In: N. Victor, W. Lehmacher & W. van Eimeren (eds). Explorative Datenanalyse. Lecture Notes in Medizinischer Informatik und Statistik, 1980, Band 26. Springer: Heidelberg.

KOVARIANZSELEKTION ALS EXPLORATIVE METHODE

N. WERMUTH
Psychologisches Institut
Universität Mainz

Zusammenfassung

Die Theorie der Kovarianzselektion - insbesondere die der Untergruppe des multiplikativen Modelle - wird kurz beschrieben. Es wird gezeigt, inwiefern jedes multiplikative Kovarianzselektionsmodell einem System von Regressionsgleichungen und einem Modell der Pfadanalyse entspricht. Anhand eines vorgegebenen Datensatzes wird schließlich verdeutlicht, wie man Kovarianzselektion zur Datenexploration verwenden kann.

1 Einführung

Kovarianzselektion (Dempster 1972) ist zunächst eine bloße mathematisch-statistische Theorie. Sie sagt aus, daß der Maximum-Likelihood-Schätzer einer Kovarianz- oder Korrelationsmatrix existiert und eindeutig definiert ist, dann, wenn den Kovarianzen bestimmte Restriktionen auferlegt werden. Kovarianzselektion nutzt jedoch auch bei einer explorativen Datenanalyse. Mit Hilfe eines Suchverfahrens kann man Kovarianzselektionsmodelle finden, die rechnerisch gut mit den Beobachtungen übereinstimmem. Jedes gut passende Modell kann sodann als Zusammenhangshypothese für die gemessenen Variablen zur inhaltlichen Diskussion gestellt werden und ist darüberhinaus an neuen Beobachtungen inferenzstatistisch überprüfbar.

Die Modelle der Kovarianzselektion sind unter der Voraussetzung einer multivariaten Normalverteilung abgeleitet worden. Es bleibt zu klären, wie empfindlich die Methode gegenüber Verletzungen dieser Annahme ist. Fest steht, daß Kovarianzselektionsmodelle unter völlig anderen Verteilungsannahmen formuliert werden können (e.g. Darroch, Lauritzen und Speed 198o, Wermuth 1979, 1976 a).

Aus verschiedenen Gründen ist Kovarianzselektion für Anwendungen attraktiv. Jedes Modell läßt sich mit Hilfe des wohlbekannten Begriffs der partiellen Korrelation vollständig beschreiben und außerdem graphisch anschaulich darstellen. Bei einer großen Modelluntergruppe ergeben sich weitere Vorteile. Für jedes sogenannte multiplikative Kovarianzselektionsmodell gilt, a) daß es äquivalent zu (mindestens) einem System von Regressionsgleichungen ist, b) daß man den Schätzwert der Kovarianzmatrix mittels der gewöhnlichen Methode der kleinsten Quadrate berechnen kann, und c) daß es bildlich durch einen gerichteten Graphen darstellbar ist.
Letztere Beschreibung korrelativer Zusammenhänge hat eine lange Tradition. Sie wurde von Wright in die Biologie, genauer in die Genetik, bereits im Jahr 1923 unter dem Namen Pfadanalyse eingeführt und für die Sozialwissenschaften von Duncan (1966) wiederentdeckt. Daß allerdings die von Wright vorgeschlagene Schätzmethode für - vom Modell implizierte - Korrelationskoeffizienten zu Fehlern führen kann, bleibt auch in neueren Beschreibungen der Pfadanalyse (Li 1975, Duncan 1975) unerwähnt. Folgt man unkritisch Wrights Vorschlag, so läuft man Gefahr, eine zu gute Übereinstimmung zwischen Modellannahmen und Beobachtungen auszuweisen. Dieser Fehler ist jedoch ausgeschlossen, wenn das Modell der Pfadanalyse zugleich ein (multiplikatives) Kovarianzselektionsmodell ist (Wermuth 198o). So gesehen, bietet Kovarianzselektion die mathematisch-statistische Rechtfertigung für eine bestimmte Gruppe von Pfadanalysen.
Im folgenden beschreiben wir zunächst die Theorie der Kovarianzselektion allgemein, danach die multiplikativen Modelle und schließlich anhand eines Datenbeispiels, wie Kovarianzselektion explorativ verwendet werden kann.

