Kanonická korelační analýza (*Canonical correlation analysis*)

Kanonická korelační analýza patří mezi metody vícerozměrné analýzy dat. Jejím cílem je nalézt vztahy mezi dvěma skupinami proměnných. Z určitého pohledu se jedná o rozšíření vícenásobné lineární regrese, při níž se hledá vztah jedné (tzv. vysvětlované) proměnné se skupinou (tzv. vysvětlujících) proměnných. Kanonická korelační analýza byla navržena v roce 1935 Hotellingem a jejím principem je hledání lineární kombinace jedné skupiny *p* proměnných, která nejlépe koreluje s lineární kombinací druhé skupiny *q* proměnných. Příkladem může být hledání vztahů mezi proměnnými popisujícími vlohy studentů a proměnnými odrážejícími výsledky v jednotlivých předmětech nebo zjišťování, jak souvisí parametry počasí (průměrné denní srážky, vlhkost, počet hodin slunečního záření) s výnosem plodin (výška rostlin, hmotnost po usušení, počet listů).

V průběhu kanonické korelační analýzy se postupně hledají lineární kombinace proměnných z každé skupiny (podobně jako u analýzy hlavních komponent – *principal component analysis* – PCA), tedy vytváří se nové proměnné, tzv. kanonické proměnné (*canonical variables*), které vedou k maximálním vzájemným korelacím mezi skupinami (na rozdíl od PCA, kde se vytváří nové proměnné za účelem vysvětlení co nejvíce variability původních dat, která se uvažují jako celek). Jedná se o krokový proces, kdy se v prvním kroku hledá lineární kombinace první skupiny proměnných a lineární kombinace druhé skupiny proměnných, jejichž korelace je maximální. Tyto lineární kombinace tvoří první pár kanonických proměnných, které jsou základem nového souřadnicového systému. Jejich korelaci označujeme jako tzv. první kanonickou korelaci (*canonical correlation*). V dalších krocích se hledají další lineární kombinace skupin proměnných, tedy další kanonické proměnné, tak, aby měly co největší vzájemnou druhou, třetí, ... korelaci a přitom byly nekorelované s kanonickými proměnnými získanými v předchozích krocích.

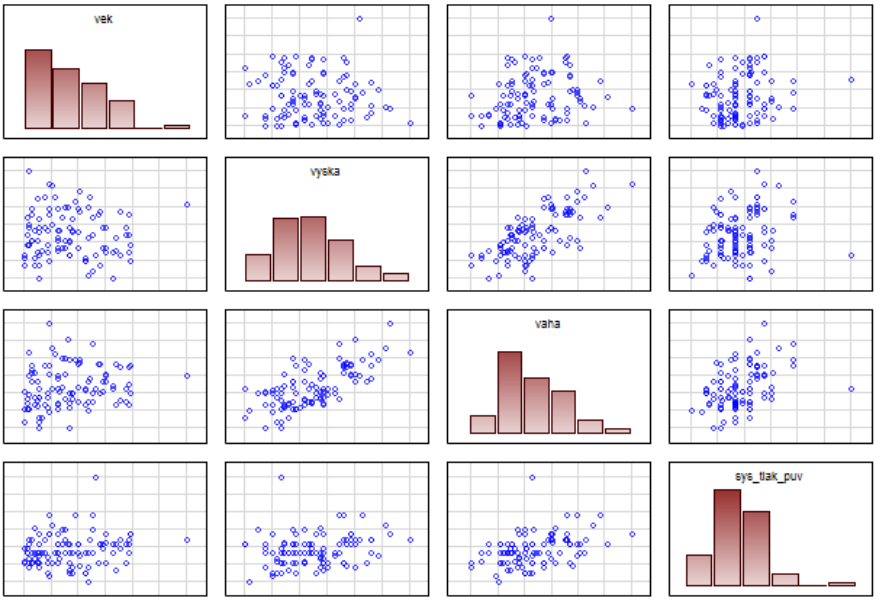
Kanonická korelační analýza je zvláště užitečná v situacích, kdy jsou proměnné uvnitř skupin korelované, takže nemá smysl vyhodnocovat korelace jednotlivých proměnných odděleně, protože by se zanedbala jejich vzájemná vnitřní korelace. Kromě nalezení vztahů v datech se tato metoda využívá i při snižování dimenze dat, pokud jsou skupiny původních proměnných velké a účelem je nalézt malý počet nových kanonických proměnných, které postihují v maximální míře korelace mezi původními skupinami proměnných.

