

Kanonická analýza kontingenčních tabulek*

JOSEF LAUBER

Vysoká škola ekonomická, Praha

JÍŘÍ LINHART

Ústav pro filozofii a sociologii ČSAV,

Praha

V této stati chceme čtenáře seznámit s podstatou jedné z technik analýzy kontingenčních tabulek, dosud — pokud je nám známo — v československé sociologii neužívané. Formálně jde o postup, který prostřednictvím transformace znaků kontingenční tabulky na dvourozměrné normální rozdělení maximalizuje korelační koeficient mezi vyhodnocovanými znaky a určuje pro každou z kategorií obou znaků skóre.

Techniku lze zařadit (z hlediska cíle i použitého matematického aparátu) mezi multivariační přístupy k analýze sociologických dat. Postup je formulován značně obecně, takže jej lze využít pro všechny typy kontingenčních tabulek — pro tabulky vzniklé tříděním nominálních, ordinálních i intervalových znaků, přičemž tabulky mohou mít libovolné rozměry. Interpretace výsledků není vždy jednoduchá, vyžaduje porozumění základům metody. Technika podává především určitý kompaktní pohled na strukturu vazeb mezi kategoriemi v kontingenční tabulce, takže tento pohled by měl být doplňován další, detailnější analýzou.

Matematický základ kanonické analýzy

Metoda je založena zejména na Lancasterově větě [Lancaster 1957; Kendall, Stuart 1973]:

Předpokládejme, že náhodná veličina $F = (X, Y)$ má dvourozměrné normální rozdělení s kovarianční maticí

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$

kde rozptyly $D(X) = \sigma_1^2$ a $D(Y) = \sigma_2^2$

a kovariance, resp. korelace jsou

$$\text{cov}(XY) = \rho\sigma_1\sigma_2 \quad \text{a} \quad \text{corr}(XY) = \rho.$$

Provedeme-li transformace jednotlivých složek uvažovaného dvourozměrného rozdělení F , tj.

$$X' = X'(X) \quad \text{a} \quad Y' = Y'(Y),$$

pak za velmi obecných podmínek (konečnost rozptylů veličin X' a Y') platí

$$\text{corr}(X'Y') \leq |\rho|,$$

* Ve stati je popsána metoda použitá v článku [Charvát, Linhart, Večerník 1978].

neboli korelace transformovaného rozdělení nemůže v absolutní hodnotě převýšit korelaci původního normálního rozdělení.

Lancasterova věta platí i pro nekategorizované veličiny. Uvažujme dále kontingenční tabulku (tab. 1), která obsahuje absolutní četnosti výskytu všech dvojic kategorií znaků X a Y spolu s řádkovými a sloupcovými součty (znak X nechť má r kategorií a znak Y c kategorií).

Tabulka 1

		Znak Y				
		1	2	...	c	Σ
Znak X	1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
	.	.	.	...	.	.
	.	.	.	...	.	.
	r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Σ		$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

Předmětem dalších úvah bude otázka, jakým způsobem připsat jednotlivým kategoriím znaků X a Y určité číselné hodnoty (skóre) tak, aby při daných četnostech nabyl korelační koeficient maxima.

Jinak řečeno — chceme nalézt hodnoty diskrétních náhodných veličin X a Y (pro určitost standardizovaných), tj. čísla (skóre) x_i a y_j ($i = 1, 2, \dots, r$) a ($j = 1, 2, \dots, c$) tak, aby marginální rozdělení

$$\begin{array}{l}
 X: \quad \begin{array}{c|c|c|c|c} x_i & x_1 & x_2 & \dots & x_r \\ \hline P(X = x_i) & \frac{n_{1.}}{n} & \frac{n_{2.}}{n} & & \frac{n_{r.}}{n} \end{array} \\
 Y: \quad \begin{array}{c|c|c|c|c} y_j & y_1 & y_2 & \dots & y_c \\ \hline P(Y = y_j) & \frac{n_{.1}}{n} & \frac{n_{.2}}{n} & & \frac{n_{.c}}{n} \end{array}
 \end{array}$$

měla jednotkové rozptyly

$$D(X) = \sum_{i=1}^r \frac{n_{i.}}{n} x_i^2 = 1,$$

$$D(Y) = \sum_{j=1}^c \frac{n_{.j}}{n} y_j^2 = 1,$$

přičemž korelační koeficient

$$\text{corr}(X, Y) = \sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}}{n} x_i y_j \rightarrow \max$$

Bezprostřední využití uvedené Lancasterovy věty, tj. přímá konstrukce dvourozměrného normálního rozdělení, nepřichází v úvahu. Nalezení optimálních hodnot

x_i a y_j (vzhledem ke kritériu maximalizace korelačního koeficientu) můžeme řešit jako úlohu hledání vázaného extrému.