2 Kovarianzselektion

Gegeben seien p Zufallsvariable, die einer um Null zentrierten, nicht degenerierten Normalverteilung folgen: $(X_1, X_2, \ldots, X_p) \sim N(\underset{\sim}{0}, \underset{\sim}{\Sigma})$; $\underset{\sim}{\Sigma}$ positiv definit. Für die Variablen X_i und X_j bezeichne σ_{ij} die Kovarianz und σ^{ij} die Konzentration. Dann sind σ_{ij} und σ^{ij} das Element in Position (i,j) der Kovarianzmatrix $\underset{\sim}{\Sigma}$ und der inversen Kovarianzmatrix $\underset{\sim}{\Sigma}^{-1}$, die auch Konzentrationsmatrix genannt wird. Weiterhin sei $\tilde{I}$ die Indexmenge aller $\binom{p}{2}$ Paare: $\tilde{I} = \{(i,j) \mid 1 \leq i < j \leq p\}$ und I bezeichne eine beliebige Teilmenge von $\tilde{I}$. Ein Nullmuster in der Konzentrationsmatrix $\underset{\sim}{\Sigma}^{-1}$ bedeute, daß die Konzentrationen aller in I genannten Paare gleich Null sind ($\sigma^{ij} = 0$ für alle $(i,j) \in I$).

2a Theorie

Dempster zeigte 1972, daß der Maximum-Likelihood-Schätzer (ML-Schätzer) $\hat{\Sigma}$ für ein vorgegebenes Nullmuster in der Konzentrationsmatrix eindeutig bestimmt ist durch:

$$\hat{\sigma}_{ij} = s_{ij} \text{ für } i=j \text{ und } (i,j) \notin I$$
$$\hat{\sigma}^{ij} = 0 \quad \text{für } (i,j) \in I \qquad (2.1)$$

wobei s_{ij} die beobachtete Kovarianz der Variablen X_i und X_j darstellt, also das Element in der Position (i,j) der beobachtete Kovarianzmatrix $\underset{\sim}{S}$. Auf den Determinanten von $\hat{\underset{\sim}{\Sigma}}$ und $\underset{\sim}{S}$ basiert ein Test für die Güte der Anpassung der Modellannahmen an die Beobachtungen. Für einen großen Stichprobenumfang n und r Paare in I ist der folgende Likelihood-Quotient Chiquadrat- verteilt mit r Freiheitsgraden:

$$LQ\text{-}\chi^2 = -2n \ln \frac{|\hat{\underset{\sim}{\Sigma}}|}{|\underset{\sim}{S}|} \sim \chi^2_r \qquad (2.2)$$

Ein programmierter Algorithmus zur Berechnung des ML-Schätzers $\hat{\underset{\sim}{\Sigma}}$ wurde 1977 veröffentlicht (Wermuth und Scheidt).

2b Konzentration und partielle Korrelation

Es besteht eine enge Beziehung zwischen der Konzentration σ^{ij} eines Variablenpaares und dem partiellen Korrelationskoeffizienten gegeben alle übrigen p-2 Variablen, den wir mit $\rho_{ij.K}$ für $K = \{1, \ldots, p\} \setminus \{i,j\}$ bezeichnen. Insbesondere gilt unter den zuvor genannten Annahmen:

$$\sigma^{ij} = 0 \leftrightarrow \rho_{ij.K} = 0 \qquad (2.3)$$

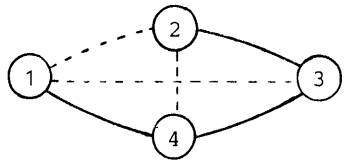
Das bedeutet, daß ein Kovarianzselektionsmodell, das durch ein Nullmuster in den Konzentrationen gekennzeichnet ist, sich gleichzeitig mittels einer Folge von partiellen Nullkorrelationen beschreiben und interpretieren läßt.