Předpoklady kanonické korelační analýzy

Kanonická korelační analýza předpokládá pouze lineární závislost mezi proměnnými i mezi skupinami proměnných. Je tudíž nutné vyšetřit grafy každého páru proměnných a prověřit linearitu a odlehlé hodnoty. Nápomocný může být např. maticový graf, jenž slouží ke znázornění závislostí dvojic proměnných pomocí bodových (tečkových) grafů, které jsou uspořádány do matice (Obr. 1). Při nelineárním vztahu by měla být jedna nebo obě proměnné vhodně transformovány.

Kanonická korelační analýza nevyžaduje předpoklad normality proměnných, normalita je požadována pouze pro provedení testů statistické významnosti kanonických korelací. Tato metoda tedy může být použita i pro nenormálně rozdělené proměnné, pokud forma rozdělení (např. silně zešikmená) nezkresluje korelaci s ostatními proměnnými. Vzhledem k tomu, že je metoda založena na výpočtu korelací, není tudíž vhodné ji aplikovat na kategoriální data s malým počtem kategorií.

Dalším důležitým předpokladem je dostatečně velký soubor vstupních dat, abychom zabránili problémům příliš malého výběru a z toho plynoucího možného zkreslení výsledků analýzy. Ideálně by mělo být alespoň 10 pozorování (subjektů či objektů) na 1 proměnnou. Rovněž je nutné předem ověřit, zda nejsou v datech chybějící hodnoty, protože většina implementací metody CCA ve statistických softwarech vyřadí z analýzy všechny subjekty, u nichž je alespoň jedna chybějící hodnota, což může vést k velkému snížení velikosti souboru a případně i zkreslení analýzy, pokud data nechybí náhodně.



Obr. 1. Ukázka maticového grafu.

Výpočet kanonické korelační analýzy

Označme první skupinu *p* proměnných jako $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ a druhou skupinu *q* proměnných jako $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_q$.

V kanonické korelační analýze se tvoří první dvojice kanonických proměnných $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$ tak, že $\boldsymbol{U}_1$ se vytvoří jako lineární kombinace proměnných $\boldsymbol{y}_i$ a $\boldsymbol{V}_1$ se vytvoří z proměnných $\boldsymbol{x}_i$:

$$\boldsymbol{U}_1 = a_{11}\,\boldsymbol{y}_1 + a_{12}\,\boldsymbol{y}_2 + \cdots + a_{1q}\,\boldsymbol{y}_q;$$

$$\boldsymbol{V}_1 = b_{11}\,\boldsymbol{x}_1 + b_{12}\,\boldsymbol{x}_2 + \cdots + b_{1p}\,\boldsymbol{x}_p.$$

Koeficienty a_{1i} a b_{1i}, tzv. kanonické váhy (*canonical coefficients*), se vyhledávají tak, aby kanonické proměnné $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$ byly maximálně korelované. Korelaci ρ_1 mezi $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$ nazýváme první kanonickou korelací a je to nejsilnější možná korelace mezi lineárními kombinacemi obou skupin proměnných.

Druhá dvojice kanonických proměnných $\boldsymbol{U}_2$ a $\boldsymbol{V}_2$ se obdobně tvoří jako lineární kombinace proměnných $\boldsymbol{y}_i$ resp. $\boldsymbol{x}_i$ tak, aby měly opět co největší korelaci, tzv. druhou kanonickou korelaci ρ_2, a přitom splňovaly podmínku, že $\boldsymbol{U}_2$ i $\boldsymbol{V}_2$ jsou nekorelované s $\boldsymbol{U}_1$ a $\boldsymbol{V}_1$. Druhá kanonická korelace je vždy menší nebo rovna první kanonické korelaci.