Označíme

$$f(x_1, x_2, \dots, x_r, y_1, y_2, \dots, y_c) = \sum_i \sum_j \frac{n_{ij}}{n} x_i y_j = \text{corr}(XY),$$

$$g_1(x_1, x_2, \dots, x_r) = \sum_i \frac{n_{i.}}{n} x_i^2 - 1 = 0,$$

$$g_2(y_1, y_2, \dots, y_c) = \sum_j \frac{n_{.j}}{n} y_j^2 - 1 = 0,$$

dostaneme úlohu nalezení maxima funkce f za podmínek $g_1 = g_2 = 0$, kterou lze řešit pomocí Lagrangeových multiplikátorů. Jelikož znalost charakteru podmínek a vztahů, které hodnoty x_i a y_j splňují, je interpretačně důležitá, uvedeme hlavní obraty použité při řešení uvedené extremalizační úlohy.

Vytvoříme nejprve Lagrangeovu funkci:

$$\varphi = f(x_1, x_2, \dots, x_r, y_1, y_2, \dots, y_c) + \lambda g_1(x_1, x_2, \dots, x_r) + \mu g_2(y_1, y_2, \dots, y_c).$$

Nutné podmínky pro extrém jsou:

$$(1) \quad \frac{\partial \varphi}{\partial x_i} = \sum_j \frac{n_{ij}}{n} y_j + 2\lambda \frac{n_{i.}}{n} x_i = 0 \quad (i = 1, 2, \dots, r),$$

$$(2) \quad \frac{\partial \varphi}{\partial y_j} = \sum_i \frac{n_{ij}}{n} x_i + 2\mu \frac{n_{.j}}{n} y_j = 0 \quad (j = 1, 2, \dots, c).$$

Ze vztahů (1) a (2) plyne platnost rovnic

$$\sum_i \frac{\partial \varphi}{\partial x_i} x_i = 0, \quad \sum_j \frac{\partial \varphi}{\partial y_j} y_j = 0,$$

z nichž po dosazení dostaneme

$$\sum_i \sum_j \frac{n_{ij}}{n} x_i y_j + 2\lambda D(X) = 0,$$

$$\sum_i \sum_j \frac{n_{ij}}{n} x_i y_j + 2\mu D(Y) = 0,$$

neboli

$$\text{corr}(XY) = -2\lambda D(X) = -2\mu D(Y),$$

takže při standardizovaných veličinách ($D(X) = D(Y) = 1$) musí platit

$$(3) \quad \text{corr}(XY) = R = -2\lambda = -2\mu.$$

Podmínky (1) a (2) zapíšeme využitím (3) ve formě soustavy lineárních rovnic

$$(4) \quad \sum_j n_{ij} y_j = R n_{i.} x_i, \quad (i = 1, 2, \dots, r),$$

$$(5) \quad \sum_i n_{ij} x_i = R n_{.j} y_j, \quad (j = 1, 2, \dots, c).$$

Je možno dokázat [Kendall 1973], že soustava lineárních rovnic (4) a (5) má netriviální (nenulové) řešení pouze pro hodnoty R , které odpovídají odmocninám vlastních čísel matice, jež vznikne jednoduchou transformací výchozí kontingenční tabulky.

Konkrétně pro existenci netriviálního řešení (4) a (5) musí platit

$$(6) \quad NN'u = R^2u,$$

kde N je matice s prvky

$$(7) \quad n_{ij}^* = \frac{n_{ij}}{\sqrt{n_{i,n_j}}} \quad (i = 1, 2, \dots, r), \\ (j = 1, 2, \dots, c),$$

N' je transponovaná matice N ,

u je sloupceový vektor se složkami $x_i \sqrt{n_i}$ ($i = 1, 2, \dots, r$).

Ze vztahu (6) je patrné, že pro existenci netriviálního řešení je třeba, aby platila rovnost

$$\det [NN' - R^2J] = 0 \quad (J - \text{jednotková matice}),$$

neboli R^2 bylo vlastní číslo matice NN' .