2c Beispiel

Für p = 4 sei das Nullmuster in der Konzentrationsmatrix mit $I = \{(1,2),(1,3),(2,4)\}$ festgelegt. Dann liegt ein Kovarianzselektionsmodell vor, in dem $\sigma^{12} = \sigma^{13} = \sigma^{24} = 0$ und zugleich $\rho_{12.34} = \rho_{13.24} = \rho_{24.13} = 0$ gilt. Der Schätzer der Kovarianzmatrix weicht nur in den Positionen $(i,j) \in I$ von der beobachteten Kovarianzmatrix ab; er ist nach (2.1) eindeutig bestimmt durch folgende Angaben:

$$\hat{\underset{\sim}{\Sigma}} = \begin{pmatrix} s_{11} & ? & ? & s_{14} \\ & s_{22} & s_{23} & ? \\ & & s_{33} & s_{34} \\ & & & s_{44} \end{pmatrix} \qquad \hat{\underset{\sim}{\Sigma}}^{-1} = \begin{pmatrix} ? & 0 & 0 & ? \\ & ? & ? & 0 \\ & & ? & ? \\ & & & ? \end{pmatrix}$$

Zeichnen wir nicht festgelegte Konzentrationen mit durchgezogenen, und Nullkonzentrationen mit gestrichelten Linien, so sieht der zum Modell mit $I=\{(1,2),(1,3),(2,4)\}$ gehörige Graph folgendermaßen aus:



Das Bild soll veranschaulichen, daß in diesem Modell die Kovarianzen der Paare (1,2) (1,3) und (2,4) durch die Beziehungen zwischen den übrigen Variablen hervorgebracht werden. Dies wird dadurch bewirkt, daß die entsprechenden Konzentrationen zu Null gesetzt sind.

3 Multiplikative Kovarianzselektionsmodelle

Während der ML-Schätzer $\hat{\underset{\sim}{\Sigma}}$ bei der Kovarianzselektion im allgemeinen mit iterativen Verfahren bestimmt werden muß, kann er bei der Untergruppe der multiplikativen Modelle mittels der Methode der kleinsten Quadrate in geschlossener Form (vergleiche 3.3) angegeben werden.

3a Theorie

Anhand des Nullmusters in den Konzentrationen - oder anhand der Indexmenge I - läßt sich leicht feststellen, ob ein multiplikatives Modell vorliegt.
Bei allen multiplikativen Modellen - und nur bei diesen - lassen sich die Variablen so anordnen, daß das Nullmuster in den Konzentrationen reduzibel wird (Wermuth 198o), daß heißt, das I die folgenden Bedingung erfüllt:

I ist reduzibel wenn für jedes $(i,j)\in I$ gilt, daß $(h,i)\in I$ oder $(h,j)\in I$ für alle $h = 1, \ldots, i-1$. (3.1)

So kennzeichnet etwa $I = \{(1,2),(2,4),(3,4)\}$ kein reduzibles Nullmuster, da weder $\sigma^{13} = 0$ noch $\sigma^{14} = 0$, aber $\sigma^{34} = 0$ ist; einfaches umnumerieren aber er-

gibt ein reduzibles Nullmuster, so zum Beispiel I ={(1,2),(1,3),(2,4)}.
Weitere Diskussionen multiplikativer Modelle findet man im Zusammenhang mit Kontigenztafeln bei Darroch, Lauritzen und Speed (1980), Sundberg (1975) und Goodman (1970).
Zur Berechnung des ML-Schätzers $\hat{\Sigma}$ bestimmt man für ein vorgegebenes reduzibles Nullmuster und ein festes i zunächst die Koeffizienten $\hat{b}_{ij}$ als Lösung der folgenden Normalgleichungen:

$$s_{ij} = \Sigma_l \hat{b}_{il} s_{lj} \text{ für } (i,j) \notin I \text{ und } (i,l) \notin I \tag{3.2}$$

und man erhält $\hat{\Sigma}$ nun aufbauend in der Reihenfolge i = p-1, p-2,...,1 aus:

$$\hat{\sigma}_{ij} = \begin{cases} s_{ij} & \text{für } i = j \text{ und für } (i,j) \notin I \\ \Sigma_l \hat{b}_{il} \hat{\sigma}_{lj} & \text{für } (i,j) \in I \text{ und } (i,l) \notin I \end{cases} \tag{3.3}$$

Dieses Ergebnis (Wermuth 1980) basiert darauf, daß es eine Zerlegung der Kovarianzmatrix in Dreiecksmatrizen gibt, deren Elemente als Regressionskoeffizienten interpretiert werden können.

