KOMPARACE PODMÍNĚNÝCH DISTRIBUCÍ KONTINGENČNÍ TABULKY KLASICKÝM MNOHOROZMĚRNÝM ŠKÁLOVÁNÍM

Jan Řehák
ÚFS ČSAV Praha
Ivana Loučková
KAS FFUP Olomouc

Podmíněne řádkové distribuce mají v dvourozměrných /i vícerozměrných/ kontingenčních tabulkách samostatný význam. Při odmítnutí hypotézy jejich homogenity pomocí některého ze známých testů je další analytickou úlohou určit vztahy mezi distribucemi a to nejen jejich seskupení do homogenních podskupin, ale také podle stupně jejich vzájemné heterogenity. Jedním z možných postupů je grafické znázornění distribucí jako bodů v eukleidovském prostoru /co nejmenší dimenze/, přičemž bližší body ukazují na podobnější a vzdálenější body na rozdílnější distribuce.

Úlohu můžeme formulovat takto:

Je dáno R distribucí $f_{(r)} = (f_{k/r})_{k=1}^{K}$, $r=1, \ldots, R$ a míra vzdálenosti $D(f_{(r)}, f_{(r')}) = D_{rr'}$; máme nalezt R bodů $X_r = (X_{r1}, X_{r2}, \ldots, X_{rM}) \in E_M$ a dimenzi M tak, aby

$$(1) \qquad D(f_{(r)}, f_{(r')}) = \left[\sum_m (X_{rm} - X_{r'm})^2\right]^{\frac{1}{2}} = d(X_r, X_{r'})$$

respektive pro dané L nalézt takové body $X_r = (X_{r1}, \ldots, X_{rL})$, aby

$$(2) \qquad \frac{\sum_r \sum_{r'} \left[\sum_m^L (X_{rm} - X_{r'm})^2\right]^{\frac{1}{2}}}{\sum_r \sum_{r'} D^2(f_{(r)}, f_{(r')})} = \min .$$

Tuto úlohu lze řešit pomocí metody klasického mnohorozměrného škálování, mají-li $D^2(f,g)$ tvar kvadratické formy.

Zobrazení $f_{(r)} \longrightarrow X_r$, $X_r \in E_M$ hledáme pomocí algoritmu kanonického rozkladu symetrických čtvercových matic /viz metoda DISTAN; J. Řehák, I. Loučková [1983] /. Pro obecnější míry nepodobnosti resp. podobnosti lze využít algoritmů MINISSA mnohorozměrného škálování /viz J.C. Lingoes [1973] /.

1. Vzdálenosti distribucí

Vzdálenosti mezi řádkovými distribucemi můžeme určit podle definice typu proměnné pomocí matice skórů $D = \|d_{ij}\|$ /viz J. Řehák, B. Řeháková [1979], [1982].

Pro n o m i n á l n í proměnné / p r o s t o u klasifikací/

$$(3) \qquad D_N^2(f_{(r)}, f_{(r')}) = \sum_{k=1}^{K} (f_{k/r} - f_{k/r'})^2 .$$

Pro o r d i n á l n í proměnné / u s p o ř á d a n o u klas./

(4) $$D_0^2\left(f_{(r)}, f_{(r')}\right) = \sum_{k=1}^{K}\left(F_{k/r} - F_{k/r'}\right)^2,$$

$F_{k/r}$ = distribuční funkce.

Pro z o b e c n ě n ý typ proměnné

(5) $$D^2\left(f_{(r)}, f_{(r')}\right) = \left(f_{(r)} - f_{(r')}\right)' D \left(f_{(r')} - f_{(r)}\right).$$

Pro k a r d i n á l n í typ ztrácí úloha smysl, neboť distribuce jsou uspořádány pomocí průměrů /lze ovšem zavést analýzu pomocí distribučních funkcí - vzorec (4) /.