Hodnost matice N je maximálně rovna

$$m = \min(r, c),$$

takže rovněž hodnost matice NN' je nejvýše m . Obecně má matice NN' m různých vlastních čísel $\lambda_0, \lambda_1, \dots, \lambda_{m-1}$, přičemž

$$\lambda_0 = R^2 = 1$$

je vlastním číslem matice NN' vždy, neboť soustava rovnic (4) a (5) má pro $R = 1$ vždy nenulové řešení

$$x_i \equiv K, \quad y_j \equiv K \quad (i = 1, 2, \dots, r), \quad (j = 1, 2, \dots, c),$$

ale s ohledem na podmínky $D(X) = D(Y) = 1$ bychom dostali

$$x_i \equiv 1 \quad \text{a} \quad y_j \equiv 1.$$

Tyto hodnoty nemají žádnou závislost na výchozí kontingenční tabulce, takže jejich znalost nedává žádnou informaci. Vlastní vektor u^0 příslušný k jednotkovému vlastnímu číslu má složky $\sqrt{n_i}$, resp. normalizovaný vlastní vektor má složky $\sqrt{n_i/n}$, které vypovídají pouze o relativních velikostech marginálií.

Další vlastní čísla, seřazená do nerostoucí posloupnosti

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{m-1}$$

spolu s jim odpovídajícími vlastními vektory $u^1, u^2, \dots, u^{m-1}$, již poskytují užitečnou informaci.

Jelikož vlastní vektory mají složky

$$u_i = x_i \sqrt{n_i},$$

dostáváme na základě u^1

$$(8) \quad x_i = u_i^1 / \sqrt{n_i}.$$

Hodnoty skóre y_j pak získáme využitím vztahu (5), neboli

$$(9) \quad y_j = \frac{1}{R n_{.j}} \sum_i n_{ij} x_i \quad (j = 1, 2, \dots, c).$$

Stejným postupem lze ukázat, že platí rovněž vztah

$$N'Nv = R^2v,$$

kde v je vektor se složkami $y_j/\sqrt{n_{.j}}$. Výpočetně je vhodné řešit úlohu nalezení vlastních čísel a vlastních vektorů pro tu z matic NN' a $N'N$, která má menší řád a hodnoty zbývajících kanonických veličin zjistit buď pomocí (5) či (6).

Ukážeme nyní některé vazby popisované techniky na „tradičnější“ přístupy k analýze kontingenčních tabulek. Diagonální prvky matice NN' jsou rovny

$$NN'_{ii} = \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}},$$

takže pro stopu matice NN' platí

$$\text{tr } NN' = \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}}.$$

Jelikož

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = n \left\{ \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1 \right\}$$

a stopa matice je součet vlastních čísel, můžeme veličinu χ^2 vyjádřit ve tvaru

$$(10) \quad \chi^2 = n(\lambda_1 + \lambda_2 + \dots + \lambda_{m+1}).$$

Významnost veličiny χ^2 zjišťujeme pomocí běžných tabulek χ^2 rozdělení pro předem zvolenou hladinu významnosti a počet stupňů volnosti

$$df = (c - 1)(r - 1).$$

Algoritmizace výpočtů a základní vlastnosti kanonických proměnných

Výpočet hodnot x_i a y_j , které jsou připsány jednotlivým kategoriím kontingenční tabulky, vyžaduje takový objem výpočetních prací, že je nezbytné využít výpočetní techniku i pro tabulky malého rozměru. Nejnáročnější je výpočet vlastních čísel a vlastních vektorů symetrické matice NN' , resp. $N'N$. K tomuto účelu je možno doporučit například některý z algoritmů uváděných Wilkinsonem [Wilkinson 1976]. Ve vytvořeném programu ¹⁾ bylo použito QR metody, která se velmi osvědčila.

Jednotlivé kroky výpočtů a některé vlastnosti skóre x_i a y_j ukážeme na jednoduchém příkladě, který bude různým způsobem modifikován.

Kontingenční tabulka (tab. 2) obsahuje výchozí údaje pro kanonickou analýzu.

Matice N (vznikne z tab. 1 použitím vztahu (7)) má tvar

$$N = \begin{bmatrix} 0.6227524 & 0.0668153 & 0.2425356 \\ 0.3636670 & 0.5518254 & 0.1669245 \\ 0.0862796 & 0.2715377 & 0.6160411 \end{bmatrix}$$

a matice NN' obsahuje hodnoty

$$NN' = \begin{bmatrix} 0.4511083 & 0.3038300 & 0.2212856 \\ 0.3038300 & 0.4646288 & 0.2840508 \\ 0.2212856 & 0.2840508 & 0.4606835 \end{bmatrix}$$

Vlastní čísla matice NN' jsou:

$$\lambda_0 = 1.0000, \quad \lambda_1 = 0.2366, \quad \lambda_2 = 0.1398.$$

¹⁾ Program byl vytvořen pro počítač NE 4120, jenž je instalován ve Výpočetním centru VŠE Praha.