Dále se obdobně vytváří další dvojice kanonických proměnných $\boldsymbol{U}_3$ a $\boldsymbol{V}_3$ atd. s co největší třetí atd. kanonickou korelací a které jsou nekorelované se všemi ostatními kanonickými proměnnými. Maximální počet dvojic kanonických proměnných a jim odpovídajících kanonických korelací je roven menšímu z čísel *p* a *q* (tzn. menšímu z počtu proměnných ve skupinách).

Pro určení koeficientů $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \dots, a_{iq})$ a $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \dots, b_{ip})$ jednotlivých dvojic kanonických proměnných, se maximalizuje korelační koeficient:

$$\boldsymbol{\rho}_i = \frac{\boldsymbol{b}_i^T \boldsymbol{S}_{xy} \boldsymbol{a}_i}{\sqrt{\boldsymbol{b}_i^T \boldsymbol{S}_{xx} \boldsymbol{b}_i \;\; \boldsymbol{a}_i^T \boldsymbol{S}_{yy} \boldsymbol{a}_i}},$$

kde $\boldsymbol{S}_{xx} = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X}$ je kovarianční matice proměnných $\boldsymbol{x}_i$, které jsou centrované (tzn. je u nich odečten průměr) a jsou uspořádané do matice $\boldsymbol{X}$ rozměru $(n \times p)$, přičemž *n* je počet subjektů či objektů; $\boldsymbol{S}_{yy} = \frac{1}{n-1} \boldsymbol{Y}^T \boldsymbol{Y}$ je kovarianční matice centrovaných proměnných $\boldsymbol{y}_i$ uspořádaných do matice $\boldsymbol{Y}$ rozměru $(n \times q)$; a $\boldsymbol{S}_{xy} = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{Y} = \boldsymbol{S}_{yx}^T$ je matice kovariancí centrovaných proměnných $\boldsymbol{x}_i$ a $\boldsymbol{y}_i$.

Pro maximalizaci korelačních koeficientů se využívá metoda tzv. Lagrangeových součinitelů. Postupným odvozováním se dospěje k tomu, že pokud provedeme rozklad matice $\boldsymbol{S}_{xx}^{-1} \boldsymbol{S}_{xy} \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx}$ na vlastní čísla a vlastní vektory, získáme matici $\boldsymbol{\lambda}$, která na diagonále obsahuje vlastní čísla odpovídající čtvercům kanonických korelací. Dále získáme matici $\boldsymbol{V}$, jejíž sloupce jsou vlastní vektory, přičemž první sloupec odpovídá hodnotám koeficientů $\boldsymbol{b}_1$ první kanonické proměnné $\boldsymbol{V}_1$, druhý sloupec koeficientům $\boldsymbol{b}_2$ druhé kanonické proměnné $\boldsymbol{V}_2$ atd. Koeficienty $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ... kanonických proměnných $\boldsymbol{U}_1, \boldsymbol{U}_2, \dots$ se pak určí ze vztahu $\boldsymbol{a}_i = \frac{1}{\sqrt{\lambda_i}} \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx} \boldsymbol{b}_i$.

Souřadnice subjektů v novém prostoru, tzv. kanonická skóre (*canonical scores*), lze vypočítat jako $\boldsymbol{V} = \boldsymbol{XB}$ a $\boldsymbol{U} = \boldsymbol{YA}$, kde $\boldsymbol{B}$ obsahuje jako sloupce koeficienty $\boldsymbol{b}_i$ a $\boldsymbol{A}$ je tvořena sloupcovými vektory koeficientů $\boldsymbol{a}_i$. Dále lze vypočítat tzv. matice zátěží $\boldsymbol{L}_x = \boldsymbol{S}_{xx} \boldsymbol{B}$ a $\boldsymbol{L}_y = \boldsymbol{S}_{yy} \boldsymbol{A}$, což jsou ve skutečnosti korelace mezi kanonickými proměnnými a původními proměnnými.

Pro ověření výsledků kanonické korelační analýzy je vhodné vytvořit dva dílčí podsoubory subjektů či objektů, provést analýzu s každým podsouborem odděleně a následně porovnat zátěže kanonických proměnných atd. Když je nalezen velký rozdíl, je nutno analyzovat, čím je způsoben.