3b Konzentrationen und Regressionskoeffizienten

Ein Satz aus der Matrizenrechnung (Anderson 1958) besagt, daß für jede positiv definite Matrix Σ eine obere Dreiecksmatrix B und eine Diagonalmatrix D existiert, so daß $B\Sigma B^T = D$ und $\Sigma^{-1} = B^T D^{-1} B$ gilt. Ist Σ eine Kovarianzmatrix ohne Restriktion, so lassen sich die Elemente der Matrizen B und D als folgende Regressionskoeffizienten und Residualvarianzen interpretieren: $b_{ij} = -a_{ij.r}$ mit $r = \{i+1, \ldots, p\} \setminus \{j\}$ und $d_{ii} = \sigma_{ii.i+1, \ldots, p}$.
Bei p = 4 Variablen (und I = ∅) zum Beispiel entsprechen die Elemente in B den Regressionskoeffizienten in einem System, das in der Ökonometrie ein vollständiges rekursives Gleichungssystem mit unabhängigen Fehlern U_i genannt wird. (Goldberger 1964):

$$\begin{aligned} X_1 &= a_{12.34} X_2 + a_{13.24} X_3 + a_{14.23} X_4 + U_1 \\ X_2 &= a_{23.4} X_3 + a_{24.3} X_4 + U_2 \\ X_3 &= a_{34} X_4 + U_3. \end{aligned} \tag{3.4}$$

Bei einer Kovarianzmatrix mit einem Nullmuster in der Konzentrationsmatrix $\hat{\Sigma}^{-1}$ ergibt sich dasselbe Nullmuster in der Dreiecksmatrix B

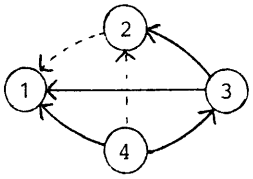
wie in $\underset{\sim}{\Sigma}^{-1}$ dann und nur dann, wenn das Nullmuster reduzibel ist (Wermuth 1980). Die Elemente von $\underset{\sim}{B}$ entsprechen sodann Regressionskoeffizienten in einem unvollständigen, rekursiven Gleichungssystem, in dem jede Variable X_i auf die Variablen X_j für $(i,j) \in I$ regressiert wird.
Für $p = 4$ und $I = \{(1,2),(2,4)\}$ etwa ist $b_{12} = b_{24} = 0$, und - ausgehend von (3.4) - auch $a_{12.34} = a_{24.3} = 0$, sodaß sich das folgende unvollständige System ergibt:

$$\begin{aligned} X_1 &= a_{13.4}\, X_3 + a_{14.3}\, X_4 + U_1 \\ X_2 &= a_{23}\, X_3 + U_2 \\ X_3 &= a_{34}\, X_4 + U_3. \end{aligned} \qquad (3.5)$$

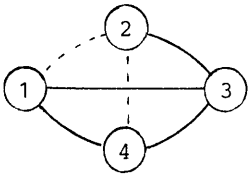
3c Graphische Darstellung

Für alle Modelle mit reduziblem Nullmuster gibt es außer einem ungerichteten Graphen (wie in 2c beschrieben) einen gerichteten Graphen, also eine Pfadanalysendarstellung. Für das Modell mit $I = \{(1,2),(2,4)\}$ sind sind diese beiden Graphen Bild B und A.

A)



B)



Beide Graphen verdeutlichen unterschiedliche Interpretationen des zugrundeliegenden Kovarianzselektionsmodells. Bild A veranschaulicht die Interpretation von System (3.5). Es zeigt, daß X_1 von X_2 nur mittelbar abhängt, dagegen von X_3 und X_4 direkt beeinflußt wird. Ähnlich ist X_2 nur mittelbar von X_4 abhängig, dagegen beeinflußt X_3 die Variable X_2 und X_4 beeinflußt die Variable X_3 direkt. Bild B weist nur die Nullkonzentrationen oder die Nullkorrelationen $\rho_{12.34} = \rho_{24.13} = 0$ aus, die besagen, daß X_1 und X_2 bedingt unabhängig sind gegeben X_3 und X_4, und daß X_2 und X_4 bedingt unabhängig sind gegeben X_1 und X_3.

