4 PROBABILISTISCHE CLUSTERANALYSEVERFAHREN

4.1 Einleitende Übersicht

Die probabilistischen Clusteranalyseverfahren unterscheiden sich von den im vorausgehenden Kapitel behandelten deterministischen Verfahren dadurch, daß ein Objekt jedem Cluster nur mit einer bestimmten Wahrscheinlichkeit $\pi(k/g)$ angehört.[1] $\pi(k/g)$ ist die Wahrscheinlichkeit, mit der Objekt g dem Cluster k angehört. Wir werden diese Wahrscheinlichkeit im folgenden als Zuordnungswahrscheinlichkeit bezeichnen. Auch die Bezeichnung Rekrutierungswahrscheinlichkeiten ("recruitment probabilities") ist üblich (*Lazarsfeld und Henry* 1968: 36-39). In diesem Kapitel werden folgende probabilistische Verfahren behandelt:

- Latente Profilanalyse (Analyse latenter Klassen für quantitative Variablen, Abschnitt 4.2)
- Analyse latenter Klassen für nominalskalierte Variablen (Abschnitt 4.3)
- Analyse latenter Klassen für ordinalskalierte Variablen (Abschnitt 4.4)
- Analyse latenter Klassen für gemischte Variablen (Abschnitt 4.5)

Alle Verfahren lassen sich als Verallgemeinerung des K-Means-Verfahren entwickeln. Die Verfahren eignen sich somit - wie das K-Means-Verfahren - nur für eine objektorientierte Clusteranalyse. Technisch bestehen die Modifikationen im folgenden:

1. Der Schritt 2 des Algorithmus der K-Means-Verfahren, in dem jedes Objekt g dem Cluster zugeordnet wird, zu dem die quadrierte euklidische Distanz minimal ist, wird dahingehend geändert, daß die *Zuordnungswahrscheinlichkeiten* $\pi(k/g)$ berechnet werden. Dazu sind in Abhängigkeit vom Meßniveau bestimmte Verteilungsannahmen erforderlich. Ferner geht in die Berechnung die Annahme der *lokalen Unabhängigkeit* ein (siehe dazu später).
2. Die Klassenzentren $\bar{x}_{gj}$ und Klassenanteilswerte $\pi(k)$ (Schritt 3 des Algorithmus des K-Means-Verfahrens) werden als *Maximum-Likelihood-Schätzer* berechnet. Das

[1] Zur Beschreibung der Modellannahmen werden im folgenden griechische Buchstaben verwendet, für deren geschätzte Werte arabische.

heißt, sie werden so bestimmt, daß die empirische Verteilung der Objekte bestmöglich durch das Modell reproduziert wird.

Mit Ausnahme dieser beiden Modifikationen erfolgt die Berechnung nach dem Algorithmus der K-Means-Verfahren. Dieser Algorithmus für probabilistische Clusteranalyseverfahren wird in der Literatur als *EM-Algorithmus* (=Expected-Maximum-Likelihood-Estimator) bezeichnet und geht auf *Goodman* (1974) zurück. Der EM-Algorithmus hat sich seit seiner Einführung in zahlreichen Anwendungssituationen bewährt (siehe z.B. *Bock und Aitkin* 1981, *DeSoete und DeSarbo* 1991, *Langeheine und Van de Pol* 1990, *Rigdon und Tsutakawa* 1983, *Van de Pol und de Leeuw* 1986 u.a.). Das Konvergenzverhalten zur Lösung einer Schätzaufgabe wurde von *Dempster, Laird und Rubin* (1977) untersucht.

Das Konzept der *lokalen Unabhängigkeit* ist für die in diesem Kapitel behandelten Verfahren zentral. Es ist aus der Analyse latenter Strukturen mit der Analyse latenter Klassen als Submodell von *Lazarsfeld und Henry* (1968) bekannt (siehe dazu auch z.B. *Denz* 1982) und geht von folgender Modellvorstellung aus:

1. Den Daten liegen K unbekannte (=nicht beobachtete) Klassen zugrunde. Diese werden als latente Klassen bezeichnet und
2. erklären die Zusammenhänge zwischen den untersuchten beobachteten (=manifesten) Variablen. Werden die (latenten) Klassen also als Kontrollvariablen in die Analyse eingeführt, verschwinden die empirischen Zusammenhänge. Die manifesten Variablen sind innerhalb jeder Klasse unabhängig.

Wegen der Beziehung zur Analyse latenter Klassen von *Lazarsfeld und Henry* (1968) wurde für die in diesem Kapitel behandelten Verfahren die Bezeichnung Analyse latenter Klassen gewählt (Ausnahme: Latente Profilanalyse). Anstelle von "Clustern" wird von "latenten Klassen" oder kurz von "Klassen" gesprochen, obwohl man sich selbstverständlich unter den "Klassen" "Cluster" vorstellen kann. Die behandelten Verfahren lassen sich vorstellen als:

1. *Verallgemeinerung des K-Means-Verfahrens:* Die Annahme einer deterministischen Zuordnung der Objekte zu den Klassen wird fallengelassen.
2. *Verallgemeinerung der klassischen Analyse latenter Klassen von Lazarsfeld und Henry* (1968): Neben dichotomen Variablen können nominalskalierte Variablen mit beliebig vielen Ausprägungen, ordinalskalierte und/oder quantitative Variablen untersucht werden.
3. *Submodelle von Mischverteilungsverfahren:* Mischverteilungsverfahren (*Kaufmann und Pape* 1984: 420-445, *Wolfe* 1970 u.a.) gehen allgemein von folgender Pro-

blemstellung aus: Die empirische Verteilung der Objekte in den untersuchten Variablen ist eine Mischung von Wahrscheinlichkeitsverteilungen, z.B. von K m-dimensionalen Normalverteilungen mit den Mittelwertsvektoren $\boldsymbol{\mu}_k$ und den Kovarianzmatrizen $\mathbf{COV}_k$ mit den Elementen $\sigma(k)^2_{jj^*}$ ($\sigma(k)^2_{jj^*}$ = Kovarianz ($j \neq j^*$) bzw. Varianz ($j=j^*$) zwischen den Variablen j und j* in der k-ten Normalverteilung). Aufgabe von Mischverteilungsverfahren ist die Schätzung des Mischungsverhältnisses und der Parameter der einzelnen Wahrscheinlichkeitsverteilungen. Werden bestimmte zusätzliche Annahmen getroffen, können die Mischverteilungen durch die hier behandelten Verfahren geschätzt werden. So z.B. geht die latente Profilanalyse von der Modellvorstellung aus, daß K m-dimensionale Normalverteilungen vorliegen, wobei die Variablen innerhalb der (latenten) Klassen unabhängig sind.

4. Schließlich können die probabilistischen Clusteranalyseverfahren als *Clusteranalyseverfahren interpretiert werden, die eine Modellierung zufälliger Meßfehler erlauben* (*Espeland und Handelman* 1989, *Van de Pol und de Leeuw* 1986 u.a.).

Ein weiterer Unterschied zu den deterministischen Verfahren des Kapitels 3 besteht darin, daß das durch unterschiedliche Skaleneinheiten und Meßniveaus bedingte *Problem der Nichtvergleichbarkeit nicht auftritt*, da zur Berechnung der Zuordnungswahrscheinlichkeiten $\pi(k/g)$ mit im Intervall [0,1] normierten Wahrscheinlichkeiten $p(x_{gj}/k)$ gerechnet wird.

Umgekehrt haben natürlich die *probabilistischen Clusteranalyseverfahren* auch bestimmte Nachteile. Diese bestehen zum einen darin, daß größere Stichproben für eine konvergente Lösung - insbesondere für die latente Profilanalyse - benötigt werden als für die K-Means-Verfahren. Wenn wir das in Abschnitt 3.7.1 durchgeführte Rechenexperiment zur Veranschaulichung des Konvergenzverhaltens aufgreifen (siehe Tabelle 3.7.1), so wird bei zufälligen Startwerten selbst bei einer Stichprobengröße von 10000 für die latente Profilanalyse keine konvergente Lösung gefunden.[1] Mit der Zahl der Variablen verbessert sich das Konvergenzverhalten entscheidend (siehe Tabelle 4.1.1). Wie beim K-Means-Verfahren konvergieren die Schätzungen ab einer Stichprobe von 500. Bei den in der Tabelle 4.1.1 wiedergegebenen Simulationsergebnissen wurde von folgenden Annahmen ausgegangen: Es liegen zwei gleich große Klassen vor. Klasse 1 besitzt in den Variablen X_1 und X_2 eine Normalverteilung mit einem Mittelwert von -1 und einer Varianz von 1, Klasse 2 eine Normalverteilung mit einem Mittelwert von +1 und einer Varianz von ebenfalls 1. Ab einer Stichprobe von n=500 werden die wahren Klassenmittelwerte somit relativ gut geschätzt.

---

[1] Ein ähnlich schlechtes Konvergenzverhalten berichten *Kaufmann und Pape* (1984: 433).

Allgemein hängt das Konvergenzverhalten bei der latenten Profilanalyse von dem Überlappungsanteil ab (*Kaufmann und Pape* 1984: 433). Um so geringer der Überlappungsanteil ist, desto besser konvergiert das Verfahren asymptotisch. Das heißt, zur Schätzung der Modellparameter wird eine kleinere Stichprobe benötigt. Das Konvergenzverhalten der latenten Klassenanalyse für ordinale und nominale Variablen ist allgemein besser, da bei diesen beiden Verfahren weniger Parameter zu schätzen sind.

