FAKTOROVÁ ANALÝZA KONTINGENČNÍCH TABULEK

Jan ŘEHÁK (Ústav pro filosofii a sociologii ČSAV)
Ivana LOUČKOVÁ (Katedra aplikované sociologie FFUP, Olomouc)

Základní statistickou úlohou v kontingenční tabulce T o rozměru $R \times S$ je testování hypotézy nezávislosti resp. hypotézy homogenity H_N vs H_Z resp. H_{hom} vs H_{het}. Tento test však je pro praxi podrobné analýzy dat málo užitečný (kromě tabulek 2x2), neboť a) v případě přijetí H_N (H_{hom}) není vyloučena platnost některé speciální alternativy, na kterou je omnibusová statistika málo citlivá, a na kterou lze přijít buď formulací specifických alternativ či speciálními rozklady obecné statistiky pro nezávislost, b) v případě zamítnutí H_N (H_{hom}) nás zajímá podrobnější informace. Tu lze získat např. měřením stupně závislosti, seskupováním řádků (resp.sloupců), hledáním trendů a regresní analýzou, rozklady tabulky a slučováním kategorií (COLLAPS), specifikací závislosti pomocí testování reziduí v polích(znaménková schémata), faktorovými rozklady (LINDA).

I. Specifikace závislosti pomocí znaménkového schématu

Testování reziduí od modelu nezávislosti lze provádět podle znaménkového schématu, což je grafická pomůcka pro rychlou orientaci ve struktuře závislostí (Řehák, Řeháková [1978]).

1. Určíme tři hladiny významnosti $\alpha_1, \alpha_2, \alpha_3$ (např. 0,05 ; 0,01 a 0,001).

2. Určíme standardizovaná rezidua

$$z_{ij} = \sqrt{n}\,\frac{f_{ij} - f_{i+}f_{+j}}{\sqrt{f_{i+}f_{+j}(1-f_{i+})(1-f_{+j})}} \sim N(0,1) \qquad (1)$$

kde $f_{ij} = n_{ij}/n$, $f_{i+} = \sum_j f_{ij}$, $f_{+j} = \sum_i f_{ij}$, n_{ij} jsou prvky tabulky $T = \|n_{ij}\|$.

3. Provedeme testování významnosti z_{ij} postupně na hladinách $\alpha_1, \alpha_2, \alpha_3$ a přiřadíme znaménka (podle kladné nebo záporné hodnoty z_{ij}) : 0 ≙ nevýznamné ; +, - ≙ významné na α_1; ++, -- ≙ významné na α_2 ; +++, --- ≙ významné na α_3. Testování se provádí buď jednotlivě pro každé pole, nebo simultánně pomocí Bonferroniho nerovnosti, nebo simultánně pomocí Holmova sekvenčního postupu. Postup lze schematicky znázornit

$$T = \|n_{ij}\| \longrightarrow Z = \|z_{ij}\| \longrightarrow S = \|s_{ij}\|$$

kde $s_{ij} = ---,--,-,\ 0,+,++,+++$.

Poznámka : Místo statistik z_{ij} lze použít kterékoliv jiné statistiky charakterizující stupeň závislosti v poli, např. některou ze statistik typu logaritmických interakcí. Postup lze obecně aplikovat na rezidua od kteréhokoliv modelu pro kategorizovaná data, jsou-li známa rozložení reziduí.

Znaménkové schéma má svá interpretační omezení (jako ostatně každá statistická metoda), která zde plynou ze závislosti reziduí. V mnoha aplikačních situacích však pomáhá k jednoduché orientaci v chování závislosti a k určení významu a specifikaci H_z (resp.H_{het}). Ve složitějších případech však použijeme jiné metody, např. faktorový rozklad reziduí.

II. Lineární dekompoziční algoritmus pro rektangulární matice dat

Dříve než popíšeme faktorovou analýzu kontingenční tabulky T, uvedeme základní věty, na nichž je celý postup založen.

Věta: a) Je-li Y matice typu $R \times S$, pak existuje matice A typu $R \times M$, B typu $S \times M$, D typu $M \times M$, diagonální, $d_1 \geq d_2 \geq \dots \geq d_M$, $M = h(Y)$ tak, že platí

(2) $$Y = A D B^{\prime},$$

přičemž A je tvořena M charakteristickými vektory matice $YY^{\prime}$, B je tvořena M charakteristickými vektory matice $Y^{\prime}Y$, $d_i = \sqrt{\lambda_i}$, kde λ_i je i-té charakteristické číslo matice $YY^{\prime}$ nebo $Y^{\prime}Y$, $(\lambda_i > 0)$.