Tyto vzdálenosti umožňují zavést pro komparační závěry celou řadu výhodných měr:

a) vzdálenost $f_{(r)}$ a směsi ostatních distribucí $\bar{f}_{(r)} = \frac{1}{(1 - w_r)} \sum_{i \neq r} w_i f_{(i)}$, $w = (w_1, \ldots, w_R)'$ je marginální sloupcové rozložení v tabulce nebo jinak určené váhy: $D\left(f_{(r)}, \bar{f}_{(r)}\right)$ resp. $D\left(f_{(r)}, \bar{f}_{(r)}\right) / \max D(f,g)$;

b) vzdálenost $f_{(r)}$ a marginální distribuce $f = \sum_i w_i f_{(i)}$, $D\left(f_{(r)}, f\right)$ resp. $D\left(f_{(r)}, f\right) / \max D(f,g)$;

c) míru neshody distribucí $\dfrac{\sum_r \sum_{r'} D^2\left(f_{(r)}, f_{(r')}\right)}{\max \sum\sum D^2\left(f_{(r)}, f_{(r')}\right)}$ a vzdálenost dvou skupin distribucí a pod.

2. Zobrazení $f_{(r)} \longrightarrow x_{(r)}$

Předpokládejme, že pro nějaké M existují x_r tak, že platí vztah (1):

$$D\left(f_{(r)}, f_{(r')}\right)^2 = d^2\left(x_r, x_{r'}\right), \qquad d \text{ je metrika v } E_M .$$

Dvojitým centrováním této matice dostaneme matici C o rozměru RxR a o prvcích

(6) $$c_{rr'} = -\frac{1}{2}\left(D^2\left(f_{(r)}, f_{(r')}\right) - \frac{1}{R}\sum_j D^2\left(f_{(j)}, f_{(r')}\right) - \frac{1}{R}\sum_i D^2\left(f_{(r)}, f_{(i)}\right) + \right.$$

$$\left. + \frac{1}{R^2}\sum_i\sum_j D^2\left(f_{(i)}, f_{(j)}\right)\right) = \sum_m \left(x_{rm} - \bar{x}_{\cdot m}\right)\left(x_{r'm} - \bar{x}_{\cdot m}\right) = \left\langle x_r, x_{r'} \right\rangle ;$$

specificky: $$c_{rr} = \sum_m \bar{x}_{rm}^2 = d^2\left(x_m, 0\right) = D^2\left(f_r, f\right),$$

$$c_{rr'} = \sum_m \bar{x}_{rm}\,\bar{x}_{r'm} = \left(f_{(r)} - f\right)' D \left(f - f_{(r')}\right),$$ kde $\bar{x}_r$ je vektor x_r centrovaný do těžiště bodů $x_1, \ldots, x_R$, t.j. $\bar{x}_r = \frac{1}{R}\sum_j x_j$, $f = \frac{1}{R}\sum_j f_{(j)}$.

(7) Platí, že $C = \bar{X}'\bar{X}$, kde X je matice RxM tvořená řádky $\bar{x}_r$.

Dále $c_{ii} = \left\|\bar{x}_i\right\|^2$, stopa $C = \sum \left\|\bar{x}_i\right\|^2 = \sum_r\sum_{r'} d^2\left(\bar{x}_r, \bar{x}_{r'}\right) = \sum_r\sum_{r'} d^2\left(x_r, x_{r'}\right) =$

$= \sum_r\sum_{r'} D^2\left(f_{(r)}, f_{(r')}\right)$.