Vlastní vektory:

u^0	u^1	u^2
0.5629	0.6328	0.5318
0.6134	0.1114	-0.7819
0.5540	-0.7663	0.3254

Z vektoru u^1 získáme využitím vztahu (8) hodnoty proměnných x_i ($i = 1, 2, 3$):

$$x_1 = 0.05, \quad x_2 = 0.0081, \quad x_3 = -0.0615.$$

Tabulka 2

	1	2	3	Σ
1	110	10	40	160
2	70	90	30	190
3	15	40	100	155
Σ	195	140	170	505

Hodnoty složek vlastních vektorů jsou určeny vždy až na multiplikativní konstantu. Vypočtené hodnoty jsou obvykle normovány, tj. například

$$\sum_{i=1}^3 (n_i^0)^2 = 0,5629^2 + 0,6134^2 + 0,5540^2 = 1.$$

Pro platnost vztahů $D(X) = D(Y) = 1$, tj. speciálně pro

$$D(X) = \sum_{i=1}^r \frac{n_i}{n} x_i^2 = 1$$

je nutné, aby platilo

$$\sum_{i=1}^r \frac{n_i}{n} \frac{(u_i^1)^2}{n_i} = \frac{(u_i^1)^2}{n} = 1,$$

takže je třeba složky normovaných vlastních vektorů vynásobit faktorem $\sqrt{n}$, resp. touto konstantou vynásobit kanonické proměnné, získané z normovaných vektorů. Po této úpravě dostaneme hodnoty kanonických proměnných

$$x_1 = 1,1236, \quad x_2 = 0,1820, \quad x_3 = -1,3820.$$

Hodnoty proměnných y_j získáme ze vztahu (5), konkrétně

$$y_1 = 1,2180, \quad y_2 = -0,4067, \quad y_3 = -1,0629.$$

Hodnotu χ^2 určíme podle vztahu (10)

$$\chi^2 = n(\lambda_1 + \lambda_2 + \dots + \lambda_{m-1}) = 505(0,2366 + 0,1398) = 190,1,$$

což je v souladu s výsledkem získaným „klasickým“ výpočtem

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - o_{ij})^2}{o_{ij}}.$$

Korelační koeficient při skóre x_i a y_j pro údaje v tab. 2 je roven $R^2 = 0,2366 =$

$= 0,4864^2$, což je opět v souladu s hodnotou získanou přímým výpočtem, v němž je každá dvojice hodnot (x_i, y_j) zastoupena s četností n_{ij} .

Tabulka 3

	!1	!2	!3	Σ
1	0,6875	0,0625	0,2500	1,000
2	0,3684	0,4737	0,1579	1,000
3	0,0968	0,2581	0,6451	1,000

Vztahy (4) a (5), které hodnoty x_i a y_j splňují, mohou pomoci při jejich interpretaci. Na základě rovnic (4) a (5) totiž platí

$$\sum_j \frac{n_{ij}}{n_{i.}} y_j = R x_i \quad (i = 1, 2, \dots, r),$$

$$\sum_i \frac{n_{ij}}{n_{.j}} x_i = R y_j \quad (j = 1, 2, \dots, c),$$

kde $\frac{n_{ij}}{n_{i.}}$ a $\frac{n_{ij}}{n_{.j}}$ jsou prvky matic relativních četností (řádkových a sloupcových).

Pro kontingenční tabulku (tab. 2) je matice „řádkových“ relativních četností (tab. 3) a matice „sloupcových“ relativních četností (tab. 4).

