Über die Aufdeckung von Ernährungsmustern in epidemiologischen Studien

Hans-Peter Altenburg
Deutsches Krebsforschungszentrum Heidelberg
Abt. Klinische Epidemiologie / C020
69009 Heidelberg
hp.altenburg@dkfz.de

Zusammenfassung

Analysen von Ernährungsmustern sollen helfen, Kombinationen von Nahrungsmitteln und Inhaltsstoffen zu identifizieren, die in einer bestimmten Population vorkommen können. Es soll gezeigt werden, wie mit Hilfe des SAS Systems solche Muster in einer ernährungs-epidemiologischen Studie gefunden werden können. Verschiedene explorative Ansätze für eine Mustererkennung werden dabei gegenübergestellt: Variablen orientierte, Subjekt orientierte bzw. Ergebnis orientierte Methoden. Traditionell werden hierfür Verfahren der Cluster- bzw. der Faktoranalyse verwendet. Das SAS System stellt hierfür eine ganze Reihe von Prozeduren zur Verfügung wie z.B. CLUSTER, FASTCLUS, VARCLUS, TREE, FACTOR oder PRINCOMP. Es wird dabei gezeigt, wie ein spezielles Problem aus mehreren, unterschiedlichen Blickwinkeln angegangen werden kann. Die Vorgehensweise wird dabei an Hand eines speziellen Beispiels aus einer Subkohorte der EPIC-Studie demonstriert.

.

Keywords: Cluster-Analyse, Dendogramm, Hauptkomponentenanalyse, Faktoranalyse.

1 Einleitung

Ernährungsgewohnheiten spielen eine bedeutsame Rolle bei der Untersuchung von Zusammenhängen mit der Entstehung von Krebs sowie anderen chronischen Erkrankungen. Hierbei sind nicht nur Einzelfaktoren wichtig, sondern auch Kombinationen verschiedener Nahrungsmittel und Einnahmemengen. Beispielsweise hat Obst eine protektive Wirkung bei der Entstehung von Lungenkrebs, jedoch ist diese Wirkung größer bei Rauchern als bei Nichtrauchern. Die Ursache liegt in diesem Beispiel in den unterschiedlichen Verzehrgewohnheiten zwischen Rauchern und Nichtrauchern. Ernährungsmuster können helfen sowohl Erkrankungsrisiken zu identifizieren als auch Zusatzinformationen über die Personen liefern oder zwischen unterschiedlichen Zentren bzw. Erkrankungen diskriminieren helfen.
Im folgenden sollen einige der methodischen Probleme sowie die Realisierung bei der Erkennung von Mustern mit Hilfe des SAS Systems diskutiert werden. Methoden zur Mustererkennung lassen sich in drei Hauptgruppen unterscheiden:

- Subjekt-orientierte Verfahren
  - Cluster-Analyse
- Objekt-orientierte Verfahren
  - Faktor-Analyse oder Hauptkomponentenanalyse
- Outcome-orientierte Verfahren.

Im Folgenden soll aber nur auf die ersten beiden Methodengruppen und entsprechende SAS-Prozeduren zu ihrer Realisierung eingegangen werden. In der Ernährungsepidemiologie werden meist nur Cluster- oder Faktoranalyseverfahren verwendet, je nachdem, ob Subjekt- oder Variablen-orientiert bzw. beide kombiniert vorgegangen werden soll.