Pro kanonický /spektrální/ rozklad matice c /viz C.R. Rao, [1978]/

(8) $C = \sum_m \lambda_m P_m P_m'$, kde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M > 0$ jsou charakteristická čísla matice C, P_1 , P_2 , ..., P_M jsou charakteristické sloupcové vektory k nim postupně příslušné a platí, že

a) P_m jsou normalizované orthogonální vektory, $P_i' P_j = \delta_{ij}$,

b) pro matice $U_L = \| P_1, P_2, \ldots, P_L \|$, $\Lambda_L = \mathrm{diag}\{\lambda_1, \ldots, \lambda_L\}$ je $U_L \Lambda_L U_L'$ nejlepší aproximace matice C mající hodnost L ve smyslu euklidovské normy $\min_{B:h(B)=L} \| C - B \|^2 = \| C - U_L' \Lambda_L U_L \|^2 = \lambda_{L+1} + \ldots + \lambda_M$,

c) stopa $C = \sum_r \left(\sum_m \bar{x}^2_{rm} \right) = \frac{1}{2R} \sum_r \sum_s d^2(x_r, x_s) = \sum_r d^2(f_{(r)}, f) = \frac{1}{2R} \sum_r \sum_s d^2(f_{(r)}, f_{(s)})$

d) stopa $U_L \Lambda_L U_L' = \lambda_1 + \lambda_2 + \ldots + \lambda_L$.

Je tedy vidět, že matice

(9) $X = U_M \Lambda_M^{\frac{1}{2}}$ je řešením rovnice (1) a

(10) $X = U_L \Lambda_L^{\frac{1}{2}}$ je řešením úlohy (2), a že řádky x_r matice X jsou souřadnice orthonormální báze v E_M resp. v E_L. Podíl m-té souřadnice na rozplýlení bodů x_r je možno měřit koeficientem determinace m-té škály

(11) $$r_m^2 = \frac{\lambda_m}{\sum_M \lambda_m} ,$$

Podíl m-té souřadnice na L-rozměrné škále hodnotu

(12) $$r_m^2(L) = \frac{\lambda_m}{\sum_L \lambda_m} ,$$

Podíl L-rozměrné škály /prvních l dimenzí/ na úplném rozptýlení bodů

(13) $$R_L^2 = \frac{\sum_L \lambda_m}{\sum_M \lambda_m}$$

(Obdobně platí (13) pro jakýkoliv soubor L dimenzí : $\sum_L$ se zamění sumou příslušných charakter. čísel λ_j).

Řešení (9) je, až na násobení vektorů P_m čísly ± 1 jednoznačné v tom smyslu, že posloupnost $U_L' \Lambda_L U_L$ jsou postupně nejlepší aproximace matice C, L=1,2, ..., M vzhledem k euklidovské normě.

Výslednou bázi je možno rotovat, tj. přejít orthonormální transformací T k jiné bázi s vyjádřením vektorů

(14) $X^* = XT$

3. Numerický příklad

Níže uvedená kontingenční tabulka vyjadřuje komparaci 4 skupin pokusných zvířat, které byly vystaveny experimentálnímu působení v různých podmínkách.

Kategorie závislé proměnné odpovídají osmi kombinacím třech zjišťovaných symptomů A, B, C : ABC, AB$\bar{C}$, A$\bar{B}$C, $\bar{A}$BC, A$\bar{B}\bar{C}$, $\bar{A}$B$\bar{C}$, $\bar{A}\bar{B}$C, $\bar{A}\bar{B}\bar{C}$. Komparační analýzu pomocí programu DISTAN provedeme jednak pro běžný nominální typ a jednak pro zobecněný typ daný maticí skórů $D = \| d_{ij} \|$, d_{ij} = počet symptomů u nichž se vyskytuje rozdíl u kategorie "i" a kategorie "j". /např.: $d(ABC, A\bar{B}\bar{C}) = 2$, $d(A\bar{B}\bar{C}, \bar{A}\bar{B}\bar{C}) = 1$ a p./

Rozložení četnosti

skupina	Výskyty symptomů								velikost souboru
	ABC	AB$\bar{C}$	A$\bar{B}$C	$\bar{A}$BC	A$\bar{B}\bar{C}$	$\bar{A}$B$\bar{C}$	$\bar{A}\bar{B}$C	$\bar{A}\bar{B}\bar{C}$	
1	.390	.339	.085	.025	.102	.034	.008	.017	118
2	.055	.407	.251	.090	.085	.050	.040	.020	199
3	.374	.122	.033	.016	.057	.098	.033	.268	123
4	.060	.080	.120	.060	.080	.090	.300	.210	100