(3) b) $$Y = P Q^{\prime}, \quad \text{kde} \quad P = A D^{1/2}, \quad Q = B D^{1/2}$$

c) Pro každou ortogonální matici W platí

(4) $$Y = P^{*}Q^{*\prime}, \quad \text{kde} \quad P^{*} = P W, \quad Q^{*} = Q W.$$

d) Je-li Y řádkově (sloupcově) centrovaná, tj. $\sum_j y_{ij} = 0$ $(\sum_i y_{ij} = 0)$, pak $B(A)$ je sloupcově centrovaná. Je-li Y řádkově i sloupcově centrovaná (dvojitě centrovaná), pak A i B jsou sloupcově centrované (totéž platí o P,Q resp. P^{*}, Q^{*}).

Řešení rozkladu (2) není tedy jednoznačné, ale tvoří třídu vzájemně na sebe převeditelných matic pomocí ortogonálních transformací. Rozklad (2) však je optimální ve smyslu Eckart-Youngovy věty: posloupnost řešení (A_r, B_r, D_r), $r = 1,2,\dots,M$ skýtá vždy optimální přiblížení k matici Y ve smyslu nejmenších čtverců pro zvyšující se hodnosti.

Věta (Eckart, Young [1936]) : Buď Y matice typu $R \times S$, $h(Y) = M$ a $K \leq M$. Pak $\tilde{Y}_K$ typu $R \times S$, hodnosti K je nejlepší aproximací matice Y ve smyslu nejmenších čtverců, $\sum\sum (y_{ij} - y^{*}_{ij})^2 \longrightarrow \min$, je-li

(5) $$\tilde{Y}_K = A_K D_K B_K^{\prime} = P_K Q_K^{\prime}.$$

A_K, B_K vzniknou z A,B zachováním prvních K sloupců a D_K z D zachováním prvních K sloupců i řádků, $P_K = A_K D_K^{1/2}$, $Q_K = B_K D_K^{1/2}$.

Dále konstatujeme, že ortogonální transformace P_K, Q_K nemá vliv na minimální součet čtverců : je-li $P^{*}_K = P_K W_K$, $Q^{*}_K = Q_K W_K$, platí $\tilde{V}_K = P^{*}_K Q^{*\prime}_K = P_K Q_K^{\prime} = \tilde{Y}_K$, ($W_K$ je ortogonální matice typu $K \times K$).

Matici $F_K = (P_K^{\prime} / Q_K^{\prime})^{\prime}$ typu $(R+S) \times K$ nazveme hlavním faktorovým řešením dimenze K pro matici Y, k-tý sloupec F_K nazveme k-tým faktorem. Obdobně pro $F^{*}_K = (P^{*\prime}_K | Q^{*}_K)^{\prime}$, která se nazývá rotovaným řešením.

Postup lineárního dekompozičního algoritmu (LINDA) :

1. Matice $X \longrightarrow Y$ dvojitě centrované.
2. Výpočet $Y Y^{\prime}$ resp. $Y^{\prime}Y$ pro $R \leq S$ resp. $S < R$.
3. Výpočet charakteristických čísel a vektorů matice $Y Y^{\prime}$ resp. $Y^{\prime}Y$: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$, vytvoření matic A,D.
4. $B = Y^{\prime} A D^{-1}$ (resp. $A = Y B D^{-1}$).
5. $P = A D^{1/2}$, $Q = B D^{1/2}$.
6. VARIMAX rotace pro P_K, Q_K podle jedné ze tří strategií (po volbě K)

a) $P^{*}_K = P_K V_K$, kde V_K je ortogonální tak, aby P_K měla prostou strukturu (viz Harman [1973]) a $Q^{*}_K = Q_K V_K$ přizpůsobíme.

b) $Q^{*}_K = Q_K U_K$ k prosté struktuře, $P^{*}_K = P_K U_K$ přizpůsobíme (U_K je ortogonální matice).

c) $\begin{Vmatrix} P^{*}_K \\ Q^{*}_K \end{Vmatrix} = \begin{Vmatrix} P_K \\ Q_K \end{Vmatrix} . W_K$ k prosté struktuře (W_K je ortogonální).