Tabelle 4.1.1: Veranschaulichung des asymptotischen Konvergenzverhaltens der probabilistischen Clusteranalyseverfahren am Beispiel der latenten Profilanalyse

Stichproben-größe	C1	C2	Stichproben-größe	C1	C2
20	-0.90 -0.55	1.10 1.44	500	-0.98 -1.11	1.03 1.05
50	-0.84 -1.55	0.86 1.36	1000	-0.96 -1.06	0.99 0.96
100	-0.78 -0.63	0.91 0.78	2000	-0.99 -0.99	1.02 0.98
200	-1.21 -1.08	1.20 1.02	5000	-1.00 -0.96	1.00 0.99
			10000	-1.04 -0.99	1.00 1.00

Ein weiterer "Nachteil" der Verfahren kann darin gesehen werden, daß strenggenommen vor der Analyse die *Identifikation des zu schätzenden Modells* untersucht werden muß. Ein zu schätzendes Modell M wird dann als identifiziert bezeichnet, wenn die zu schätzenden Modellparameter P eindeutig bestimmt sind. Das heißt, es darf kein anderes Modell M* mit anderen Modellparametern P* geben, das dieselben Modelldaten produziert. Eine notwendige Bedingung für die Eindeutigkeit (=Identifikation) eines Modells ist, daß die Zahl der empirischen Informationen größer der Zahl der zu schätzenden Modellparameter ist (siehe dazu die Ausführungen). Bei Anwendung der latenten Profilanalyse kann immer davon ausgegangen werden, daß das untersuchte Modell identifiziert ist, da die notwendige Bedingung (mehr empirische Informationen als Modellparameter) i.d.R. erfüllt ist und Mischungen von m-dimensionalen Verteilungen identifizierbar sind (*Kaufmann und Pape* 1984: 422-423). Bei den anderen Verfahren muß dies nicht unbedingt der Fall sein. Die Identifikation eines Modells kann dadurch geprüft werden, daß die gesuchten Modellparameter als Funktion der empirischen Daten ausgedrückt werden (*Bacher* 1990: 92-95, *Lazarsfeld und Henry* 1968: 59-68). Mitunter erfordert dieses Vorgehen aber komplexe mathematische Operationen. Deshalb wurden Verfahren zur computerunterstützten Identifikationsprüfung entwickelt (*Van de Pol, Langeheine und de Jong* 1989: 29). Dabei wird geprüft, ob lineare Abhängigkeiten zwischen den geschätzten Parametern bestehen. Ist dies der Fall,

ist das Modell nicht identifiziert. Umgekehrt kann allerdings aus dem Fehlen von linearen Abhängigkeiten nicht abgeleitet werden, daß ein Modell identifiziert ist.

4.2 Latente Profilanalyse

4.2.1 Modellansatz und Algorithmus

Das Modell der latenten Profilanalyse bzw. der Analyse latenter Klassen für quantitative Variablen wurde im Rahmen der Analyse latenter Strukturen (*Lazarsfeld und Henry* 1968: 228-239, *Gibson* 1966) entwickelt. Die Modellannahmen[1] sind:

1. Es liegen K latente Klassen vor.
2. Diese besitzen Anteilswerte von $\pi(k)$ in der Grundgesamtheit.
3. Jede Klasse k besitzt in jeder Klassifikationsvariablen j eine Normalverteilung[2] mit dem Mittelwert (=Klassenzentrum) μ_{kj} und der Varianz σ^2_{kj}.
4. In jeder Klasse k sind die Variablen j und j* unabhängig.

Die Normalverteilung in den Variablen können wir uns wie folgt zustandegekommen vorstellen: Der empirisch beobachtete Wert x_{gj} eines Objekts g aus k in der Variablen X_j setzt sich aus einem zufälligen Fehlerterm ε_{gj} und dem Klassenmittelwert μ_{kj} zusammen: $x_{gj} = \mu_{kj} + \varepsilon_{gj}$. Der zufällige Fehlerterm ε_{gj} ist die Realisierung einer normalverteilten Zufallsvariablen ξ_{kj} mit Erwartungswert 0 und Varianz σ^2_{kj}. Er kann durch zufällige Meßfehler und/oder zufällige individuelle Unterschiede in den einzelnen Variablen entstehen. Da die Abweichungen $\varepsilon_{gj} = x_{gj} - \mu_{kj}$ zufällig auftreten, sind sie in zwei Variablen j und j* unabhängig. Wäre dies nicht der Fall, würde es wenig Sinn machen, von Zufälligkeit zu sprechen. Formal ausgedrückt sind somit die Ausnahmen:

1. $x_{gj} = \mu_{kj} + \varepsilon_{gj}$ für alle g, die Elemente aus k sind.
2. ε_{gj} ist die Realisierung einer normalverteilten Zufallsvariablen ξ_{kj}.
3. ξ_{kj} besitzt einen Erwartungswert von 0 und eine Varianz von σ^2_{kj}.
4. Die Zufallsvariablen ξ_{kj} sind unabhängig. Es gilt also $COV(\xi_{kj},\xi_{kj*}) = 0$.

wobei x_{gj} der empirische Wert des Objekts g aus der Klasse k in der Variablen j ist. μ_{kj} ist der Klassenmittelwert in der Variablen j und ε_{gj} der zufällige Fehlerterm.

Aus diesen Modellannahmen lassen sich folgende Aussagen über die Mittelwerte, Varianzen und Kovarianzen der Gesamtpopulation ableiten (*Lazarsfeld und Henry* 1968: 229-231):

---

[1] Zur Beschreibung der Modellannahmen werden im folgenden griechische Buchstaben verwendet, für deren geschätzte Werte arabische.

[2] In dem klassischen Ansatz der latenten Profilanalyse ist eine Verteilungsannahme nicht erforderlich, da die Modellparameter nicht über die Maximum-Likelihood-Methode geschätzt werden.

(4.2.1) $$\mu_j = \sum_k \pi(k) \cdot \mu_{kj}.$$

(4.2.2) $$\sigma_{jj^*} = \sum_k \pi(k) \cdot \left(\mu_{kj} - \mu_j\right) \cdot \left(\mu_{kj^*} - \mu_{j^*}\right).$$

(4.2.3) $$\sigma_j^2 = \sigma_{jj} = \sum_k \pi(k) \cdot \sigma_{kj}^2 + \sum_k \pi(k) \cdot \left(\mu_{kj} - \mu_j\right)^2 .$$

Der Gesamtmittelwert einer Variablen ist gleich dem gewichteten Mittelwert der Klassenzentren. Die Kovarianz zwischen zwei Variablen j und j* hängt nur von den Abweichungen der Klassenzentren von den Gesamtmittelwerten ab, die Varianz einer Variablen dagegen auch noch von den Fehlerstreuungen (siehe dazu auch Abschnitt 3.2.7).

Lazarsfeld und Henry (1968: 231-233) sowie *Gibson* (1966) entwickelten auf der Grundlage der Modellgleichungen (4.2.5) bis (4.2.7) sogenannte "accounting-equations", die eine Schätzung der Modellparameter mit Hilfe einer Eigenwertzerlegung ermöglichen. Diese hat mehrere Nachteile (*Van de Pol und de Leeuw* 1986). So z.B. können negative Zuordnungswahrscheinlichkeiten entstehen.

Eine andere Schätzmethode besteht in der Verwendung des in Abschnitt 4.1 erwähnten EM-Algorithmus. Die Modellparameter werden so geschätzt, daß die Likelihood-Funktion

(4.2.4) $$L = \prod_g \sum_k \pi(k) \cdot \pi(g/k)$$

bzw. ihr Logarithmus

(4.2.5) $$l = \ln(L) = \sum_g \ln \sum_k \pi(k) \cdot \pi(g/k)$$

ein Maximum wird. $\pi(k)$ ist dabei der "wahre" Anteil der Klasse k und $\pi(g/k)$ ist die (bedingte) Wahrscheinlichkeit des Auftretens des Merkmalsvektors der Person g in der Klasse k. Die bedingte Wahrscheinlichkeit $\pi(g/k)$ ist wegen der lokalen Unabhängigkeit (siehe Abschnitt 4.1) gleich dem Produkt der (bedingten) Auftrittswahrscheinlichkeiten $\pi(x_{gi}/k)$ des Wertes von g in den Variablen j für die Klasse k:

(4.2.6) $$\pi(g/k) = \prod_j \pi(x_{gj}/k).$$

Die Auftrittswahrscheinlichkeit $\pi(x_{gj}/k)$ ist bei der latenten Profilanalyse als Wert der Dichtefunktion der Normalverteilung mit dem Mittelwert μ_{kj} und der Varianz σ_{kj}^2 definiert:

(4.2.7) $$\pi(x_{gj}/k) = \varphi(x_{gj}/\mu_{kj}, \sigma_{kj}^2) = \frac{1}{\sigma_{kj} \cdot \sqrt{2 \cdot \pi}} \cdot e^{-0.5 \cdot \left(x_{gj} - \mu_{kj}\right)^2 / \sigma_{kj}^2},$$

wobei $\varphi(...)$ die Dichtefunktion der Normalverteilung ist. Betrachten wir dazu ein Beispiel: Die Klasse k soll in der Variablen j einen Mittelwert μ_{kj} von -1 und eine Varianz σ_{kj}^2 von 2 haben. Das Objekt g besitzt in der Variablen j einen Wert von -2. Die Auftrittswahrscheinlichkeit des Wertes von -2 in der Klasse k ist entsprechend (4.2.11)

gleich dem Wert der Dichtefunktion der Normalverteilung mit den Modellparametern der Klasse k:

$$\pi(x_{gj} = -2 / k) = \varphi(x_{gj} = -2 / \mu_{kj} = -1, \sigma_{kj}^2 = 2) = \frac{1}{1.41 \cdot \sqrt{2 \cdot \pi}} \cdot e^{-0.5 \cdot (-2-(-1))^2 / 2} = 0.2197 .$$

Der Wert von -2 tritt also mit einer Wahrscheinlichkeit von 0.2197 in der Klasse k auf. Dieser Wert ergibt sich auch, wenn mit (-2-(-1))/$\sqrt{2}$ = 0.707 der z-Wert berechnet und die Dichtefunktion der Standardnormalverteilung dividiert mit der Standardabweichung verwendet wird.