2 Subjekt-orientierte Verfahren

2.1 Cluster-Analyse

Mit einer Cluster-Analyse werden Gruppen von Individuen gesucht, die sich bezüglich dem, was sie essen ähnlich sind. „Ernährungsmuster“ sind dann hauptsächlich Beschreibungen dieser Gruppen. In der Regel besteht die Datenmatrix aus p Variablen (Spalten), die an n Individuen (Zeilen) beobachtet wurden (n×p-Matrix: X). Die Aufgabe besteht dann darin, die Spalten der Datenmatrix geeignet zu ordnen („clustern“). In einem ersten Schritt wird also die Datenmatrix transformiert. Hierzu wird eine Transformation durchgeführt und eine Abstands- bzw. Ähnlichkeitsmatrix D, berechnet. Als Ähnlichkeits- oder Unähnlichkeitsmaß (Similarity, Dissimilarity) kann z.B. der Euklidische Abstand oder ein anderes Maß verwendet werden. Bei diskreten oder dichotomen Variablen wird der Anteil der Individuen verwendet, für welche die Variablen / Attribute sich unterscheiden. Es gibt eine große Anzahl (mehr als 40) von Varianten für diese Kennzahlen (siehe Hubálek (1982) für dichotome Variablen oder auch Jäger et al. (2001)). Bei stetigen Merkmalen wird der Korrelationskoeffizient r benutzt, z.B. Ähnlichkeit: r^2, $|r|$ oder Unähnlichkeit: $1-r^2$, $1-|r|$. In der SAS Sample Library existiert auch ein SAS-Macro %DISTANCE, was Abstandsmaße berechnet. Ein wichtiger zu beachtender Punkt wäre auch eine evtl. Standardisierung der Daten durchzuführen bevor die Ähnlichkeitsmatrix bestimmt wird. Details hierzu findet man in Everitt (1993).

Hierarchische Klassifikation

Die hierarchische Klassifikation ist das gängigste Clustering-Verfahren. Hierbei werden die Daten sukzessive in Gruppen / Cluster zusammengefasst: P_n, P_{n-1}, ... , P_1, wobei P_n aus n einzelnen Clustern, und P_1 aus einem Cluster mit allen n Individuen besteht. Es gibt quasi eine aufsteigende Ordnung von Unähnlichkeit. Die grafische Darstellung erfolgt in einem Dendogramm.

Das folgende Beispiel zeigt ein SAS-Programm für eine Cluster-Analyse mit Hilfe der Prozedur CLUSTER mit einer anschließenden Erzeugung eines Dendogrammes (Prozedur TREE):

Beispiel-Programm (Clusteranalyse):

```
%LET dset=lung.ransampl_s ;
PROC CLUSTER DATA=&dset OUTTREE=mv_t METHOD=WARD ;
VAR meat_gr vegetgr ;
ID cntr_c ;
RUN ;
PROC TREE DATA=mv_t ;
ID cntr_c ;
RUN ;
```

Bei der WARD-Methode wird die Summe der Abstände innerhalb der Cluster berechnet. In jedem Schritt sind die beiden zusammengeführten Cluster diejenigen, die zum kleinsten Anwachsen der Gesamtsumme der quadrierten „within"-Cluster-Abstände führen. Sie versucht die Intra-Cluster-Varianz und damit die Cluster-Separabilität zu minimieren. Die Methode produziert in der Regel gute Lösungen, obwohl sie dazu neigt kleinere Cluster zu erzeugen.

Wenn eine Ähnlichkeitsmatrix bereits vorliegt, so kann diese auch direkt zur Eingabe in PROC CLUSTER genutzt werden (siehe hierzu auch das SAS-Macro von Jäger et al. (2001). Bei großen Datenmengen, wie dies auch bei Auswertungen ernährungs-epidemiologischer Studien aus dem Umfeld der EPIC-Studie der Fall sein dürfte, kann es bei der direkten Anwendung der Prozedur CLUSTER zu Problemen kommen, da die benötigte Zeit proportional zum Quadrat (oder sogar eine noch höhere Potenz) des Stichprobenumfanges ist. Als Ausweg sei hier die Verwendung der Prozedur FASTCLUS empfohlen, um initiale Cluster zu finden (z.B. 50). Anschließend wird dann die Prozedur CLUSTER auf diese 50 Cluster angewandt. Hierbei ist jedoch zu beachten, dass FASTCLUS das „k-Means-Modell" verwendet. Vorsicht ist auch bei vorkommenden Ausreißern angebracht.

Hierarchische Cluster-Algorithmen führen nicht direkt zu einer Zerlegung in Klassen. Vielmehr erhält man eine „globale Visualisierung" der Struktur der Unähnlichkeiten, um etwa die Anzahl der Cluster in den Daten vorzuschlagen. Eine Partitionierung erfordert das Abschneiden des (Dendogramm-) Baumes an einer bestimmten Höhe, d.h. man muss das Dendogramm ansehen, um die Anzahl von Zeilen-Pattern der Datenmatrix zu finden.

