bedeutung. Der Regressionskoeffizient von 0,045 in Richtung auf die Wahrnehmung ist dagegen statistisch gleich Null und somit besteht keine erklärende lineare Beziehung zwischen den latenten Variablen Unternehmenscharakteristik und Wahrnehmung in unserem Pfadmodell.
Der mit -0,16 angegebene Einfluß der Ansprüche auf die Wahrnehmungen ist auch statistisch von Null verschieden (signifikanter T-Test auf 1%-Niveau), allerdings erfaßt auch er nur 2,6 % der Varianz in den Wahrnehmungen und dürfte deshalb auch ohne Bedeutung in diesem Modell sein. Damit kann auch die Annahme A4, daß die Ansprüche Einflüsse auf die Wahrnehmungen ausüben, nicht durch das Mod ell verifiziert werden.
Ein weiteres Vorgehen muß sein, andere Anspruchsfaktoren in das Modell aufzunehmen und zu prüfen, ob sich die Modellparameter ändern. Sollte das Modell auch dann keine Erklärungskraft erreichen, sind für die Erklärung der Ansprüche und Wahrnehmungen von Standortfaktoren andere Indikatoren zu suchen.

10. Schlußfolgerungen und Ausblick

In der vorliegenden Arbeit wurde der Versuch unternommen, die Behandlung kategorialer Daten in Pfadmodellen mittels korrespondenzanalytischer Methoden zu lösen.
Ausgangspunkt ist ein von H. Wold entwickeltes Pfadmodell, welches Zusammenhänge zwischen verschiedenen Blöcken manifester metrischer Variablen über die Konstruktion von latenten Variablen darstellt und mißt. Die Fragestellung ist dabei ähnlich der der Kanonischen Korrelationsanalyse, allerdings gehen in das Pfadmodell kausale Strukturen ein.
Die Vorgehensweise des ursprünglichen PLS-Algorithmus auf kategoriale Daten übertragend und Eigenschaften der Korrespondenzanalyse nutzend lassen sich dann auch bewertbare latente Variable für kategoriale Variablenblöcke finden.
Die Abschnitte zur Homogenitätsanalyse und ihrer Verwandtschaft mit der Hauptkomponentenanalyse sowie ihre Anwendung im Pfadmodell sollen das Optimierungsziel des PLS-Algorithmus zur Konstruktion von latenten Variablen verdeutlichen.
Zudem, was keinen Eingang in die vorliegende Arbeit fand, lassen sich hier äquivalente Formulierungen der Homogenitätsanalyse und des PLS-Basisalgorithmus in Form von Eigenwertproblemen aufstellen. Mit deren Hilfe wiederum läßt sich der Iterationsalgorithmus in PLS zur Bestimmung der latenten Variablen in einer Eigenwertgleichung darstellen und es lassen sich die Optimierungseigenschaften des PLS-Algorithmus in verschiedenen Gewichtungsformen (Schritt-2 des Iterationsalgorithmus) erläutern.
Ebenfalls mit Hilfe dieser Eigenwertformulierung kann das noch immer nicht befriedigend gelöste Problem der Konvergenz des PLS-Basis-Algorithmus behandelt werden. Hier gibt es berechtigte Hoffnungen, in Kürze schlüssige Antworten vorlegen zu können.
Sehr wahrscheinlich ist der PLS-Algorithmus im Fall kategorialer Daten auch erfolgreich, wenn als Startgewichte nicht die Ergebnisse einer Korrespondenzanalyse benutzt werden, sondern auch

---

[24] Die Voraussetzungen zur Anwendung des t-Tests zur Prüfung des Regressionskoeffizienten sind im Prinzip nicht gegeben. Er wird hier nur als Indiz herangezogen. Dies gilt auch für die folgenden Testangaben.

hier beliebige Startgewichte zugelassen werden (s. Lohmöller, 1989). Die Idee die dahinterliegt ist aber sehr wohl die in Kapitel 7 und 8 dargelegte.
Was im weiteren geklärt werden muß ist die Bewertung der Varianzaufklärung der latenten Variablen in den zugehörigen Blöcken, um einen Anhaltspunkt für die Güte des aufgestellten Modells zu erhalten. Nötig sind weitere Erfahrungen im praktischen Einsatz des Modells.
Bisher haben Missing-Werte noch keinen Eingang in den oben beschriebenen Modellansatz gefunden, hier bedarf es weiterer Überlegungen, wie diese einzubeziehen sind.