Die gesuchten Parameter $\pi(k)$, μ_{kj} und σ_{kj}^2 werden nun so bestimmt, daß die Log-Likelihood-Funktion ein Maximum ist. Das heißt, es werden Schätzungen gesucht, die (theoretische) Modelldaten mit einer Verteilung erzeugen, die der empirischen Verteilung der Objekte bestmöglich angepaßt ist. In die Schätzung gehen die Nebenbedingungen $\sum \pi(k) = 1$ und $\pi(k)>0$ für alle k ein. Die erste Nebenbedingung besagt, daß die Summe der Klassenanteilswerte gleich 1 sein soll. Die zweite Nebenbedingung bedeutet, daß keine Klasse leer sein soll. Unter Berücksichtigung dieser Nebenbedingung sind bei K Klassen und m Variablen folgende Parameter zu schätzen:

K-1	Klassenanteilswerte $\pi(k)$ (ein Klassenanteilswert ist wegen der Nebenbedingung $\sum \pi(k) = 1$ fixiert)
K·m	Klassenzentren oder Klassenmittelwerte μ_{kj} (in jeder Klasse für jede Variable ein Klassenzentrum)
K·m	Klassenstreuungen oder Klassenvarianzen σ_{kj}^2 (in jeder Klasse für jede Variable eine Klassenvarianz)
K·(1 + 2·m) - 1	Gesamtzahl zu schätzender Parameter

Insgesamt sind also K·(1 + 2·m) - 1 Parameter zu bestimmen. Wird beispielsweise eine 6-Klassenlösung bei drei Variablen (m=3) gesucht, sind 6·(1 + 2·3) - 1 = 41 Parameter zu schätzen. Die notwendige Bedingung für ein identifiziertes Modell ist somit, daß mindestens 41 unterschiedliche Datensätze vorliegen. Die notwendige Bedingung wird i.d.R. bei der latenten Profilanalyse immer erfüllt sein. Wie bereits erwähnt, ist beim Vorliegen der notwendigen Identifikationsbedingung das Modell der latenten Profilanalyse dann auch identifiziert.

Beim EM-Algorithmus wird nun angenommen, daß die Zuordnungswahrscheinlichkeiten $\pi(k/g)$, mit der die Klasse k bei den Objekten g auftritt, bekannt sind. Dadurch vereinfacht sich die Log-Likelihood-Funktion zu (*Van de Pol und de Leeuw* 1986):

$$\begin{aligned} l &= \sum_g \sum_k \pi(k/g) \cdot \left(\ln(\pi(k)) + \ln(\pi(g/k))\right) \\ &= \sum_g \sum_k \pi(k/g) \cdot \ln(\pi(k) + \sum_g \sum_k \sum_j \pi(k/g) \cdot \pi(x_{gj}/k) . \end{aligned} \tag{4.2.8}$$

Die Schätzaufgabe zerfällt also zum einen in die Schätzung der Anteilswerte $\pi(k)$ und zum anderen in die Schätzung der Parameter der einzelnen Normalverteilungen. Die Schätzwerte sind (*Bock* 1974: 258, *Kaufmann und Pape* 1984: 432)[1] :

(4.2.9) $$p(k) = \sum_g \pi(k/g) / n,$$

(4.2.10) $$\bar{x}_{kj} = \sum_g \pi(k/g) \cdot x_{gj} \Big/ \sum_g \pi(k/g),$$

(4.2.11) $$s_{kj}^2 = \sum_g \pi(k/g) \cdot \left(x_{gj} - \bar{x}_{kj}\right)^2 \Big/ \sum_g \pi(k/g),$$

wobei der Schätzwert von $\pi(k)$ mit $p(k)$ bezeichnet wurde. Der Schätzwert von μ_{kj} wurde mit $\bar{x}_{kj}$ und jener von σ_{kj}^2 mit s_{kj}^2 bezeichnet. Die Schätzwerte kann man sich wie folgt vorstellen: Zur Berechnung der Parameter der Klasse k werden die Datensätze mit $\pi(k/g)$ gewichtet und ausgezählt. Dadurch erhält man die Mittelwerte und Varianzen der Klasse k in den Variablen. Die gewichtete Fallzahl dividiert durch die ungewichtete Fallzahl ergibt den Anteil der Klasse k.

Bei den bisherigen Ausführungen wurde davon ausgegangen, daß die Zuordnungswahrscheinlichkeiten $\pi(k/g)$ bekannt sind. Dies ist natürlich nicht der Fall. Sie können aber ihrerseits aus den geschätzten Modellparametern mit Hilfe des Satzes von Bayes (*Fisz* 1980: 40-41) berechnet werden mit:

(4.2.12) $$p(k/g) = \frac{p(k) \cdot p(g/k)}{\sum_k p(k) \cdot p(g/k)},$$

wobei $p(k/g)$ der Schätzwert von $\pi(k/g)$ ist. $p(k)$ ist der Schätzwert von $\pi(k)$ und $p(g/k)$ der Schätzwert von $\pi(g/k)$. Damit haben wir das *Grundprinzip des EM-Algorithmus* skizziert. Es besteht aus zwei Schritten:

1. *E-Schritt:* Die Zuordnungswahrscheinlichkeiten $\pi(k/g)$ werden aufgrund der geschätzten Modellparameter (=Erwartungswerte) berechnet. Die geschätzten Modellparameter werden dabei als gegeben angenommen.
2. *M-Schritt:* Die Modellparameter $\pi(k)$, μ_{kj} und σ_{kj}^2 werden aufgrund der Schätzwerte der Zuordnungswahrscheinlichkeiten nach der *Maximum-Likelihood-Methode* geschätzt. Die Zuordnungswahrscheinlichkeiten werden dabei als gegeben angenommen.

---

[1] *Bock* (1974: 258) sowie *Kaufmann und Pape* (1984: 432) leiten die Schätzungsfunktion direkt aus der Gleichung (4.2.5) ab.

Diese Schritte werden solange wiederholt, bis eine konvergente Lösung gefunden ist. Es ergibt sich folgender Schätzalgorithmus:

Schritt 1: Berechnung oder Eingabe der Startwerte.

Schritt 2: Berechnung der Zuordnungswahrscheinlichkeiten. Entsprechend der Gleichung (4.2.12) werden die Zuordnungswahrscheinlichkeiten berechnet mit:

$$p(k/g)^{(i)} = \frac{p(k)^{(i-1)} \cdot p(g/k)^{(i-1)}}{\sum_k p(k)^{(i-1)} \cdot p(g/k)^{(i-1)}},$$

wobei aufgrund der Gleichungen (4.2.10) und (4.2.11)

$$p(g/k)^{(i-1)} = \prod_j p(x_{gj}/k)^{(i-1)} = \prod_j \varphi(x_{gj}/\bar{x}_{kj}^{(i-1)}, s_{kj}^{2(i-1)})$$

ist. Der hochgestellte Index in Klammern ist der Iterationszähler. Beim ersten Durchlaufen ist i=0.

Schritt 3: Neuberechnung der Modellparameter. Die Modellparameter werden entsprechend den Gleichungen (4.2.9) bis (4.2.11) neu berechnet mit:

$$p(k)^{(i)} = \sum_g p(k/g)^{(i)} / n.$$

$$\bar{x}_{kj}^{(i)} = \sum_g p(k/g)^{(i)} \cdot x_{gj} \Big/ \sum_g p(k/g)^{(i)}.$$

$$s_{kj}^{2(i)} = \sum_g p(k/g)^{(i)} \cdot \left(x_{gj} - \bar{x}_{kj}^{(i)}\right)^2 \Big/ \sum_g p(k/g)^{(i)}.$$

Schritt 4: Prüfung der Konvergenz. Der Algorithmus wird dann abgebrochen, wenn (a) die Verbesserung der Log-Likelihood-Funktion kleiner einem vorgegebenen Schwellenwert (z.B. 10^{-7}) und/oder (b) die maximale Abweichung der aufeinanderfolgenden Schätzwerte kleiner einem zweiten Schwellenwert (z.B. 0.0001) ist.

Der Algorithmus ist mit jenem des K-Means-Verfahren strukturgleich. Das asymptotische Konvergenzverhalten ($n \to \infty$) dieses Algorithmus wurde bereits im einleitenden Abschnitt beschrieben.

(....)