Auch eine Standardisierung kann zu Problemen führen: Wenn im Vorfeld der Clustertyp bekannt ist, kann eine bestimmte Standardisierung vorgeschlagen werden, was zu einer Art von „circulus vitiosus" führt:

(1) Um eine Cluster-Analyse durchführen zu können, wird ein geeignetes Abstandsmaß benötigt.

(2) Das geeignete Abstandsmaß hängt ab von den Charakteristiken der Standardisierungsmethode.

(3) Um eine Standardisierungsmethode auszuwählen, wird der Cluster-Typ benötigt.

Methodisch gibt es aus diesem Kreis auszubrechen nur die „Trial-and Error-Methode", wobei mit verschiedenen Alternativen und unterschiedlichen

Evaluierungen der Lösungen experimentiert wird (Visuelle Inspektion, Dateninterpretation und Nützlichkeitsbetrachtungen).

Varianten der Standard-Cluster-Analyse-Verfahren
Für eine Cluster-Analyse sind auch Spalten-orientierte Algorithmen denkbar und in SAS realisiert. Die Prozedur VARCLUS mit der Option CENTROID erlaubt eine solche spalten-orientierte Methode. Eine weitere alternative Cluster-Methode wäre das Block-Muster („Block-Pattern“)-Verfahren: Sukzessives Anwenden der Prozeduren CLUSTER und VARCLUS.
Beim Optimum Partitioning Algorithmus werden eingegeben, die Definition von Unähnlichkeit und die Anzahl der gewünschten Cluster. Ziel ist es die Partitionierung in k Cluster, derart vorzunehmen, dass der Quotient von inter- / intra-Cluster-Variation maximal wird. In der Regel wird hier der k-Means-Algorithmus verwendet. Es existieren zahlreiche Varianten dieser Methode.
Conceptual Clustering verwendet ausgewählte Variablen, um Partitionen von Subjekten zu definieren. So werden etwa bei Gower‘s Predictive Clustering (Gower 1974) die Anzahl von Variablen ausgewählt. Bestimmte Werte dieser Variablen definieren die Cluster der Subjekte und als ausgewählte Variablen sind etwa diejenigen geeignet, welche am besten die anderen vorhersagen.

3 Objekt- oder Variablen-orientierte Verfahren

3.1 Variablenreduktion

Variablenorientierte Verfahren werden benutzt, um latente Strukturen in einer Datenbasis zu entdecken. Hierbei wird der Attributeraum von einer großen Anzahl von Variablen in eine geringe Zahl von Faktoren reduziert ohne dass dabei eine „abhängige“ (oder als von weiteren „unabhängigen Variablen“ abhängige) Variable erforderlich ist. Hierzu werden numerische Assoziationsmaße zur Identifizierung allgemeiner Tendenzen in der Abhängigkeitsstruktur verwendet. Ernährungsmuster sind dabei jede begründbare Beschreibung solcher Tendenzen. Das Ziel ist dann die Datenrepräsentation in einem Raum mit reduzierter Dimension. Die wichtigsten statistischen Verfahren oder Techniken um die Variablenanzahl zu reduzieren, sind Hauptkomponenten- (PCA) oder Faktorenanalyse (FA). Bei einer Spalten-Orientierung spricht man von einer R-FA, bei Zeilen-Orientierung von Q-FA. Ziel beider Methoden ist es, die Originalvariablen als eine Linearkombination von neuen „Variablen“, den Faktoren oder mehr spezifisch den Hauptkomponenten zu approximieren. Sie liefern heuristische Kriterien, um Daten durch eine kleine Zahl bedeutungsvoller Faktoren zu präsentieren. Das SAS-System stellt hierfür im Modul SAS/STAT zwei SAS-Prozeduren zur Verfügung: PROC PRINCOMP oder PROC FACTOR. Die Vorgehensweise sei im folgenden aber lediglich nur an Hand der Prozedur FACTOR dargestellt. Auch wenn wir in den folgenden Beispielprogramm für beide Methoden nur die Prozedur FACTOR verwenden, sei nochmals darauf hingewiesen, dass eine Hauptkomponentenanalyse keine Faktoranalyse ist. Beide Verfahren sind