III. Faktorová analýza kontingenčních tabulek

Výsledky části II lze jednoduše aplikovat na matici $T = \|n_{ij}\|$. V praxi používáme model: $n_{ij} \vee Pois\,(np_{ij})$, po transformaci $x_{ij} = \ln n_{ij}$ (resp. $x_{ij} = \ln(n_{ij} + \frac{1}{2})$, jsou-li v tabulce nuly) můžeme aplikovat postup

$$T \xrightarrow{\ln} X \xrightarrow{\text{dvojité centrování}} Y \xrightarrow{\text{rozklad}} F \xrightarrow{\text{rotace}} F_K^*$$

$X \longrightarrow Y$ je krok z ANOVA, očištění o marginální efekty, Y je matice reziduí

(6) $$y_{ij} = x_{ij} - \bar{x}_{i+} - \bar{x}_{+j} + \bar{x}_{++} = \text{reziduum}\,(i,j) = \ln \frac{n_{ij}\, n_{..}}{n_{i.}\, n_{.j}}$$

kde $n_{i.} = \sqrt[S]{\prod_j n_{ij}}$, $n_{.j} = \sqrt[R]{\prod_i n_{ij}}$, $n_{..} = \sqrt[RS]{\prod_i \prod_j n_{ij}}$.

Model faktorové dekompozice vede na

(7) $$x_{ij} = a_i + b_j + \sum_{k=1}^{K} d_k\, a_{ik}\, b_{jk} + \varepsilon = a_i + b_j + \sum_{k=1}^{K} p_{ik}\, q_{jk} + \varepsilon$$

kde P_K , Q_K jsou interakční faktory, a_i jsou efekty i-tého řádku, b_j jsou efekty j-tého sloupce.

y_{ij} = reziduum (i,j) = příspěvek 1.faktoru + příspěvek 2.faktoru + ... + $\varepsilon = p_{i1}q_{i1} + \ldots$

Příspěvek každého faktoru je úměrný efektům řádků a sloupců (je v každém řádku a sloupci konstantní).

Pro model platí ještě řada dalších vlastností, jako

a) $\sum_{i=1}^{M} d_i = \sum y_{ij}^2$.

b) $d_1 : d_2 : \ldots : d_M$ určují podíly vysvětlení reziduální variance pomocí jednotlivých faktorů.

c) $\sum_{i=1}^{K} d_i = \sum_{i=1}^{K} d_i^*$ pro každou rotaci a každé K.

d) $d_m / \sum d_m$ = koeficient determinace m-tého faktoru (m=1,2,...,M), obdobně $d_m^* / \sum d_m^*$.

IV. Faktorové dekompozice různých typů

1. Faktorová analýza matice dat Y (typu $N \times S$) - klasická - vychází ze sloupcové standardizace matice Y na Z a rozkladu matice Z podle (2) , přičemž však zjišťujeme pouze B resp. Q a posléze B_K^* resp. Q_K (viz Harman [1973]).

2. Kanonická analýza kontingenčních tabulek je hledání skórů x_r pro řádky a y_s pro sloupce kontingenční tabulky T typu $R \times S$ tak, aby $\text{corr}(X,Y)$ byla maximální. O tom viz Lancaster [1957] , [1963] , Guttman [1941], [1959], Fisher [1940], Kendall, Stuart [1973], Hirschfeld [1936] . V této souvislosti lze poznamenat, že z_{ij} v (1) jsou $\sqrt{n}$ násobky korelačních koeficientů.

(8) $$z_{ij} = \sqrt{n}\; R_{ij} = \sqrt{n}\; \text{corr}(A_i, B_j)$$

3. Korespondenční analýza vychází z hledání vztahů mezi řádky a sloupci Y , která nemusí odpovídat kontingenční tabulce. Viz Benzécri [1969], Hill [1974] .

4. Model FANOVA-faktorová analýza v ANOVA odpovídá rozkladu reziduí v ANOVA se dvěma vstupy (viz Golleb [1968a] , [1968b] , [1968c]).