4.2.2 Modellprüfgrößen

4.2.2.1 Bestimmung der Klassenzahl

Zur Bestimmung der Klassenzahl wird das Verfahren wiederum mit einer unterschiedlichen Anzahl von Klassen durchgerechnet. Für die Wertedaten von *Denz* (1989) ergeben sich die in der Tabelle 4.2.2 dargestellten Werte für die Log-Likelihood-Funktion, wenn die Gesamtpunktewerte für die postmaterialistische und materialistische Wertorientierung als Klassifikationsvariablen in die Analyse einbezogen und die Startwerte mit dem Quick-Clustering-Verfahren berechnet werden.

Tabelle 4.2.1: Modellprüfgrößen der latenten Profilanalyse für die Wertedaten von Denz (Startwerte aus dem Quick-Clustering-Verfahren)

Klassenzahl	Wert der Log-Likelihood-Funktion	prozentuelle Verbesserung gegenüber Nullmodell	Informationsmaß von AKAIKE	prozentuelle Verbesserung gegenüber vorausgehender Lösung
K	L_K	$PV0_K$	IA_K	PV_K
1	-374.101	0.000	-378.101	KW
2	-340.546	8.969	-349.546	8.969
3	-321.242	14.129	-335.242	5.669
4	-314.870	15.833	-333.870	1.984
5	-301.851	19.313	-325.851	4.135
6	-300.303	19.727	-329.303	0.513
7	-298.240	20.278	-332.240	0.687
8	-298.567	20.191	-337.567	-0.110
9	-283.048	24.339	-327.048	5.198
10	-282.748	24.419	-331.748	0.106
11	-286.485	23.420	-340.485	-1.322
12	-266.264	28.826	-325.264	7.059

Aus den Werten der Log-Likelihood-Funktion[1] L_K lassen sich folgende Modellprüfgrößen berechnen:

Prozentuelle Verbesserung $PV0_K$ gegenüber dem Nullmodell der 1-Klassenlösung: Diese wird analog zu ETA[2] berechnet mit

$$(4.2.13) \qquad PV0_K = 1 - \frac{|L_K|}{|L_1|} \quad \text{bzw.} \quad PV0_K(\text{in }\%) = 100 \cdot PV0_K,$$

wobei $|L_1|$ der Absolutbetrag der Log-Likelihood-Funktion für die 1-Klassenlösung und $|L_K|$ der Absolutbetrag der Log-Likelihood-Funktion der K-Klassenlösung ist. Für die 5-Klassenlösung beispielsweise beträgt die prozentuelle Verbesserung 19.3 Prozent,

[1] Allgemein bedeutet ein kleinerer negativer Wert eine bessere Modellanpassung.

da der Absolutbetrag $|L_1|$ = 374.101 und $|L_5|$ = 301.851 ist. $PV0_5$ ist daher gleich 1 - 301.851/374.101=0.193 (=19.3 Prozent).

Prozentuelle Verbesserung PV_K gegenüber der vorausgehenden Klassenlösung: Diese Maßzahl ist analog dem PRE-Koeffizienten beim K-Means-Verfahren definiert mit

(4.2.14) $$PV_K = 1 - \frac{|L_K|}{|L_{K-1}|} \quad \text{bzw.} \quad PV_K(\text{in \%}) = 100 \cdot PV_K .$$

Für die 5-Klassenlösung ergibt sich mit $|L_{5-1}| = |L_4|$ =314.870 und $|L_5|$ =301.851 ein Wert von 1 - 301.851/314.870=0.04135 (=4.1 Prozent.).

Informationsmaß von Akaike (*Akaike* 1974, *Kaufmann und Pape* 1984: 443): Dieses ist definiert mit

(4.2.15) $$IA_K = L_K - m_K,$$

wobei m_K die Zahl der zu schätzenden Parameter ist. Für die latente Profilanalyse ohne Restriktionen ist m_K gleich (K-1)+m·K+m·K, da (K-1) Klassenanteilswerte und jeweils $m \cdot K$ Klassenmittelwerte bzw. Klassenvarianzen zu schätzen sind (m=Zahl der Variablen). Für K=5 ergibt sich in unserem Beispiel eine Zahl zu schätzender Parameter von (5-1) + 2·5 + 2·5 = 24. Das Informationsmaß für die 5-Klassenlösung ist daher –301.851 - 24 = 325.851. Das Informationsmaß von Akaike berücksichtigt somit wie die F-MAX-Statistik die Tatsache, daß bei einer größeren Klassenzahl in der Tendenz "automatisch" eine bessere Modellanpassung erzielt wird.

Chi-Quadrat-Test für die Likelihood-Quotienten-Teststatistik: Wegen der Maximum-Likelihood-Schätzung ist die Teststatistik $-2\cdot(L_{K-1} - L_K)$ approximativ Chi-Quadrat-verteilt mit df = m_K - m_{K-1} Freiheitsgraden. Bei kleinen Stichproben ist diese Approximation schlecht, es sollte daher die modifizierte Likelihood-Quotienten-Teststatistik nach *Wolfe* verwendet werden (*Kaufmann und Pape* 1984: 443):

(4.2.16) $$LQ_K(\text{Wolfe}) = -\tfrac{2}{n} \cdot (n - 1 - m - K/2) \cdot (L_{K-1} - L_K).$$

Diese Testgröße ist approximativ Chi-Quadrat-verteilt mit 2·m Freiheitsgraden. Mit dieser Teststatistik wird geprüft, ob die K-Klassenlösung eine signifikant bessere Modellanpassung erbringt als die (K-1)-Klassenlösung. Sie entspricht somit den Bealschen F-Werten. Sollen allerdings die Klassenlösungen mit K-h (h>1) Klassen mit der K-Klassenlösung verglichen werden, muß die gewöhnliche Likelihood-Quotienten-Statistik $-2\cdot(L_{K-h} - L_K)$ verwendet werden. Für den Vergleich der 4- und 5-Klassenlösung ergibt sich ein Wert von

$$LQ_5(\textit{Wolfe}) = -\tfrac{2}{221} \cdot (221 - 1 - 2 - \tfrac{2}{5}) \cdot \big(-314.870 - (-301.851)\big) = 25.39 ,$$

da n = 221, m = 2, K = 5 sowie L_4 = -314.870 und L_5 = -301.851 sind. Bei 2·2 Freiheitsgraden ist dieser zu einem Signifikanzniveau von 100 Prozent von 0 verschieden. Die gewöhnliche Likelihood-Quotienten-Statistik ist in dem Beispiel gleich 26.038 und besitzt 5 Freiheitsgrade. Auch sie ist zum Niveau von 100 Prozent signifikant von 0 verschieden. Die 5-Klassenlösung verbessert somit die 4-Klassenlösung signifikant.

Übersicht 4.2.1: Beziehung zwischen den Modellprüfgrößen der latenten Profilanalyse und jenen des K-Means-Verfahrens

Modellprüfgrößen der latenten Profilanalyse	Modellprüfgrößen des K-Means-Verfahren	Anwendung zur Bestimmung der Klassenzahl
Prozentuelle Verbesserung gegenüber 1-Klassenlösung (=$PV0_K$)	Erklärte Streuung (=ETA_K^2)	Es werden nur jene Lösungen ausgewählt, für die $PV0_K$ einen bestimmten Wert überschreitet.
Prozentuelle Verbesserung gegenüber vorausgehender Lösung (=PV_K)	PRE-Koeffizient (=PRE_K^2)	Es wird (werden) jene Lösung(en) ausgewählt, bei der (denen) PV_K im Vergleich zu der voraus-gehenden Lösung relativ groß ist.
Informationsmaß von Akaike (=IA_K)	Maximaler F-MAX-Wert (=$F\text{-}MAX_K$)	Es wird jene Lösung mit dem maximalen Informationsmaß (=kleinster negativer Wert) ausgewählt.
Likelihood-Quotienten-Statistiken (gewöhnliche LQ-Statistik und LQ_K(Wolfe))	Bealsche F-Werte	Es wird jene Lösung ausgewählt, die (a) im Vergleich zu allen vorausgehenden Lösungen signifikant und (b) im Vergleich zu allen nachfolgenden Lösungen nicht signifikant ist.

Die bereits erwähnte Analogie der Modellprüfgrößen zu den Modellprüfgrößen des K-Means-Verfahrens gilt auch für das Vorgehen bei der Bestimmung der Klassenzahl (siehe Übersicht 4.2.1). Wendet man die einzelnen Strategien an, würden wir uns für folgende Lösungen entscheiden:

Prozentuelle Verbesserung PV_K gegenüber vorausgehender Lösung: Für die 2-, 3-, 5-, 9- und 12-Klassenlösung, da hier die prozentuelle Verbesserung relativ groß ist. Absolut betrachtet sind die prozentuellen Verbesserungen aber gering (<10 Prozent).

Informationsmaß von Akaike: Für die 12-Klassenlösung, da hier der Wert des Informationsmaßes mit -325.264 am größten ist (=kleinster negativer Wert). Allerdings ist der Wert der 5-Klassenlösung mit -325.851 nur geringfügig kleiner.

Likelihood-Quotienten-Test: Für eine Analyse mit einer größeren Klassenzahl, da die 12-Klassenlösung (=12-Spalte) gegenüber den vorausgehenden Klassenlösungen signifikant ist.

Prozentuelle Verbesserung gegenüber Nullmodell der 1. Klassenlösung: Mitunter würden wir uns hier für die 9-Klassenlösung entscheiden, da für sie - im Vergleich zur 2-, 3- und 5-Klassenlösung - die prozentuelle Verbesserung größer 20 Prozent ist.

Wir wollen im folgenden zunächst die 5-Klassenlösung weiter untersuchen.