lediglich Methoden zur Reduktion der Variablenanzahl und tendieren dazu empirisch zusammen zu hängen. Der wichtigste Unterschied liegt in der Annahme der zugrunde liegenden kausalen Struktur. Die Faktoranalyse nimmt an, dass die Kovariation in den Beobachtungsvariablen von der Anwesenheit von einer oder mehreren latenten Variablen abhängt, die die Beobachtungsvariablen beeinflussen. Diese latenten Faktoren werden zwar als existent angenommen, können aber nicht direkt gemessen werden. Eine explorative Faktoranalyse hilft einem Wissenschaftler die Anzahl und Struktur von latenten Variablen zu identifizieren.
Im Gegensatz dazu macht eine Hauptkomponentenanalyse keine Annahmen über ein zugrundeliegendes kausales Modell. Sie ist lediglich ein Variablenreduktionsverfahren, das dem größten Varianzanteil in einer Menge von Beobachtungsvariablen Rechnung trägt.

3.2 Hauptkomponenten-Analyse

Die Hauptkomponentenanalyse stellt eine Linearkombination (LK) der Originalvariablen ohne spezielle Interpretation dar. Sie ist im Prinzip nur eine Linearkombination von optimal gewichteten Beobachtungsvariablen. So erklärt z.B. die erste Hauptkomponente den größten Betrag der totalen Variation (totale Variation aus Kovarianzmatrix / Distanz zwischen den Variablen). Die Hauptkomponenten werden aus der Datenmatrix extrahiert in der Ordnung der erklärten Variablen. Eine Entscheidungsregel könnte dann etwa so aussehen: Wähle die k Hauptkomponenten, die einen bestimmten Anteil (etwa 80%) der totalen Variation erklären. Normalerweise ist k deutlich kleiner als die Anzahl der Originalvariablen.
Eine Hauptkomponentenanalyse wird durchgeführt auf der Basis einer Matrix von Pearson’schen Korrelationskoeffizienten. Die Daten müssen deshalb Annahmen für diese Kennzahl erfüllen, wie z.B. Intervall-skalierte und normal verteilte Variablen auf der Grundlage einer Zufallsstichprobe, lineare Beziehungen zwischen den Variablen sowie das Vorliegen einer bivariaten Normalverteilung bei jedem Paar der Beobachtungsvariablen.

Beispiel-Programm (Hauptkomponentenanalyse):

```
%LET dset=dsetname ;
%LET varlist= variablenliste ;
PROC FACTOR DATA=&dset
        METHOD=PRIN
        MINEIGEN=1.
        SCREE
        ROTATE=VARIMAX
        ROUND
        FLAG=.35 ;
VAR &varlist ;
RUN ;
```

Dieses Beispielprogramm enthält alle für die Standardanwendung einer Hauptkomponentenanalyse wesentlichen Optionen und Kriterien um die bedeutsamen Komponenten aus einer Datenbasis zu ermitteln. Diese sind das Eigenwertkriterium (siehe SAS-OUTPUT), der Scree-Test, welcher die Eigenwerte grafisch gegen die Nummer in der Reihenfolge aufträgt (man suche nach dem „Break“ im Schaubild(!)), den Varianzanteil (siehe Ausgabe der Prozedur) und schließlich die Interpretierbarkeit der gefundenen Hauptkomponenten. Ein Beispieldatensatz mit acht Variablen (Meat, Fish, Potatoes, Vegetables, Leafy Vegetables, Legumes, Eggs, und Fruits) liefert dann für das oben angegebene Programm mit einer Subkohorte der EPIC-Studie folgende Ausgabe:

SAS-OUTPUT:

Teil 1: Tabelle der Eigenwerte

```
Principal Component Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE


Eigenvalues of the Correlation Matrix: Total = 8  Average = 1

        Eigenvalue    Difference    Proportion    Cumulative

   1    2.16789226    0.67389651        0.2710        0.2710
   2    1.49399575    0.57058536        0.1867        0.4577
   3    0.92341039    0.04552879        0.1154        0.5732
   4    0.87788160    0.05674065        0.1097        0.6829
   5    0.82114095    0.12783964        0.1026        0.7855
   6    0.69330131    0.03214743        0.0867        0.8722
   7    0.66115388    0.29993003        0.0826        0.9548
   8    0.36122385                      0.0452        1.0000

2 factors will be retained by the MINEIGEN criterion.
```