Kovarianzselektion ermöglicht im allgemeinen (wie in Bild B) nur Aussagen über Assoziationen und (bedingte) Unabhängigkeiten zwischen p Variablen. Ein Modell mit reduziblem Nullmuster dagegen erlaubt gleichzeitig (wie in Bild A) eine Interpretation als ein System von Ziel- und Einflußgrößen. Im folgenden Datenbeispiel ist ein rekursives System aus inhaltlichen Überlegungen vorgegeben.

4 Datenbeispiel

Miller schlug 1976 ein Abhängigkeitsmodell für eheliche Beziehungen in der Form der Pfadanalyse vor. Für 16o Ehepaare gab er Korrelationskoeffizienten für sechs Variable an:

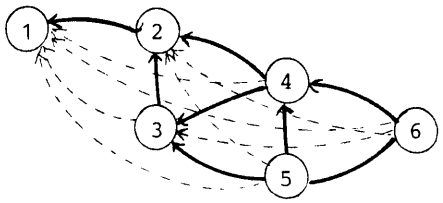
Korrelationen x 1ooo nach Miller (1976)

Variablen	1	2	3	4	5
1 Zufriedenheit in der Ehe	1ooo				
2 Zeit füreinander	37o				
3 Altersunterschied der Kinder	- 62	- 16			
4 Anzahl der Kinder	47	-211	- 41		
5 Sozioökonomischer Status	132	289	48	- 217	
6 Ehedauer	127	-1oo	216	552	24o

Als Modell wird dazu von Miller postuliert:

$$
\begin{aligned}
X_1 &= a_{12}\, X_2 &&&&&&&&&& + U_1 \\
X_2 &= a_{23.4}\, X_3 &&+ a_{24.3}\, X_4 &&&&&&&& + U_2 \\
X_3 &= a_{34.5}\, X_4 &&+ a_{35.4}\, X_5 &&&&&&&& + U_3 \\
X_4 &= a_{45.6}\, X_5 &&+ a_{46.5}\, X_6 &&&&&&&& + U_4
\end{aligned}
\qquad (4.1)
$$

Der dazu gehörige Graph ist:



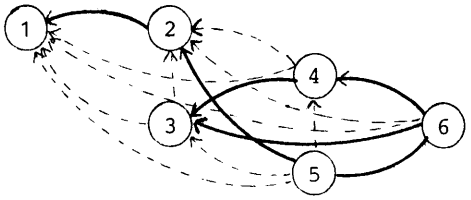
Da das Nullmuster in den Regressionskoeffizienten des rekursiven Systems (4.1) reduzibel ist, entspricht es einem Kovarianzselektionsmodell mit demselben Nullmuster $I=\{(1,3),(1,4),(1,5),(1,6),(2,5),(2,6),(3,6)\}$ und man kann einen Likelihoodquotiententest - nach (2.2) und (3.3) - durchführen. Das Testergebnis besagt, daß das Modell nicht mit den Beobachtungen zu vereinbaren ist. (LQ-χ^2= 3o,17 bei 7 Freiheitsgraden).

An dieser Stelle ist es sinnvoll zu explorieren, welche anderen Hypothesen mit den Daten besser übereinstimmen. Eine solche Möglichkeit bietet ein Suchverfahren nach multiplikativen Modellen (Wermuth 1976b, 1978, 198o), für das das Fortran-Programm COVSEL vorliegt, welches maximal dreißig Variable berücksichtigen kann. Dieses Programm weist für Millers Daten das Modell mit dem Nullmuster $I=\{(1,3),(1,4),(1,5),(1,6),(2,3),(2,4),$