4.2.2.2 Modellprüfgrößen für eine bestimmte Klassenlösung

Für eine bestimmte Klassenlösung, z.B. für die 5-Klassenlösung, können zur Beschreibung die entsprechenden Modellprüfgrößen verwendet werden. Darüber hinaus können - wie beim K-Means-Verfahren - varianzanalytische Maßzahlen verwendet werden, insbesondere die erklärte Streuung. In unserem Beispiel ergibt sich eine erklärte Streuung von 60.9 Prozent. Die erklärte Streuung bei der latenten Profilanalyse ist i.d.R. kleiner als beim K-Means-Verfahren, da die Fehlerstreuung nicht minimiert wird. Beim K-Means-Verfahren ergibt sich eine erklärte Streuung von 72.5 Prozent für die 5-Clusterlösung, wenn mit Quick-Clustering-Startwerten gerechnet wird.

4.2.2.3 Zufallstestung einer Klassenlösung

Wie beim K-Means-Verfahren kann auch bei der latenten Profilanalyse mit Hilfe des Nullmodells einer homogenen, normalverteilten Population geprüft werden, ob eine bestimmte Klassenlösung überzufällig ist. Dazu werden wiederum Zufallsdatenmatrizen für das homogene Nullmodell erzeugt. Für diese wird geprüft, wie gut sie durch die berechnete Klassenlösung reproduziert werden können. Es wird also wiederum das Modell mit vorgegebener Klassenstruktur verwendet. Ergibt sich eine annähernd gleich gute Reproduktion - gemessen durch den Wert der Log-Likelihood-Funktion -, wird man die Lösung als Zufallsprodukt betrachten. Führt man 20 Simulationen durch, ergeben sich die in der Tabelle 4.2.3 dargestellten Werte.

Tabelle 4.2.2: Simulationswerte für die Log-Likelihood-Funktion aus einer Zufallstestung

-452.339	-417.273	-447.587	-475.383	-432.936	-433.723	-504.192
-461.704	-430.683	-433.043	-461.709	-407.436	-441.347	-492.683
-503.631	-504.743	-537.952	-446.600	-481.406	-494.642	

Alle berechneten Log-Likelihood-Werte sind unter der Annahme einer homogenen Population deutlich kleiner dem empirischen Wert von -301.851. Der Mittelwert der Simulationswerte ist gleich -463.051, die Standardabweichung hat einen Wert von 33.953. Konstruieren wir eine z-Teststatistik mit $z=(t-E(t))/\sigma(t)$, wobei t der Wert der empirischen Log-Likelihood-Funktion ($t=L_K$), E(t) der Mittelwert der Log-Likelihood-Werte der simulierten Daten und $\sigma(t)$ deren Standardabweichung ist, ergibt sich ein Wert von 4.75. Dieser ist größer einem kritischen Schwellenwert von 2. Wir können daher die 5-Klassenlösung als überzufällig betrachten.

4.2.3 Beschreibung und Interpretation einer Klassenlösung

Bei der Beschreibung und Interpretation einer Klassenlösung wird analog wie beim K-Means-Verfahren vorgegangen. Das bedeutet u.a.:

1. Für jede Variable kann geprüft werden, ob sie signifikant zur Trennung der Klassen beiträgt. Dazu wird die durch eine Variable erklärte Streuung und ein entsprechender F-Wert berechnet. Da im Unterschied zum K-Means-Verfahren nicht die Streuungsquadratsumme in den Klassen minimiert wird, ist die Durchführung eines Signifikanztests für den F- Wert angemessener.
2. Es können die paarweisen Unterschiede zwischen den Klassen berechnet werden.
3. Die Variablen innerhalb einer Klasse können zu Variablengruppen zusammengefaßt werden.
4. Es können z-Werte zur Beantwortung der Frage, ob signifikante Abweichungen von den Gesamtmittelwerten vorliegen, berechnet werden.
5. Zur Beschreibung und Validitätsprüfung können Deskriptionsvariablen in die Analyse einbezogen werden.

Wir wollen hier nicht die einzelnen Schritte durchgehen, sondern die 5-Klassenlösung der latenten Profilanalyse mit der 5-Klassenlösung des K-Means-Verfahrens vergleichen. Die bei beiden Verfahren berechneten Klassenzentren und Klassengrößen enthält die Tabelle 4.2.4. Die Klassenlösungen stimmen hinsichtlich der Klassenzentren sehr gut überein. Die Klasse C1 läßt sich als Anti-Postmaterialisten interpretieren, die allerdings bei beiden Verfahren nur einen Anteil von 0.9 Prozent besitzt. Die zweite Klasse läßt sich als Konsenstypus mit geringerem Interesse und einer leichteren materialistischen Wertepräferenz interpretieren. Klasse 3 könnte als Cluster der Anti-Materialisten bezeichnet werden. Wie bei den Anti-Postmaterialisten ist der Anteilswert dieser Klasse allerdings gering. Klasse C4 läßt sich als gemäßigte Postmaterialisten interpretieren. Sie besitzt den größten Anteilswert (latente Profilanalyse=70.1 Prozent,

K-Means=55.5 Prozent). Die letzte Klasse schließlich läßt sich als Klasse der Postmaterialisten bezeichnen.

Bezüglich der Klassenanteilswerte treten etwas größere Unterschiede auf. So z.B. besitzt die Klasse C5 bei der latenten Profilanalyse einen Anteil von 12.7 Prozent, beim K-Means-Verfahren dagegen von 25.7 Prozent. Dies ist auf den unterschiedlichen Modellansatz zurückzuführen. Beim K-Means-Verfahren werden die Objekte deterministisch den Klassen zugeordnet. Besitzt beispielsweise ein Objekt die Zuordnungswahrscheinlichkeit 0.24 für die erste Klasse und die Zuordnungswahrscheinlichkeiten von 0.19 für die anderen fünf Klassen, wird es deterministisch der ersten Klasse zugeordnet (*Kaufmann und Pape* 1984: 449). Die beim K-Means-Verfahren berechneten Anteilswerte der Klassen vermitteln daher nur eine sehr grobe Vorstellung über die Größe der Klassen, wenn Überlappungen vorliegen.

Bei beiden Verfahren sind zwei Klassen bzw. zwei Cluster schwach besetzt, nämlich die als Anti-Postmaterialisten und Anti-Materialisten bezeichneten Cluster. Dies ist eine Konsequenz des Startwertverfahrens. Das Quick-Clustering führt dazu, daß als Startcluster maximal getrennte Cluster berechnet werden. Welche Ergebnisse bei einem anderen Startwertverfahren erzielt werden, wird zu Ende dieses Abschnitts beschrieben. Wir wollen uns zunächst mit dem Problem der Überlappungen beschäftigen.

Tabelle 4.2.3: Ergebnisse der 5-Klassenlösung der latenten Profilanalyse und des K-Means-Verfahrens für die Wertedaten von Denz (Startwertverfahren für beide Verfahren=Quick-Clustering)

		C1	C2	C3	C4	C5
		Latente Profilanalyse				
Anteilsw.	in %	0.9	14.0	2.3	70.1	12.7
Mittelw.	GMAT	2.58	2.05	4.20	1.83	2.98
	GPMAT	4.08	2.28	1.95	1.49	1.22
Standardabw.	GMAT	0.42	0.44	0.55	0.43	0.38
	GPMAT	0.25	0.37	0.33	0.27	0.20
		K-Means-Verfahren				
Anteilsw.	p(k) in %	0.9	15.1	2.8	55.5	25.7
Mittelw.	GMAT	2.58	2.01	4.22	1.66	2.74
	GPMAT	4.08	2.29	1.83	1.46	1.39
Standardabw.	GMAT	0.42	0.37	0.44	0.31	0.35
	GPMAT	0.25	0.32	0.45	0.25	0.26

GMAT = Gesamtpunktwert für Materialismus, GPMAT = Gesamtpunktwert für Postmaterialismus

Es wurde bereits darauf hingewiesen, daß der Überlappungsanteil entscheidend die Konvergenz und Stabilität der Ergebnisse der latenten Profilanalyse beeinflußt. Eine Grobabschätzung des Überlappungsanteils kann dadurch durchgeführt werden, daß die Zuordnungswahrscheinlichkeiten dichotomisiert und alle Ausprägungskombinationen

berechnet werden. Als Dichotomisierungsschwelle kann man dabei 1/K wählen, also jenen Wert, der sich ergibt, wenn ein Objekt jeder Klasse mit der gleichen Wahrscheinlichkeit angehört. Für die 5-Klassenlösung ergibt sich folgendes Bild (siehe Tabelle 4.2.5):

(...)

4.3 Analyse latenter Klassen für nominalskalierte Variablen

4.3.1 Modellansatz und Algorithmus

Im Unterschied zur latenten Profilanalyse setzt die Analyse latenter Klassen für nominalskalierte Variablen - wie ihr Name sagt - nur nominalskalierte Variablen voraus. Die Modellannahmen sind:

1. Es liegen K latente Klassen (=Muster) mit den Mittelwerten $\pi(k)$ vor.
2. Die Wahrscheinlichkeit, daß in der latenten Klasse k die nominalen Variablen j mit der Ausprägung i auftritt, ist gleich $\pi(i(j)/k)$.
3. Wegen der Annahme der lokalen Unabhängigkeit ist die Auftrittswahrscheinlichkeit des Merkmalsvektors eines Objekts g, wenn die Klasse k vorliegt, gleich:

(4.3.1) $$\pi(g/k) = \pi(x_{g1}/k)\cdot\pi(x_{g2}/k)\cdot K\ \cdot\pi(x_{gm}/k) = \prod_j \pi(x_{gj}/k),$$

wobei x_{gj} der Wert des Objekts g in der nominalen Variablen j ist. $\pi(x_{gj}/k)$ ist die bedingte Wahrscheinlichkeit des Auftretens der Ausprägung des Objekts g in der Variablen j für die Klasse k.