Teil 2, den Scree-Plot lassen wir hier weg, da er hier keine Zusatzinformationen liefert. Die beiden wichtigen Hauptkomponenten lassen sich bereits aus dem ersten Teil erkennen. Wichtiger ist dagegen Teil 3, wo nach einer orthogonalen Rotation (zur Erzeugung unkorrelierter Bestandteile) die bedeutsamen Hauptkomponenten (jetzt als Faktoren bezeichnet) mit den ihnen zugeordneten Beobachtungsvariablen aufgelistet sind:

Teil 3: Rotiertes Faktormuster

```
Principal Component Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components

                          Factor Pattern

                                                 Factor1      Factor2

QG01      POTATOES AND OTHER TUBERS                -28           57 *
QG02      VEGETABLES                                82 *         -6
QG0201    LEAFY VEGETABLES (EXCEPT CABBAGES)        80 *          5
QG03      LEGUMES                                   55 *          5
QG04      FRUITS                                    62 *        -22
QG07      MEAT AND MEAT PRODUCTS                     7           74 *
QG08      FISH AND SHELLFISH                        17           50 *
QG09      EGGS AND EGG PRODUCTS                     25           57 *

Printed values are multiplied by 100 and rounded to the nearest
integer.  Values greater than 0.4 are flagged by an '*'.
```

Die Anzahl der wichtigen Komponenten (bei gleicher Parameterkonstellation) kann sich ändern, wenn die Beobachtungsvariablen geändert werden. Fügt man beispielsweise noch eine weitere Variable (nämlich Fat) hinzu, so ergeben sich auf einmal drei bedeutsame Hauptkomponenten. Es sei nur Teil 3 der Ausgabe aufgelistet:

Teil 3: Rotiertes Faktormuster bei zusätzlicher Variable Fat

```
Initial Factor Method: Principal Components

                              Factor Pattern

                                             Factor1      Factor2      Factor3

QG01      POTATOES AND OTHER TUBERS            -21           66 *         31
QG02      VEGETABLES                            82 *        -10           13
QG0201    LEAFY VEGETABLES (EXCEPT CABBAGES)    80 *         -4           -7
QG03      LEGUMES                               54 *         -2          -10
QG04      FRUITS                                61 *        -22           23
QG07      MEAT AND MEAT PRODUCTS                13           73 *        -5
QG08      FISH AND SHELLFISH                    18           39          -42 *
QG09      EGGS AND EGG PRODUCTS                 27           44 *        -56 *
QG10      FAT                                   25           48 *         64 *

Printed values are multiplied by 100 and rounded to the nearest integer.  Values
greater than 0.4 are flagged by an '*'.
```

Die Anzahl der extrahierten Hauptkomponenten ist stets gleich der Anzahl der analysierten Variablen. Die erste Komponente zeigt den größten Anteil der totalen Variation auf, die zweiten Komponente den zweitgrößten Anteil usw. Jedoch spielen nur ein paar wenige der ersten Komponenten eine bedeutsame Rolle für die Interpretation. Das „Eigenwert≥1“-Kriterium oder der SCREE-Test sind lediglich Hilfsmittel zur Identifikation der wichtigsten Komponenten. Orthogonale Rotationen der Datenstruktur werden durchgeführt, um unkorrelierte Komponenten zu erhalten. Bei der Interpretierbarkeit sollte darauf geachtet werden, dass mindestens drei Variablen einen Beitrag für eine Komponente liefern, und dass die Variablen einer Komponente eine