Anhang

Symboltabelle

y , y_j	-	manifeste Variable, j=1, ..., P (P = Anzahl der MV im Modell)
$\mathbf{y}$, $\mathbf{y}_j$	-	Datenvektoren der MV y bzw. y_j
$\mathbf{y}_{j(m)}$	-	in Klammern gesetzte Indizes bezeichnen den entsprechenden MV-Block
y_{ij}, $y_{ij(m)}$	-	i-tes Element des Vektor $\mathbf{y}_j$, i=1, ..., N (N = Anzahl der Objekte/Beobachtungen)
$\boldsymbol{0}$, $\boldsymbol{0}_{(m)}$	-	latente Variable, m=1, ..., M (M = Anzahl der MV-Blöcke im Modell)
$\boldsymbol{0}$, $\boldsymbol{0}_{(m)}$	-	Vektor der Länge N der LV $\boldsymbol{0}$ bzw. $\boldsymbol{0}_{(m)}$
$\boldsymbol{0}_i$, $\boldsymbol{0}_{i(m)}$	-	i-tes Element des entsprechenden Vektors, i=1, ..., N
$\boldsymbol{0}^*$, $\boldsymbol{0}^*_{(m)}$	-	sogenannte Umgebungsvariable (s. S. 10)
$\boldsymbol{T}$, $\boldsymbol{T}_{(m)}$	-	Gewichtsvektor, die Länge ist aus dem jeweiligen Kontext ersichtlich
C_m	-	Indexmenge der mit $\boldsymbol{0}_{(m)}$ direkt im Pfadmodell verbundenen LV
C_m^{Pr}	-	Indexmenge der "Vorgänger" von $\boldsymbol{0}_{(m)}$
C_m^{Su}	-	Indexmenge der "Nachfolger" von $\boldsymbol{0}_{(m)}$
$\mathbf{D}_{\mathbf{Y}'\mathbf{Y}}$	-	Diagonalmatrix mit den Elementen der Hauptdiagonale von $\mathbf{Y}'\mathbf{Y}$ auf der Hauptdiagonale und sonst Nullen
$\mathbf{G}_j$	-	Indikatormatrix des Vektor $\mathbf{y}_j$
$\mathbf{Q}$	-	Matrix/Tabelle relativer Häufigkeiten
Q_{ij}	-	Elemete der Matrix Q
$\mathbf{R}$	-	Zeilenprofilmatrix
$\mathbf{C}$	-	Spaltenprofilmatrix

Abkürzungsverzeichnis

LV	**l**atente **V**ariable
MV	**m**anifeste **V**ariable
PLS	**P**artial **L**east **S**quares
LISREL	**L**inear **S**tructural **REL**ationships
PCA	**P**rincipal **C**omponent **A**nalysis
MGS	**M**eß**g**leichungs**s**ystem
SGS	**S**truktur**g**leichungs**s**ystem
GGS	**G**ewichtungs**g**leichungs**s**ystem
ALS	**A**lternating **L**east **S**quare

Liste der im Beispiel verwendeten Variablen- und Kategorienbezeichnungen

Variable:		Kategorie	
Vollname	Abkürzung	Vollname	Abkürzung
Branche	BR	verarb. Gewerbe	verG
		Baugewerbe	Bau
		Handel/Banken/Vers.	HBV
		sonstige Dienstleist.	sonst
Anzahl Beschäftigte	BG	bis 50 Beschäftigte	KB
		51-500 Beschäftigte	MB
		über 500 Beschäftigte	GB
Umsatz	UM	unter 5 Mio. DM	< 5
		5 bis 100 Mio. DM	5-100
		über 100 Mio. DM	> 100
Finanzielle Förderg.	FF		
Gutes Invest.-klima	GI		
Geringe Steuerbel.	GS		
Ansprüche		sehr wichtig	-A1
		wichtig	-A2
		unwichtig	-A3
Wahrnehmung		sehr stark	-W1
		stark	-W2
		schwach	-W3
		sehr schwach	-W4

Literaturverzeichnis

Balderjahn, Ingo / Aleff, Hans Jörg (1996)
"Die Wirtschaftsregion Brandenburg: Grundlagen für ein Standortmarketing", Verlag für Berlin-Brandenburg, Potsdam

Bertholet, Jean-Luc / Wold, Herman (1984)
Recent Developments on Categorical Data Analysis by PLS Modeling, ICUS Seminar, Washington

Gifi, Albert (1990)
Nonlinear Multivariate Analysis, John Wiley & Sons, New York u. a.