Die Modellparameter sind somit:

1. Die Anteilswerte $\pi(k)$ der latenten Klassen k (k=1,...,K)
2. Die bedingten Auftrittswahrscheinlichkeiten $\pi(j(i)/k)$ für das Auftreten der Ausprägungen i in den nominalen Variablen j in den latenten Klassen k.

Die bedingten Auftrittswahrscheinlichkeiten entsprechen den Klassenzentren der latenten Profilanalyse. Im Unterschied zur latenten Profilanalyse stellen die Klassenvarianzen keine Modellparameter dar, da sie mit $\pi(j(i)/k)\cdot(1-\pi(j(i)/k))$ aus den bedingten Auftrittswahrscheinlichkeiten berechnet werden können. Die Schätzung der Modellparameter erfolgt wiederum über den EM-Algorithmus. Dazu werden die

nominalen Variablen in ihre Dummies aufgelöst. Wir wollen mit $x_{gj(i)}$ den Wert des Objekts g in der i-ten Dummy-Variablen (=Ausprägung) der nominalen Variablen j bezeichnen. $x_{gj(i)}$ ist gleich 1, wenn Objekt g in der nominalen Variablen j die Ausprägung i besitzt, andernfalls 0. Mit $\pi(k/g)$ sollen wiederum die (wahren) Zuordnungswahrscheinlichkeiten der Objekte g zu den Klassen k und mit p(k) und p(j(i)/k) die Schätzwerte der gesuchten Modellparameter bezeichnet werden. Unter Verwendung dieser Notation ergeben sich folgende Schätzgleichungen (*Van de Pol und de Leeuw* 1986)[1] :

(4.3.2) $$p(k) = \sum_g \pi(k/g) \Big/ n.$$

(4.3.3) $$p(j(i)/k) = \sum_g \pi(k/g) \cdot x_{gj(i)} \Big/ \sum_g \pi(k/g).$$

Die beiden Schätzgleichungen (4.3.2) und (4.3.3) sind vollkommen strukturgleich jenen der latenten Profilanalyse (Gleichungen 4.2.9 und 4.2.11). Die Auftrittswahrscheinlichkeiten $\pi(j(i)/k)$ sind die Klassenzentren der Dummies der nominalen Variablen. Im Unterschied zur latenten Profilanalyse werden die Wahrscheinlichkeiten $\pi(x_{gj}/k)$ des Auftretens der Werte des Objektes g in den Variablen j, wenn die Klasse k vorliegt, nicht über die Dichtefunktion der Normalverteilung berechnet, sondern mit:

(4.3.4) $$\pi(x_{gj}/k) = \pi(j(i)/k) \quad \text{für} \quad x_{gj(i)} = 1.$$

Die Wahrscheinlichkeit $\pi(x_{gj}/k)$ ist also gleich der Auftrittswahrscheinlichkeit der Ausprägung i der nominalen Variablen j, die das Objekt besitzt. Hat beispielsweise das Objekt g in der nominalen Variablen j die Ausprägung 1, so ist die Wahrscheinlichkeit $\pi(x_{gi}/k)$ gleich der Auftrittswahrscheinlichkeit der Ausprägung 1 in der Klasse k, also gleich $\pi(j(1)/k)$. Das zugrundeliegende Verteilungsmodell ist das einer Polynomialverteilung (*Fisz* 1980: 195-197).

Die Zuordnungswahrscheinlichkeiten $\pi(k/g)$ werden wie bei der latenten Profilanalyse über den Satz von Bayes bestimmt. Der EM-Algorithmus sieht folgendermaßen aus:

Schritt 1: Berechnung oder Eingabe von Startwerten für die Modellparameter.

Schritt 2: Berechnung der Zuordnungswahrscheinlichkeiten nach Bayes mit[1]

---

[1] *Van de Pol und de Leeuw* (1986) entwickeln die Schätzgleichungen für das allgemeine latente Markovmodell (siehe dazu auch *Langeheine und Van de Pol* 1990). Dieses enthält als Submodell die Analyse latenter Klassen für nominalskalierte Variablen. Die Schätzgleichungen für die Analyse latenter Klassen sind beispielsweise auch in *Andersen* (1991: 426-429) wiedergegeben. *Langeheine* (1988) gibt einen Überblick über neue Ansätze der Analyse latenter Klassen.

[1] Aus Gründen der einfacheren Schreibweise wurde auf den Index für die i-te Iteration verzichtet.

(4.3.5) $$p(k/g) = \frac{p(k) \cdot p(g/k)}{\sum p(k) \cdot p(g/k)},$$

wobei die Wahrscheinlichkeit p(g/k) des Auftretens des Objekts g in der Klasse k entsprechend der Gleichung (4.3.1) geschätzt wird. Die dafür benötigten Auftrittswahrscheinlichkeiten $\pi(x_{gj}/k)$ werden entsprechend Gleichung (4.3.7) berechnet.

Schritt 3: Schätzung der Modellparameter entsprechend den Gleichungen (4.3.2) und (4.3.3). Für Zuordnungswahrscheinlichkeiten $\pi(k/g)$ werden die Schätzwerte p(k/g) aus dem zweiten Schritt verwendet.

Schritt 4: Prüfung der Konvergenz analog der latenten Profilanalyse.

In die Schätzung gehen folgende Nebenbedingungen ein:

(4.3.6) $$\sum_k \pi(k) = 1.$$

(4.3.7) $\pi(k) > 0.$

(4.3.8) $$\sum_i \pi(j(i)/k) = 1 \text{ für alle Variablen i und Klassen k.}$$

Die beiden ersten Nebenbedingungen haben wir bereits bei der latenten Profilanalyse eingeführt. Sie besagen, daß die Summe der Klassenanteilswerte gleich 1 und keine Klasse leer sein soll. Die dritte Nebenbedingung bedeutet, daß die Summe der Auftrittswahrscheinlichkeiten der Ausprägungen einer Variablen in einer Klasse k gleich 1 sein soll.

Unter Berücksichtigung der Nebenbedingungen sind somit bei K Klassen und m nominalen Variablen folgende Modellparameter zu schätzen:

K-1	Klassenanteilswerte $\pi(k)$. Wegen der Nebenbedingung $\Sigma\pi(k) = 1$ ist ein Klassenanteilswert fixiert.
$K \cdot \left(\sum_{j=1}^{m} (m_j - 1) \right)$	Auftrittswahrscheinlichkeiten $\pi(j(i)/k)$. Wegen der Nebenbedingung (4.3.6) sind in jeder Klasse für jede Variable m_j-1 Auftrittswahrscheinlichkeiten zu schätzen (m_j = Zahl der Ausprägungen der Variablen j), insgesamt also die angegebene Zahl.
$K \cdot \left(\sum_j m_j - m + 1 \right) - 1$	Gesamtzahl zu schätzender Parameter

Wird also für drei nominale Variablen mit jeweils drei Ausprägungen eine 2-Klassenlösung gesucht, sind $2 \cdot (\Sigma 3 - 3 + 1) - 1 = 13$ Modellparameter zu schätzen. Diesen stehen $3 \cdot 3 \cdot 3 = 27$ unterschiedliche Datenvektoren gegenüber. Die notwendige Bedingung für ein identifiziertes Modell, daß mehr oder zumindest gleichviele empirische Informationen (=unterschiedliche Merkmalsvektoren) als Modellparameter vorliegen, ist somit erfüllt. Die allgemeine Bedingung lautet: Die Zahl der Ausprägungskombinationen der nomina-

len Variablen muß größer/gleich der Zahl der zu schätzenden Parameter sein. Sind nicht alle Ausprägungskombinationen besetzt, kann das Modell empirisch nicht identifiziert sein. In unserem Beispiel wäre dies der Fall, wenn nur 12 der 27 möglichen Kombinationen empirisch besetzt sind. In diesem Fall macht eine Schätzung keinen Sinn. Vor einer Analyse sollte daher immer geprüft werden, ob die notwendige Identifikationsbedingung erfüllt ist. Praktisch kann diese Bedingung z.B. mit dem Quick-Clustering-Verfahren überprüft werden. Dazu wird die Clusterzahl gleich der Zahl zu schätzender Parameter gesetzt. Wir wollen diese Zahl mit m_K bezeichnen. Findet das Quick-Clustering-Verfahren keine m_K Cluster, ist die notwendige Bedingung der Modellidentifikation nicht erfüllt.