vergleichbare konzeptionelle Bedeutung besitzen. Wichtig wäre auch, dass unterschiedliche Komponenten verschiedenes Messen und ein rotiertes Faktorenmuster eine einfachere Struktur aufweist.
Wenn eine Analyse vollständig durchgeführt wurde, ist es oft wünschenswert zu wissen, wo die einzelnen Probanden bzgl. ihrer bedeutsamen Komponenten stehen. Hier kann ein Faktor-Score oder ein Faktor-basierter Score gute Dienste leisten: Ein Faktor-Score ist eine Linearkombination der optimal gewichteten Beobachtungsvariablen und kann ganz einfach mit Hilfe den SAS-Optionen NFACT=*k* OUT=*dset2* erhalten werden. Die Konstante *k* gibt dabei die gewünschte Anzahl von Hauptkomponenten an. Die neu erzeugte Datentabelle *dset2* enthält dann *k* neue Variablen mit diesem Faktor-Score (mit Namen factor1, factor2, ..., factor*k*). Ein Faktor basierter Score dagegen ist nur eine Linearkombination der für eine Komponente bedeutsamen Beobachtungsvariablen. Er erfordert zu seiner Erstellung einen neuen DATA-Step mit den entsprechenden Statements zur Bildung der Score-Summen.

3.3 Faktoren-Analyse

Faktorenanalyseverfahren versuchen ebenfalls die hinter einer großen Anzahl von Variablen stehende Struktur (die sog. Latenten Variablen oder Common Faktoren) zu entdecken. Die Faktorladungen werden dazu benutzt, um die Faktorenstruktur zu beschreiben. Common Faktoren (CF) sind unbeobachtbare Variablen. Jede Originalvariable ist eine Linearkombination von CFs gestört durch einen unkorrelierten Fehlerterm. Die Koeffizienten werden Faktorladungen genannt und messen die Assoziation der Variablen mit den Common Faktoren. Eine Hauptkomponente ist eine künstliche Variable, die eine Linearkombination der (optimal gewichteten) Beobachtungsvariablen darstellt. Über einen entsprechenden Scorewert läßt sich für jeden Probanden (bzw. jedes Subjekt) bestimmen, wo er hinsichtlich der Beobachtungsvariablen steht. Dagegen ist ein Common Faktor nur eine hypothetische latente Variable von der angenommen wird, dass sie für die Kovariation zwischen zwei oder mehren Beobachtungsvariablen verantwortlich ist. Da diese Faktoren nicht messbare latente Variablen sind, weiss man nie genau, wo ein bestimmtes Subjekt hinsichtlich des Faktors steht. Die Common Faktoren sind keine Linearkombinationen der Beobachtungsvariablen (wie bei der Hauptkomponentenanalyse). Die Faktorenanalyse setzt dagegen ein gegensätzliches Modell voraus, nämlich dass die Beobachtungsvariablen eine Linearkombination der latenten Variablen darstellen.
Ein weiterer Unterschied zwischen beiden Verfahren liegt in der betrachteten Varianz. Die totale Variation einer Datenstruktur setzt sich zusammen aus einer „common“, gemeinsamen Varianzkomponente und einer eindeutigen Varianzkomponente, nämlich der Varianz, die eindeutig einer bestimmten Beobachtungsvariablen zugeordnet werden kann. Während die Faktorenanalyse sich auf die gemeinsame Varianz bezieht, betrachtet die Hauptkomponentenanalyse die totale Varianz.

Die Nachteile dieses Konzeptes sind, dass die Existenz der latenten Variablen vorausgesetzt wird, d.h. sie sind u.U. „künstlich“, sie können nur als Linearkombination der Originalvariablen geschätzt werden, und es gibt keine eindeutigen Schätzwerte. Es gibt aber auch Vorteile dieses Konzeptes. Geeignete Transformationen des Common-Faktor-Raumes liefern meist akzeptable Lösungen. Eine Rotation sollte so geeignet gewählt werden, um zusätzliche externe Kriterien zu optimieren, z.B. die Interpretierbarkeit der Faktoren. Eine rotierte Schätzung eines Common Faktors repräsentiert meist die Originalvariablen, die große Koeffizienten in der Linearkombination haben. Die R-mode Faktorenanalyse ist eine „Art“ von Variablen-Clustering und eine Variable kann in mehr als einem Cluster vorkommen.