Matice D, vytvořující typ zobecněné proměnné

	ABC	AB$\bar{C}$	A$\bar{B}$C	$\bar{A}$BC	A$\bar{B}\bar{C}$	$\bar{A}$B$\bar{C}$	$\bar{A}\bar{B}$C	$\bar{A}\bar{B}\bar{C}$
ABC	0.000	1.000	1.000	1.000	2.000	2.000	2.000	3.000
AB$\bar{C}$	1.000	0.000	2.000	2.000	1.000	1.000	3.000	2.000
A$\bar{B}$C	1.000	2.000	0.000	2.000	1.000	3.000	1.000	2.000
$\bar{A}$BC	1.000	2.000	2.000	0.000	3.000	1.000	1.000	2.000
A$\bar{B}\bar{C}$	2.000	1.000	1.000	3.000	0.000	2.000	2.000	1.000
$\bar{A}$B$\bar{C}$	2.000	1.000	3.000	1.000	2.000	0.000	2.000	1.000
$\bar{A}\bar{B}$C	2.000	3.000	1.000	1.000	2.000	2.000	0.000	1.000
$\bar{A}\bar{B}\bar{C}$	3.000	2.000	2.000	2.000	1.000	1.000	1.000	0.000

Reprezentace distribucí v euklidovském prostoru

nominální případ

M = 3

λ_1 = .161

λ_2 = .148 $\quad \sum\lambda$ = .323

λ_3 = .014

Souřadnice

skupina	1	2	3
1	-.191	-.145	-.074
2	-.194	.222	.051
3	.114	-.231	.066
4	.271	.154	-.043
r^2 (v%)	49.9	45.8	4.4

Pro popis a komparaci distribucí je zapotřebí dvourozměrná reprezentace (L=2). První souřadnice odlišuje (1,2) proti (3,4), druhá (1,3) proti (2,4).

zobecněný případ

M = 3

λ_1 = .625

λ_2 = .020 $\quad \sum\lambda$ = .656

λ_3 = .011

Souřadnice

skupina	1	2	3
1	.469	.009	-.065
2	.184	-.089	.057
3	-.046	.108	.044
4	-.607	-.028	-.036
r^2 (v%)	95.3	3.1	1.6

Distribuce jsou umístěny na jednorozměrné škále; první souřadnice je řadí do pořadí a charakterizuje podíly jejich vzdálenosti.

V obou případech lze nakreslit obrázek: v nominálnim případě dvourozměrný, v zobecněném případě stačí ke zobrazeni umistit skupiny na přimku.

ZÁVĚR

Vzhledem k tomu, že pro podmíněné distribuce kategorizovaných dat je možno zavest smysluplny pojem vzdálenosti, vystačíme po geometrickém zobrazováni do euklidovského prostoru s klasickým mnohorozměrným škálováním.

Vyhodnost metody se projevi ovšem především při velkém počtu podmíněných distribuci.

LITERATURA

Lingoes J.C. [1973] : The Guttman - Lingoes Nonmetric Program Series. Ann Arbor - Michigan, Mathesis Press

Rao R.C. [1978] : Lineární metody statisticke indukce a jeji aplikace, Praha, Academia

Řehák J., Loučková I. [1983]: Klasicke mnohorozměrné škálováni /Aplikace metody DISTAN/, Sociologicky časopis XIX, s. 535-554

Řehák J., Řeháková B. [1979] : Zakladni charakteristiky proměnných s konečným počtem hodnot a distančni analyza jejich rozloženi. Sociologicky časopis XV, s. 214-233

Řehák J., Řeháková B. [1982] : Distančni přistup k analyze kategorizovaných dat a jeho aplikace na problem shody, Robust, Podkost

Řehák J., Řeháková B. [1984]: Analyza kategorizovaných dat v sociologii, Praha, Academia /v tisku/