Wir wollen den Algorithmus anhand eines Rechenbeispiels veranschaulichen. Gegeben sind zwei dichotome nominalskalierte Variablen X_1 und X_2 mit jeweils drei Ausprägungen. Es soll eine 2-Klassenlösung berechnet werden. Für den Iterationsschritt (i) sollen folgende Schätzwerte vorliegen:

	C1 (k=1)	C2 (k=2)
p(k)	0.5	0.5
p(1(1)/k)	0.8	0.0
p(1(2)/k)	0.2	0.2
p(1(3)/k)	0.0	0.8
p(2(1)/k)	0.6	0.1
p(2(2)/k)	0.2	0.2
p(2(3)/k)	0.2	0.7

Die Interpretation der Schätzwerte soll exemplarisch für die latente Klasse 1 (=C1) dargestellt werden. Die latente Klasse 1 (k=1) hat einen Anteilswert von 0.5. Die Ausprägung 1 der nominalen Variablen 1 tritt mit einer Wahrscheinlichkeit von 0.8 (=p(1(1)/1) auf, die Ausprägung 2 der nominalen Variablen 1 mit einer Wahrscheinlichkeit von 0.2 und die Ausprägung 3 mit einer Wahrscheinlichkeit von 0. Da die Auftrittswahrscheinlichkeiten als Mittelwerte der Dummies (=Anteilswerte) interpretiert werden können, ist auch folgende Interpretation möglich: In der Klasse 1 tritt mit 80 Prozent die Ausprägung 1 in der Variablen 1 auf und mit 20 Prozent die Ausprägung 2. Für die zweite Variable ist die Auftrittswahrscheinlichkeit der Ausprägung 1 gleich 0.6 (=p(2(1)/1)) usw. An den Schätzwerten lassen sich auch die dargestellten Nebenbedingungen veranschaulichen. Die Summe der Anteilswerte der beiden Klassen ist gleich 1. Da beide Klassen einen Anteilswert von 0.5 haben, ist auch die zweite Nebenbedingung, daß keine Klasse leer ist, erfüllt. Die dritte Nebenbedingung besagt, daß die Summe der Auftrittswahrscheinlichkeiten jeder Variablen in jeder Klasse gleich 1 ist. Auch diese Nebenbedingung ist erfüllt. Für die nominale Variable 1 in der Klasse 1 gilt beispielsweise:

p(1(1)/1) + p(1(2)/1) + p(1(3)/1) = 1,

wie man leicht nachrechnen kann.

Die Berechnung der Zuordnungswahrscheinlichkeiten p(k/g) zeigt Tabelle 4.3.1. Sie soll für das erste Objekt A dargestellt werden. Objekt A besitzt in der Variablen X1 die Ausprägung 1 und in der Variablen X2 ebenfalls die Ausprägung 1. Die Wahrscheinlichkeit für das Auftreten dieses Antwortmusters in der ersten latenten Klasse ist gleich 0.8·0.6 = 0.48, da p(1(1)/1) = 0.8 und p(2(1)/1) = 0.6 ist. Die Auftrittswahrscheinlichkeit des Objekts A für die latente Klasse 2 ist gleich 0.0·0.1 = 0.00.

Der Likelihood-Wert (=PGES in der Tabelle) des Objekts A ist - wie bei der latenten Profilanalyse - gleich 0.24 + 0.00 = 0.24, da p(1) = 0.5, p(2) = 0.5 und p(A/1) = 0.48 und p(A/2) = 0 ist. Daraus ergibt sich ein Log-Likelihood-Wert von -1.4271.

Die Zuordnungswahrscheinlichkeiten für das Objekt A lassen sich über den Satz von Bayes berechnen. Für p(1/A) ergibt sich eine Wert von

$$p(1/A) = \frac{0.5 \cdot 0.48}{0.5 \cdot 0.48 + 0.5 \cdot 0} = 1,$$

da p(A/1) = 0.48 und p(A/2) = 0 ist. Die Auftrittswahrscheinlichkeit der Klasse 1 für das Objekt A ist gleich 1, jene der Klasse 2 gleich 0. Das Objekt A ist also mit einer Wahrscheinlichkeit von 1 der Klasse 1 zugeordnet.

Tabelle 4.3.1: Veranschaulichung der Berechnung der Zuordnungswahrscheinlichkeiten für die Analyse latenter Klassen für nominalskalierte Variablen

Objekt	Variablen		Auftrittswahr. der Objekte			Log-Likelihood-Werte	Zuordnungswahr.	
g	X1	X2	p(g/1)	p(g/2)	PGES	ML	p(g/1)	p(g/2)
A	1	1	0.48	0.00	0.24	-1.4271	1.00	0.00
B	1	2	0.16	0.02	0.16	-1.8326	1.00	0.00
C	2	2	0.04	0.04	0.04	-3.2189	0.50	0.50
D	3	3	0.00	0.56	0.28	-1.2730	0.00	1.00
E	3	2	0.00	0.16	0.08	-2.5257	0.00	1.00
					Σ	-10.2773		

Für die anderen Objekte können die Zuordnungswahrscheinlichkeiten analog berechnet werden. Es ergeben sich die in der Tabelle 4.3.1 dargestellten Werte. Zur Neuberechnung der Modellparameter werden die nominalen Variablen in ihre Dummies aufgelöst und mit den entsprechenden Zuordnungswahrscheinlichkeiten multipliziert. Da beide Variablen drei Ausprägungen haben, wird jede nominale Variable in drei Dummies aufgelöst. Tabelle 4.3.2 zeigt exemplarisch die Berechnung der Modellparameter für die erste nominale Variable in der ersten Klasse. Da das Objekt A in der nominalen Variablen 1 die Ausprägung 1 besitzt, ist der Wert in der Dummy-Variablen $X_{1(1)}$ gleich 1. Dieser wird mit der Zuordnungswahrscheinlichkeit des Objekts A zur ersten Klasse (=1) multipliziert. Analog wird für die anderen Dummies und die weiteren Objekte vorgegangen. Berechnet man die Spaltensumme der Zuordnungswahrscheinlichkeiten für die Klasse 1 (=Spalte "p(g/1)"), ergibt sich die mit den Zuordnungswahrscheinlichkeiten gewichtete Fallzahl der Klasse 1. Der Spaltensummenwert ist gleich 2.5. Division mit der Fallzahl ergibt entsprechend Gleichung (4.3.2) den Anteilswert der Klasse 1. In unserem Beispiel ist dieser gleich 0.5. Bilden wir die Spaltensummen der mit den Zuordnungswahrscheinlichkeiten multiplizierten (gewichteten) Dummies und dividieren diese mit der gewichteten Fallzahl, erhalten wir entsprechend Gleichung (4.3.3) die Schätzwerte für die Auftrittswahrscheinlichkeiten der Ausprägungen der nominalen Variablen 1 in der Klasse 1. Es ergeben sich Werte von 0.8, 0.2 und 0.0. Sie sind also mit den Werten der vorausgehenden Iteration identisch. Dies gilt auch für die anderen Modellparameter.

Damit dürfte das Grundprinzip der Schätzung der Modellparameter hinlänglich verdeutlicht worden sein. *Abschließend ist auf eine Besonderheit des Algorithmus hinzuweisen:* Besitzt ein Schätzwert für eine Auftrittswahrscheinlichkeit p(i(j)/k) den Wert 0 oder 1, wird er während der Iteration nicht mehr geändert. Bei der Eingabe von Startwerten ist also darauf zu achten, daß keine Werte von 0 oder 1 eingegeben werden, da diese dann nicht mehr geändert werden. Für die Auswahl eines Startwertverfahrens

bedeutet dies, daß Startwerte aus dem Quick-Clustering- oder Repräsentanten-Verfahren ausscheiden, da sie ein typisches Objekt für jede Klasse auswählen. Da bei einem Objekt nur eine Ausprägung auftreten kann, ist die Auftrittswahrscheinlichkeit in dieser Ausprägung gleich 1, alle anderen Auftrittswahrscheinlichkeiten gleich 0.

Tabelle 4.3.2: Neuberechnung der Modellparameter für die erste Variable in der ersten Klasse

	Variablen		Zuordnungswahr.		Dummies multipliziert mit Zuordnungswahrscheinlichkeiten			
g	X1	X2	p(g/1)	p(g/2)	X1(1)	X1(2)	X1(3)	usw.
A	1	1	1.00	0.00	1·1.00	0·1.00	0·1.00	
B	1	2	1.00	0.00	1·1.00	0·1.00	0·1.00	
C	2	2	0.50	0.50	0·0.50	1·0.50	0·0.50	
D	3	3	0.00	1.00	0·0.00	0·0.00	0·0.00	
E	3	2	0.00	1.00	0·0.00	0·0.00	0·0.00	
			2.50		2.00	0.50	0.00	
			p(1)= 2.50/5 =0.50		p(1(1)/1)= 2.00/2.5= 0.8	p(1(2)/1)= 0.50/2.5= 0.2	p(1(3)/1)= 0.00/2.5= 0.0	

4.3.2 Modellprüfung und Interpretation

Es können die für die latente Profilanalyse entwickelten Modellprüfgrößen verwendet werden. Ihre Anwendung soll anhand der Analyse der Freizeitaktivitäten von Kindern dargestellt werden (siehe Abschnitte 2.3, 2.4.2, 2.4.3, 3.5). Für eine erste Analyse wurde angenommen, daß vier bis acht latente Klassen vorhanden sind. Die entsprechenden Testgrößen zeigt die Tabelle 4.3.3.

Tabelle 4.3.3: Ergebnisse der Analyse latenter Klassen für die Freizeitaktivitäten von Kindern

Klassenzahl	Wert der Log-Likelihood-Funktion	prozentuelle Verbesserung gegenüber Nullmodell in %	Informationsmaß von AKAIKE	Signifikanz der LQ-Statistik von Wolfe in %	prozentuelle Verbesserung geg. vorausgehender Lösung in %
K	L_K	$PV0_K$	IA_K	LQ_K(Wolfe)	PV_K
1(a)	-32024.300	0.000	- (b)	- (b)	- (b)
4	-30166.865	5.800	-30249.865	- (c)	- (c)
5	-30090.713	6.038	-30194.713	100	25.2
6	-29984.312	6.370	-30109.312	100	35.4
7	-29956.915	6.456	-30102.915	93	9.1
8	-29933.790	6.528	-30100.790	75	7.7

(a) unser Programm berechnet automatisch immer aus Modelltestgründen die 1-Klassenlösung,

(b) nicht definiert, (c) diese Werte werden nicht berechnet, da keine vorausgehende 3-Klassenlösung vorliegt.