5. Komparační kanonická analýza kontingenční tabulky byla zavedena Williamsem [1978] .

6. Analýza profilů byla provedena autory podle rozkladu (2),(3),(4) a (5).

7. Multidimenzionální škálování dvou typů objektů bylo pomocí metody LINBA provedeno autory Loučková, Šeblák [1982]). Vychází z toho, že matice F_K resp. F_K^* určuje projekci R

dimenzionálního eukleidovského prostoru, v němž se původně nacházejí body (kategorie) řádkové i sloupcové proměnné do K-rozměrného eukleidovského prostoru. Každá kategorie je reprezentována K-rozměrným vektorem, který je obsažen v příslušném řádku matice F_K resp. F_K^*.

V. Závěr

Faktorová dekompoziční analýza je explorační technika, která může sloužit jednak k odhalení struktury dat ať už výběrových nebo struktury parametrů (censových dat). Může též sloužit ke generování hypotéz a k přehledům i grafickému shrnutí informace o interakčních souvislostech v obdélnikových uspořádáních dat.

Literatura

1. Benzécri J.P. [1969] : Statistical Analysis as a Tool to Make Patterns Emerge from Data. In S.Watanabe (ed.): Methodologies of Pattern Recognition, New York, Academic Press.
2. Eckart C., Young G. [1936] : The Approximation of One Matrix by Another of Lower Rank. Psychometrika 1, 211-218.
3. Fisher R.A. [1940] : The Precision of Discriminant Functions. Annals of Eugenics 10,422-429.
4. Gollob H.F. [1968a] : A Statistical Model which Combines Features of Factor Analytic and Analysis of Variance Techniques. Psychometrika 33, 73-115.
5. Gollob H.F. [1968b] : Rejoinder to Tucker's "Comments on Confounding of Sources of Variation in Factor-Analytic Techniques". Psychological Butletin 70, 355-360.
6. Gollob H.F. [1968c] : Confounding of Sources of Variation in Factor-Analytic Techniques. Psychological Bulletin 70, 330-344.
7. Guttman L. [1941] : The Quantification of a Class of Attributes : A Theory and Method of Scale Construction. In Horst P. et al (eds): The Prediction of Personal Adjustment,321-348, New York, SSRC.
8. Guttman L. [1959] : Metricizing Rank-ordered and Unordered Data for Linear Factor Analysis. Sankhyà 21, 257-268.
9. Harman H.H. [1973] : Sovremennyj faktornyj analiz. Statistika, Moskva.
10. Hill M.O. [1974] : Correspondence Analysis : A Neglected Multivariate Method. Applied Statistics 23, 340-354.
11. Hirschfeld H.O. [1935] : A Connection between Correlation and Contingency. Cambridge Philosophical Society Proceedings 31, 520-524.
12. Holm S. [1979] : A Simple Sequentially Rejective Multiple Procedures. Scandinavian Journal of Statistics 6, 65-70.
13. Kendall M.G., Stuart A. [1973] : Statističeskije vyvody i svjazi. Nauka,Moskva.
14. Lancaster H.O. [1957] : Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Table. Biometrika 44, 289-292.
15. Lancaster H.O. [1963] : Canonical Correlations and Partitions of χ^2. Quarterly Journal of Mathematics 14, 220.

16. Lancaster H.O. [1969] : The Chi-squared Distribution. New York, Wiley.

17. Loučková I., Řehák J. [1979] : LINDA-soubor programů pro analýzu reziduálních odchylek v dvojrozměrných uspořádáních dat. In Dvořák P., Řehák J.(red): Současný vývoj programů pro potřeby sociologie, Štiřín 1979. Sborník ČSSR, Sekce metod a technik, 24-44.

18. Loučková I., Řehák J. [1981] : Škálování dvou množin objektů. In Řehák J.(red) : Měření a škálování, Mikulov 1981. Sborník ČSSS, Sekce metod a technik, v tisku.

19. McDonald R.P. Torii J., Nishisato S. [1979] : Some Results on Proper Eigenvalues and Eigenvectors with Applications to Scaling. Psychometrika 44, 211-227.

20. Řehák J., Řeháková B. [1978] : Analýza kontingenčních tabulek: rozlišení dvou základních typů a znaménkové schéma. Sociologický časopis XIV, 619-631.

21. Torgerson W.S. [1958] : Theory and Methods of Scaling. New York, Wiley.

22. Tukey J.W. [1949] : One Degree of Freedom for Nonadditivity. Biometrics 5, 232-242.

23. Williams J.S. [1978] : Canonical Analyds : A Factor Analytic Method Comparing Finite Discrete Distribution Functions. JASA 73, 781-786.