(2,6),(3,5),(4,5)} als noch annehmbar aus. (LQ-χ^2 = 14,2 bei 9 Freiheitsgraden, p=0,12). Die Auswahl ist dabei so vorgenommen worden, das jedes Modell, das durch eine Teilmenge von I gekennzeichnet ist, ebenfalls mit den Beobachtungen zu vereinbaren ist. Das Modell kann als Gleichungssystem

$$\begin{aligned} X_1 &= a_{12} X_2 &&&&&&+ U_1 \\ X_2 &= &&+ a_{25} X_5 &&&&+ U_2 \\ X_3 &= a_{34.6} X_4 &&&+ a_{36.4} X_6 &&&+ U_3 \\ X_4 &= &&a_{46} X_6 &&&&+ U_4 \end{aligned} \qquad (4.2)$$

und als gerichteter Graph dargestellt werden:



Demnach hat Miller vor allem den Einfluß des Sozialstatus (5) auf die Zeit füreinander (2) und den der Ehedauer (6) auf den Altersunterschied der Kinder (3) unterschätzt und möglicherweise einige andere Beziehungen ((2,3),(2,4),(3,5),(4,5)) überschätzt. Ohne auf die inhaltliche Interpretation von Millers Variablen hier weiter einzugehen, läßt sich festhalten, daß das Suchverfahren einerseits auf Abweichungen der ursprünglichen Hypothese von den Beobachtungen hinweist, andererseits eine verbesserte Hypothesenformulierung ermöglicht. Sei es, daß man an neuen Daten das Kovarianzmodell, das dem System (4.2) entspricht, prüfen möchte oder sei es, daß man dafür eine neue Hypothese formuliert, die eine Verbindung darstellt aus der ursprünglichen Hypothese (4.1) und der zu den vorliegenden Daten passenden Hypothese (4.2).

Literatur

ANDERSON, T.W., 1958:
An Introduction to Multivariate Statistical Analysis. Wiley, New York

DARROCH, J. N., Lauritzen, S. L. und Speed, T. P., 1980:
Markov Fields and Log-Linear Interaction Models for Contingency Tables. Annals of Statistics, 8, 522-539.

DEMPSTER, A. P. , 1972:
Covariance Selection. Biometrics 28, 157 - 175.

DUNCAN, O. D., 1966:
Path Analysis :Sociological Examples. American Journal of Sociology 72, 1 - 16.

ders., 1975:
Introduction to Structural Equation Models. Academic Press, New York

GOLDBERGER, A. S., 1964:
Econometric Theory, Wiley, New York

GOODMAN, L. A. 197o:
The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications. Journal of the American Statistical Association, 65, 226 - 256.

MILLER, B. C., 1976:
A Multivariate Development Model of Marital Satisfaction. Journal of Marriage an the Familiy, 38, 634 - 657.

LI, C. C., 1975: Path Analysis: A Primer. The Boxwood Press, Pacific Grove

SUNDBERG, R. 1975:
Some Results about Decomposable (or Markov-type) Models for Multidimensional Contingency Tables: Distribution of Marginals and Partitioning of Tests, Scandinavian Journal of Statistics, 2, 71 - 79.

WERMUTH, N., 1976a:
Analogies Between Multiplicative Models in Contingency Tables and Covariance Selection, Biometrics, 32, 95 - 1o8.

dies., 1976 b:
Model Search Among Multiplicative Models. Biometrics 32, 253-263.

dies., 1978:
Zusammenhangsanalysen Medizinischer Daten, Band 5, Lecture Notes in Medizinischer Informatik und Statistik, Springer, Berlin.

dies., 1979:
Datenanalyse und Multiplikative Modelle. Allgemeines Statistisches Archiv, 323 - 339.

dies., 198o:
Linear Recursive Equations, Covariance Selection, and Path Analysis. Journal of the American Statistical Association (erscheint im Dezember Heft).

WERMUTH, N. und Scheidt, E. , 1977:
Fitting a Covariance Selection Model to a Matrix, Algorithm AS 1o5, Journal of the Royal Statistical Society, Series C, Applied Statistics, 26, 88 - 92.

WRIGHT, S., 1923 The Theory of Path Coefficients: A Reply to Niles Criticism, Genetics, 8, 239 - 255.

Prof. Dr. N. Wermuth
Psychologisches Institut
Abt. Statistik der
Universität Mainz
Saarstr. 1
D - 6500 Mainz