Betrachten wir zunächst die prozentuellen Verbesserungen PV_K gegenüber der jeweils vorausgehenden Lösung, so zeigt sich, daß die 7- und 8-Klassenlösung kaum mehr den Wert der Log-Likelihood-Funktion der vorausgehenden Lösung verbessern: Die 6-Klassenlösung verbessert die 5-Klassenlösung nur mehr um 9.1 Prozent, die 8-Klassenlösung die 7-Klassenlösung um 7.7 Prozent. Diese Zunahmen in der Log-Likelihood-Funktion sind auch nicht mehr signifikant. Wir würden uns daher für die 6-Klassenlösung entscheiden. Der maximale Wert des Informationsmaßes von Akaike liegt allerdings für die 8-Klassenlösung vor. Mitunter ist somit eine weitere Analyse mit mehr Klassen erforderlich, um das tatsächliche Maximum zu bestimmen.

Wir wollen hier aber die 6-Klassenlösung weiter beschreiben. Die Ergebnisse sind in der Tabelle 4.3.4 dargestellt. Da sehr viele Variablen untersucht werden, wird man zur Erleichterung der Interpretation untersuchen, ob in den einzelnen Klassen Variablengruppen gebildet werden können. Technisch bedeutet dies, daß der in Abschnitt 3.7.3 dargestellte Algorithmus zur Bildung von Variablengruppen eingesetzt wird. Dabei ergeben sich die in der Tabelle 4.3.5 dargestellten Variablengruppen.

Die letzten drei Klassen lassen sich relativ einfach interpretieren: Klasse 4 ist der Typus der sehr aktiven Kinder, Klasse 5 der Typus der inaktiven oder deprivierten Kinder und Klasse 6 der Typus der Mittelaktiven, wobei Fernsehen und Radfahren überzufällig häufig auftreten.

Die Klassen 1 bis 3 haben dagegen ein selektives Freizeitmuster. Klasse 3 ist zum einen stark spielorientiert (alleine Spielen, mit Freunden spielen, Computerspiele). Zum anderen wird Musik gehört, mit der Familie etwas unternommen sowie Sport betrieben und Rad gefahren. Wir können uns darunter einen Typus von Kindern vorstellen, für den das Spielen im Vordergrund steht.

Die Klasse 2 dagegen ist dadurch gekennzeichnet, daß nur drei Freizeitaktivitäten häufig ausgeübt werden. Der Typus läßt sich wie folgt beschreiben: Die Kinder sind zum einen in der Wohnung sehr zurückgezogen. Sie lesen in der Freizeit ein Buch, im Freien spielen sie mit Freunden oder fahren Rad. Wir wollen diesen Typus als eher zurückgezogenen Freizeittypus bezeichnen.

Die Klasse 1 ist im Unterschied zur Klasse 2 durch eine stärkere Familienorientierung gekennzeichnet. Mit der Familie wird etwas gemeinsam unternommen, u.a. ein Spaziergang, oder es wird gemeinsam gebastelt oder gemeinsam ferngesehen. Daneben wird noch ein Buch gelesen, radgefahren, Musik gehört, Sport betrieben und alleine oder mit Freunden gemeinsam gespielt. Ein Name für diesen Typus läßt sich nur schwer finden, unter Umständen könnte er als familienorientierter Typus bezeichnet werden.

Zusammenfassend läßt sich somit festhalten, daß zwar eine inhaltliche Interpretation möglich ist, daß man diese aber in einer konfirmatorischen Analyse zu verbessern versuchen wird.

Tabelle 4.3.4: 6-Klassenlösung bei einer Analyse der Freizeitaktivitäten von Kindern (n=2745)

		C1	C2	C3	C4	C5	C6
Anteilsw.	p(k)	0.283	0.188	0.157	0.063	0.095	0.213
Ausruhen	ja	0.55	0.18	0.36	0.72	0.08	0.32
Ausruhen	nein	0.45	0.82	0.64	0.28	0.92	0.68
Freunde	ja	0.93	0.79	0.89	0.92	0.49	0.66
Freunde	nein	0.07	0.21	0.11	0.08	0.51	0.34
Familie	ja	0.71	0.53	0.74	0.89	0.25	0.39
Familie	nein	0.29	0.47	0.26	0.11	0.75	0.61
Basteln	ja	0.67	0.33	0.35	0.91	0.06	0.21
Basteln	nein	0.33	0.67	0.65	0.09	0.94	0.79
Comics	ja	0.61	0.21	0.56	0.86	0.11	0.47
Comics	nein	0.39	0.79	0.44	0.14	0.89	0.53
Musizieren	ja	0.38	0.36	0.20	0.63	0.06	0.06
Musizieren	nein	0.62	0.64	0.80	0.37	0.94	0.94
Haustiere	ja	0.61	0.61	0.61	0.78	0.19	0.48
Haustiere	nein	0.39	0.39	0.39	0.22	0.81	0.52
Kino	ja	0.04	0.04	0.21	0.51	0.05	0.08
Kino	nein	0.96	0.96	0.79	0.49	0.95	0.92
Konzert	ja	0.18	0.17	0.22	0.56	0.05	0.04
Konzert	nein	0.82	0.83	0.78	0.44	0.95	0.96
Musikhören	ja	0.93	0.63	0.81	0.97	0.23	0.65
Musikhören	nein	0.07	0.37	0.19	0.03	0.77	0.35
Kirche	ja	0.55	0.66	0.38	0.73	0.27	0.29
Kirche	nein	0.45	0.34	0.62	0.27	0.73	0.71
Fernsehen	ja	0.84	0.47	0.93	0.97	0.25	0.84
Fernsehen	nein	0.16	0.53	0.07	0.03	0.75	0.16
Computer	ja	0.31	0.09	0.78	0.80	0.14	0.48
Computer	nein	0.69	0.91	0.22	0.20	0.86	0.52
Buch	ja	0.92	0.79	0.62	0.94	0.21	0.52
Buch	nein	0.08	0.21	0.38	0.06	0.79	0.48
Vereinsv.	ja	0.10	0.07	0.30	0.56	0.07	0.03
Vereinsv.	nein	0.90	0.93	0.70	0.44	0.93	0.97
Radfahren	ja	0.92	0.88	0.93	1.00	0.42	0.85
Radfahren	nein	0.08	0.12	0.07	0.00	0.58	0.15
Spazieren	ja	0.88	0.56	0.55	0.98	0.18	0.38
Spazieren	nein	0.12	0.44	0.45	0.02	0.82	0.62
alleine Sp.	ja	0.81	0.44	0.75	0.96	0.10	0.57
alleine Sp.	nein	0.19	0.56	0.25	0.04	0.90	0.43
Parties	ja	0.18	0.12	0.30	0.70	0.05	0.09
Parties	nein	0.82	0.88	0.70	0.30	0.95	0.91
Sport	ja	0.69	0.57	0.92	0.91	0.33	0.48
Sport	nein	0.31	0.43	0.08	0.09	0.67	0.52

Tabelle 4.3.5: Variablengruppen in den 6 Klassen

Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Klasse 6
Ausruhen, Comics, Haustiere, Kirche (0.58)	Basteln, Fernsehen (0.41)	Musizieren, Kino, Konzert, Vereinsv., Parties (0.27)	Ausruhen, Musizieren, Haustiere, Kino, Konzert, Kirche, Computer, Vereinsv., Parties (0.67)	Ausruhen, Basteln, Comics, Musizieren, Kino, Konzert, Computer, alleine Sp., Parties (0.09)	Buch, Sport, Comics, Haustiere, alleine Sp. (0.50)
Musizieren (0.42)	Familie, Haustiere, Musikhören, Kirche, Spaziereng. (0.59)	Familie, Musikhören, Computer, alleine Sp. (0.73)	Freunde, Familie, Basteln, Comics, Musik, Fernsehen, Buch, Radfahren, Spazieren, Sport (0.94)	Familie Haustiere, Musik, Kirche, Fernsehen, Buch, Spazieren, Sport (0.28)	Freunde (0.67)
Freunde, Musikhören, Buch, Radfahren, Spazieren (0.91)	Musizieren, Comics, Konzerte (0.19)	Comics, Haustiere, Buch, Spazieren (0.59)		Freunde, Radfahren (0.46)	Ausruhen, Kirche, Spazieren, Computer (0.33)
Vereinsv. (0.09)	Freunde, Buch (0.81)	Ausruhen, Basteln, Kirche (0.41)			Musizieren, Kino, Konzert, Vereinsv., Parties (0.05)
Computer (0.30)	Computer, Vereinsv., Parties (0.10)	Freunde, Fernsehen, Radfahren, Sport (0.91)			Basteln (0.18)
Fernsehen, alleine Sp. (0.83)	Radfahren (0.90)				Fernsehen, Radfahren (0.82)
Konzert, Parties (0.19)	Kino (0.04)				
Kino (0.04)	alleine Sp. (0.44)				
	Sport (0.57)				