Glang, Manfred (1988)
Maximierung der Summe erklärter Varianzen in linear-rekursiven Strukturgleichungsmodellen mit multiplen Indikatoren: Eine Alternative zum Schätzmodus B des Partial-Least-Squares-Verfahrens, Dissertation (Uni Hamburg)

Greenacre, Michael J. (1984)
"Theory and Applications of Correspondence Analysis", Academic Press, London

Greenacre, Michael J. (1994)
"The Correspondence Analysis and ist interpretation", in: Greenacre, M./Blasis, J.: Correspondence Analysis in the Social Sciences, Academic Press, 1994.

Heuser, Harald (1986)
"Funktionalanalysis", B. G. Teubner, Stuttgart1

Jöreskog, Karl G. / Sörbom, Dag (1989)
"LISREL 7 - User's Refernce Guide", Scientific Software, Inc., Mooresville.

Kukuk, Martin (1991)
Latente Strukturgleichungsmodelle und rangskalierte Daten, Konstanzer Dissertationen, Bd. 330, Hartung-Gorre Verlag, Konstanz

Kühnel, Steffen (1993)
Lassen sich ordinale Daten mit linearen Strukturgleichungsmodellen analysieren, ZA-Information, 33

Lebart, L./ Morineau, A./ Kenneth, M. W. (1984)
"Multivariate Descriptive Statistical Analysis - Correspondence Analysis and Related Techniques for Large Matrices", John Wiley & Sons, New York u. a.

Lohmöller, Jan-Bernd (1984)
"Das Programmsystem LVPLS für Pfadmodelle mit Latenten Variablen", ZA-Information Nr. 15

Lohmöller, Jan-Bernd (1989)
"Latent Variable Path Modeling with Partial Least Squares", Physica-Verlag, Heidelberg

Mathes, Harald (1993a)
"Der PLS-Ansatz für die Analyse von Pfadmodellen", Mathematical Systems in Economics, Bd. 131, Verlag Anton Hain, Frankfurt am Main.

Mathes, Harald (1993b)
"Global Optimization Criteria of the PLS-Algorithm in Recursive Path Models with Latent Variables", in: Haage, K. / Bartholomew, D. J. / Deistler, M. (Eds.): "Statistical Modelling and Latent Variables"

Schnell, Rainer (1996)
"Graphisch gestützte Datenanalyse", Oldenbourg Verlag, München.

van der Heijden, P. G. M. /Mooijaart, A. / Takane, Y. (1994)
"Correspondence Analysis and Contingency Table Models", in: Greenacre, M./Blasis, J.: Correspondence Analysis in the Social Sciences, Academic Press, 1994.

Wold, Herman (1982)
Soft Modeling: The Basic Design and Some Extensions. In: Jöreskog, K. G./Wold, H. "Systems under indirect observation: Causality-structure-prediction, Part II, North-Holland, Amsterdam

UNIVERSITÄT POTSDAM
Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe
ISSN 0949-068X

Nr.1	1995	Strohe, Hans Gerhard: Dynamic Latent Variables Path Models - An Alternative PLS Estimation -
Nr.2	1996	Kempe, Wolfram: Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundesländern - Eine semiparametrische Regressionsanalyse -
Nr.3	1996	Strohe, Hans Gerhard: Statistik im DDR-Wirtschaftsstudium zwischen Ideologie und Wissenschaft
Nr.4	1996	Berger, Ursula: Die Landwirtschaft in den drei neuen EU-Mitgliedstaaten Finnland, Schweden und Österreich - Ein statistischer Überblick -
Nr.5	1996	Betzin, Jörg: Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten

Bezugsquelle: Universität Potsdam
Lehrstuhl Statistik der Wirtschafts- und Sozialwissenschaftlichen Fakultät
Postfach 90 03 27, D-14439 Potsdam
Tel. (+49 331) 977-32 25
Fax. (+49 331) 977-32 10
eMail: strohe@rz.uni-potsdam.de