Beispiel-Programm (Faktor-Analyse):

```
%LET dset=datasetname ;
%LET varlist=qg01 qg02 qg03 qg04 qg05
        qg06 qg07 qg08 qg09 qg10
        qg11 qg12 qg13 qg14 qg15 qg16 ;


TITLE1 'Factor Analysis' ;

PROC FACTOR DATA=&dset          METHOD=PRIN
            SCREE NFACT=3       ROUND
            PRIORS=SMC  ROTATE=Varimax FLAG=0.3 ;
VAR &varlist ;
RUN ;
```

Die bedeutsamen Komponenten können jetzt nur noch an Hand der Kriterien Scree-Test („Break“ im Schaubild!), Varianzanteil und Interpretierbarkeit ausgewählt werden. Die folgende Ausgabe zeigt das Ergebnis einer initialen Faktorenextraktion mit der Prozedur FACTOR und 16 Haupternährungsgruppen der EPIC-Studie (Teilstichprobe).

SAS-OUTPUT:

Teil 1: Tabelle der Eigenwerte

```
Factor Analysis

The FACTOR Procedure
Initial Factor Method: Principal Factors

                                    Prior Communality Estimates: SMC

      QG01          QG02          QG03          QG04          QG05          QG06
QG07          QG08

0.18576949    0.30519912    0.16054745    0.23143257    0.05857486    0.13511459
0.20184401    0.07855510

      QG09          QG10          QG11          QG12          QG13          QG14
QG15          QG16
```

```
0.11233983    0.20018049     0.15168119     0.09393953     0.20612389     0.10605832
0.16941939     0.08300745


Eigenvalues of the Reduced Correlation Matrix: Total = 2.4797873 Average = 0.1549867

        Eigenvalue    Difference    Proportion    Cumulative

   1    1.20720989    0.06982018        0.4868        0.4868
   2    1.13738971    0.52646600        0.4587        0.9455
   3    0.61092371    0.18500788        0.2464        1.1918
   4    0.42591583    0.13534787        0.1718        1.3636
   5    0.29056796    0.13711252        0.1172        1.4808
   6    0.15345544    0.05828123        0.0619        1.5427
   7    0.09517421    0.07669456        0.0384        1.5810
   8    0.01847965    0.05580151        0.0075        1.5885
   9    -.03732186    0.09106440       -0.0151        1.5734
  10    -.12838626    0.00226284       -0.0518        1.5217
  11    -.13064910    0.04459317       -0.0527        1.4690
  12    -.17524227    0.00851514       -0.0707        1.3983
  13    -.18375741    0.02859764       -0.0741        1.3242
  14    -.21235504    0.04574963       -0.0856        1.2386
  15    -.25810467    0.07540785       -0.1041        1.1345
  16    -.33351252                     -0.1345        1.0000

2 factors will be retained by the MINEIGEN criterion.
```

Wie bei der Hauptkomponentenanalyse haben die extrahierten Faktoren zwei wichtige Eigenschaften. Jeder Faktor trägt maximal zu dem Varianzanteil bei, der von anderen Faktoren noch nicht bereits berücksichtigt wurde und jeder Faktor ist unkorreliert zu den vorher bereits extrahierten Faktoren.

Teil 3: Rotiertes Faktormuster

```
                        Rotated Factor Pattern

                                        Factor1      Factor2      Factor3

QG01    POTATOES AND OTHER TUBERS           -22           38 *         18
QG02    VEGETABLES                           59 *          5           -5
QG03    LEGUMES                              44 *          7           -5
QG04    FRUITS                               53 *         -7           -9
QG05    DAIRY PRODUCTS                       -3            3           25
QG06    CEREALS AND CEREAL PRODUCTS          26           18           20
QG07    MEAT AND MEAT PRODUCTS               -2           54 *         11
QG08    FISH AND SHELLFISH                    6           26          -15
QG09    EGGS AND EGG PRODUCTS                 9           33 *          4
QG10    FAT                                  22           34 *         14
QG11    SUGAR AND CONFECTIONERY              -4            5           48 *
QG12    CAKES                                 1           -5           36 *
QG13    NON ALCOHOLIC BEVERAGES             -28           13           30 *
QG14    ALCOHOLIC BEVERAGES                 -10           36 *         -2
QG15    CONDIMENTS AND SAUCES                23            8           38 *
QG16    SOUPS, BOUILLON                      26           -3           12

Printed values are multiplied by 100 and rounded to the nearest integer.
Values greater than 0.3 are flagged by an '*'.
```