Tabulka 4

	1	2	3
1	0,5641	0,0714	0,2353
2	0,3590	0,6429	0,1765
3	0,0769	0,2857	0,5882
Σ	1,0000	1,0000	1,0000

Vztahy (4) a (5) můžeme pak vyjádřit symbolicky takto:

$$y_1 \begin{bmatrix} \text{1. sloupec} \\ \text{tab. 3} \end{bmatrix} + y_2 \begin{bmatrix} \text{2. sloupec} \\ \text{tab. 3} \end{bmatrix} + y_3 \begin{bmatrix} \text{3. sloupec} \\ \text{tab. 3} \end{bmatrix} = R \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

$$x_1[\text{1. řádek tab. 4}] + x_2[\text{2. řádek tab. 4}] + x_3[\text{3. řádek tab. 4}] = \\ = R \cdot [y_1, y_2, y_3],$$

přičemž

$$R = \max_{\{x_i\}, \{y_j\}} \sum_i \sum_j n_{ij} x_i y_j.$$

Z výše uvedených úvah je patrné, že hodnoty kanonických proměnných x_i a y_j jsou nezávislé na lineárním vzrůstu četností v kontingenční tabulce, neboť závisí

pouze na relativních četnostech. Provedeme-li permutaci na pořadí řádků a sloupců tabulky, jsou souhlasně změněna pořadí hodnot x_i a y_j .

Například zaměníme-li u kontingenční tabulky s hodnotami v tab. 2 druhý řádek s prvním a zároveň druhý sloupec s prvním, což odpovídá změnám v pořadí kategorií znaků, dostaneme kontingenční tabulku (tab. 5).

Tabulka 5

	1	2	3	Σ
1	90	70	30	190
2	10	110	40	160
3	40	15	100	155
Σ	140	195	170	505

Pro údaje v tab. 5 dostaneme výpočtem následující hodnoty:

$$\lambda_0 = 1.0000, \quad \lambda_1 = 0.2366, \quad \lambda_2 = 0.1398$$

$$x_1 = 0.1820 \quad y_1 = -0.4067$$

$$x_2 = 1.1236 \quad y_2 = 1.2180$$

$$x_3 = -1.3820 \quad y_3 = -1.0629$$

Ze srovnání těchto hodnot s výsledky pro tab. 2 je patrné, že se vlastní čísla nezměnila a u hodnot skóre x_i a y_j došlo jen ke změnám pořadí: $x_1 \leftrightarrow x_2$, $y_1 \leftrightarrow y_2$.

Hodnota χ^2 a maximální hodnota korelačního koeficientu jsou pro tab. 2 a tab. 5 shodné. Rozborem těchto výsledků dojdeme například k závěru, že znaky v tab. 2 (před permutací) jsou ve smyslu lineární regrese těsněji svázány, i když hodnota χ^2 je pro obě tabulky stejná.

Konkrétní příklady aplikace kanonické analýzy kontingenčních tabulek nalezneme v článku Charvát, Linhart a Večerník, uvedené v seznamu literatury

Literatura

- Charvát, F. — Linhart, J. — Večerník, J.: *Sociální mobilita v socialistických a kapitalistických zemích*. Sociologický časopis 1978, č. 1, s. 40.
 Kendall, M. G. — Stuart, A.: *The Advanced Theory of Statistics*. Vol. 2: *Inference and Relationship*. London, Griffin 1973.
 Lancaster, H. O.: *Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Tables*. Biometrika 1957, 44.
 Wilkinson, J. H. — Reinsch, C.: *Handbook for Automation Computation. Linear Algebra*. Berlin, Springer Verlag 1973. (ruský překlad: Moskva, Mašinostrojenije 1976).

Резюме

Даубер Й. - Лингарт И.: Канонический анализ таблиц сопряженности признаков

В статье описывается математическая модель канонического анализа, который был введен Ланкастером (Lancaster 1957) и который применяется к таблицам сопряженности признаков.

Целью метода является нахождение нумерических значений (скорее) X и Y для строк и столбцов так, чтобы их коэффициент линейной корреляцией Пирсона приобрел максимальное значение. В таблице можно иметь любое число строк и столбцов и переменные в таблице могут репрезентовать номинальный, ординальный или кардинальный характер.

Авторы вводят математическую модель и ее основные свойства. Применение метода можно найти в статье Харват, Лингарт, Вечерник (Charvát, Linhart, Večerník 1978), где анализируется интергенерационная мобильность в некоторых социалистических и капиталистических странах.

Summary

Lauber J. — Linhart J.: Canonical Analysis of Contingency Tables

The paper presents the mathematical model of canonical analysis for contingency tables that has been introduced by Lancaster [1957].

The purpose of the method is to assign quantitative scores X and Y to the values of both variables entering the tables, to give the maximal Pearson linear correlation coefficient of X and Y . Table may consist of arbitrary number of rows and columns that may represent nominal, ordinal or cardinal variables.

Authors present the method and its mathematical properties. They refer to an analysis of intergenerational mobility in selected socialist and non-socialist countries based on this method and described in Charvát, Linhart, Večerník [1978].