Während bei der Hauptkomponentenanalyse das Eigenwertkriterium sinnvoll war, da jede Variable eine Varianzeinheit in die Analyse einbrachte, ist es hier nicht mehr geeignet. Jede Variable trägt nun nicht mehr mit einer Varianzeinheit zur Analyse bei, stattdessen tun dies die Schätzer der (Prior) Kommunalitäten. Diese Schätzer sind aber kleiner als 1, und es macht deshalb keinen Sinn mehr ein solches Kriterium zu wählen, wenn es um die Bestimmung der bedeutsamen

Komponenten geht. Somit verbleiben hier nur noch der Scree-Test (Break“ im Schaubild), der Varianzanteil und die Interpretierbarkeit mit dem Ziel einer orthogonalen Lösung. Auch hier sollte man fragen: Gibt es mind. 3 Variablen mit signifikanten Ladungen bei jedem bedeutsamen Faktor? Haben die Variablen eines Faktors eine gemeinsame Zuordnung? Messen unterschiedliche Variablen bei verschiedenen Faktoren unterschiedliche Konstrukte? Zeigt das rotierte Faktormuster eine einfache Struktur?

4 Allgemeine Strategie

Für eine allgemeine Strategie in einer explorativen Analyse wäre etwa zu unterscheiden zwischen Haupt- (Ernährungs-) Variablen und Ernährungs-untergruppen, die aber oft nicht genügend erfasst wurden sowie Zusatzinformationen. So können etwa Ernährungsmuster dazu verwendet werden, um zwischen Männern / Frauen, Zentren, Rauchern / Nichtrauchern, etc. zu unterscheiden. Weiter sollte zwischen subjekt- und variablenorientierten Verfahren unterschieden werden.

Eine alternative Vorgehensweise wäre auch subjekt- und variablenorientierte Verfahren zu kombinieren. So kann man z.B. mit einer Faktoranalyse beginnen und anschließend eine Clusteranalyse vornehmen. Oder mit einer Clusteranalyse anfangen, mit Diskriminanz-Faktoranalyse der Originalernährungsvariablen fortfahren um die Subjektcluster vorherzusagen. Eine Visualisierung der Ergebnisse sollte die Analyse abrunden. Alternativ ist auch eine modellbasierte Faktor- / Clusteranalyse sinnvoll. Weiter wäre zu untersuchen, wie Zusatzinformationen in die Muster eingehen („supervised“ Mustererkennung). Eine interessante Aufgabe wäre auch eine geeignete Wahl der Faktorenzahl.

Speziell für Daten aus der EPIC-Studie ergeben sich besondere Probleme, wie etwa der Umfang der Studie (ca. 500000 Probanden, viele Variablen), die Multilevel Struktur (Cluster der Individuen auf Zentrum-Ebene, Cluster der Zentren (29), Cluster der Länder (9)) und die Frage des Measurement Errors.

Literatur

1. Everitt, B.S. (1993): *Cluster Analysis*. Arnold, London
2. Gordon, A.D. (1999): Classification. Chapman and Hall / CRC, New York
3. Gower, J.C. (1974): Maximal Predictive Classification. *Biometrics* **30**, 643-654
4. Hatcher, L. (1994): A Step-by-Step Appproach to Using the SAS System for Factor Analysis and Structural Equation Modelling. SAS Institute, Cary, NC
5. Hubálek, Z. (1982): Coefficients of association and similarity based on binary (presence-absence data): an evaluation. *Biological Reviews,* **57**, 669-689

6. Jäger B., Wodny M., Rudolph P.E., Patschinsky, D. (2001): Clusteranalyse mit Binärdaten. In: M. Schuhmacher et al.: Proceedings 5. KSFE, Universität Hohenheim, 8.-9. März 2001
7. Kaufman, L. and Rousseeuw, P. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York
8. Rummel, R.J. (1970): *Applied Factor Analysis*. Evanston, IL: Northwestern University Press.
9. Spector, P.E. (1992): *Summated Ratingscale Construction: An Introduction*. Newbury Park, CA: Sage