Interpretace kanonických proměnných

Kanonické proměnné jsou uměle vytvořené proměnné, které zpravidla nemají přímé vysvětlení a je nutno je interpretovat (obdobně jako u PCA). Důležitost každé proměnné se vyhodnocuje ze dvou hledisek. Určujeme intenzitu vztahu mezi kanonickou proměnnou $\boldsymbol{U}_i$ a původními proměnnými $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_q$ a rovněž kanonickou proměnnou $\boldsymbol{V}_i$ a původními proměnnými $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p$. Dále také vyjadřujeme sílu vztahu mezi oběma kanonickými proměnnými $\boldsymbol{U}_i$ a $\boldsymbol{V}_i$.

Intenzita vztahu mezi oběma kanonickými proměnnými je vyjádřena prostřednictvím kanonické korelace. Čtverec kanonických korelací (čili koeficient determinace) představuje velikost sdíleného rozptylu mezi kanonickými proměnnými. Zpravidla se analyzují pouze ty dvojice kanonických proměnných, jejichž kanonické korelace jsou statisticky významné. Pro určení statistické významnosti je možno použít např. Wilkovo lambda, k němuž lze vypočítat p-hodnotu. P-hodnota menší než 0,05 pak ukazuje na statistickou významnost kanonické korelace. Wilkovo lambda je interpretováno opačně než koeficient determinace (tzn. hodnota blízká 0 ukazuje na silný vztah, zatímco hodnota blízká 1 na slabý vztah). Kromě statistické významnosti je nutné se dívat i na samotnou velikost kanonické korelace. Obecně přijatelný návod o vhodné velikosti kanonických korelací však bohužel neexistuje.

Sílu vztahu mezi kanonickou proměnnou a původními proměnnými můžeme posuzovat pomocí kanonických vah či kanonických zátěží:

- U kanonických vah vyšetřujeme znaménko a velikost váhy. Původní proměnné s váhami stejného znaménka vykazují přímý vztah, zatímco proměnné, jejichž váhy mají opačné znaménko, vykazují inverzní vztah. Proměnné s relativně velkými váhami přispívají více do kanonických proměnných a naopak. Malá váha tedy zpravidla znamená, že odpovídající původní proměnná je v určování kanonické proměnné nevýznamná. Další možností však je, že je nízká váha způsobena tím, že je mezi původními proměnnými vysoká multikolinearita.
- Kanonické zátěže měří lineární korelaci mezi původní proměnnou a kanonickou proměnnou, tzn. odrážejí rozptyl, který sdílejí původní proměnné s kanonickou proměnnou. Multikolinearita tedy kanonické zátěže nijak nezkresluje.

Je-li skupina proměnných v jedné kanonické proměnné nekorelovaná, kanonické zátěže jsou rovny standardizovaným kanonickým vahám. Jsou-li však některé z původních proměnných v dané skupině silně korelovány, pak jsou zátěže a váhy zcela rozličné. Např. pokud jsou dvě proměnné $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ silně kladně korelovány a každá je pozitivně korelována s kanonickou proměnnou, může se stát, že jedna kanonická váha bude kladná a jedna záporná, zatímco kanonické zátěže budou obě kladné, jak by se dalo očekávat. Pokud se takto podstatně liší kanonické váhy od kanonických zátěží, je nutné zjistit příčinu.

Příklad

Chceme zjistit, zda a jak souvisí charakteristiky práce (5 proměnných: zpětná vazba, významnost úkolu, variabilita úkolů, provedení celého úkolu, autonomie) se spokojeností s prací (7 proměnných: spokojenost nadřízeného, spokojenost s budoucností práce, finanční spokojenost, spokojenost s pracovní zátěží, prestiž firmy, spokojenost s druhem práce, všeobecná spokojenost) u 784 zaměstnanců (Dunham 1997). Z dat byla vypočítána korelační matice všech proměnných:

1.0					.33	.32	.20	.19	.30	.37	.21
.49	1.0				.30	.21	.16	.08	.27	.35	.20
.53	.57	1.0			.31	.23	.14	.07	.24	.37	.18
.49	.46	.48	1.0		.24	.22	.12	.19	.21	.29	.16
.51	.53	.57	.57	1.0	.38	.32	.17	.23	.32	.36	.27
.33	.30	.31	.24	.38	1.0						
.32	.21	.23	.22	.32	.43	1.0					
.20	.16	.14	.12	.17	.27	.33	1.0				
.19	.08	.07	.19	.23	.24	.26	.25	1.0			
.30	.27	.24	.21	.32	.34	.54	.46	.28	1.0		
.37	.35	.37	.29	.36	.37	.32	.29	.30	.35	1.0	
.21	.20	.18	.16	.27	.40	.58	.45	.27	.59	.31	1.0

Dále bylo vypočítáno 5 kanonických korelací (protože minimum z 5 a 7 je 5) a kanonické proměnné:

Canonical Variate Coefficients and Canonical Correlations

	Standardized variables							Standardized variables						
	$z_1^{(1)}$	$z_2^{(1)}$	$z_3^{(1)}$	$z_4^{(1)}$	$z_5^{(1)}$	$\widehat{\rho_1^*}$		$z_1^{(2)}$	$z_2^{(2)}$	$z_3^{(2)}$	$z_4^{(2)}$	$z_5^{(2)}$	$z_6^{(2)}$	$z_7^{(2)}$
$\hat{\mathbf{a}}_1'$:	.42	.21	.17	−.02	.44	.55	$\hat{\mathbf{b}}_1'$:	.42	.22	−.03	.01	.29	.52	−.12
$\hat{\mathbf{a}}_2'$:	−.30	.65	.85	−.29	−.81	.23	$\hat{\mathbf{b}}_2'$:	.03	−.42	.08	−.91	.14	.59	−.02
$\hat{\mathbf{a}}_3'$:	−.86	.47	−.19	−.49	.95	.12	$\hat{\mathbf{b}}_3'$:	.58	−.76	−.41	−.07	.19	−.43	.92
$\hat{\mathbf{a}}_4'$:	.76	−.06	−.12	−1.14	−.25	.08	$\hat{\mathbf{b}}_4'$:	.23	.49	.52	−.47	.34	−.69	−.37
$\hat{\mathbf{a}}_5'$:	.27	1.01	−1.04	.16	.32	.05	$\hat{\mathbf{b}}_5'$:	−.52	−.63	.41	.21	.76	.02	.10

Například první kanonická korelace je 0,55, což je vcelku silná korelace, která ukazuje na to, že existuje vztah mezi spokojeností s prací a charakteristikami práce. První pár kanonických proměnných je:

$$\hat{U}_1 = .42z_1^{(1)} + .21z_2^{(1)} + .17z_3^{(1)} - .02z_4^{(1)} + .44z_5^{(1)}$$
$$\hat{V}_1 = .42z_1^{(2)} + .22z_2^{(2)} - .03z_3^{(2)} + .01z_4^{(2)} + .29z_5^{(2)} + .52z_6^{(2)} - .12z_7^{(2)}$$

Z hodnot kanonických vah je patrné, že kanonická proměnná $\widehat{\boldsymbol{U}}_1$ je založena především na 1. a 5. původní proměnné, tzn. na zpětné vazbě a autonomii. Kanonická proměnná $\widehat{\boldsymbol{V}}_1$ je reprezentována 1., 2., 5. a 6. původní proměnnou, tedy prestiží firmy, spokojeností nadřízeného a spokojeností s budoucností a druhem práce.

Dále byly vypočteny kanonické zátěže:

Sample Correlations Between Original Variables and Canonical Variables

	Sample canonical variates			Sample canonical variates	
$\mathbf{X}^{(1)}$ variables	$\hat{U}_1$	$\hat{V}_1$	$\mathbf{X}^{(2)}$ variables	$\hat{U}_1$	$\hat{V}_1$
1. Feedback	.83	.46	1. Supervisor satisfaction	.42	.75
2. Task significance	.74	.41	2. Career-future satisfaction	.35	.65
3. Task variety	.75	.42	3. Financial satisfaction	.21	.39
4. Task identity	.62	.34	4. Workload satisfaction	.21	.37
5. Autonomy	.85	.48	5. Company identification	.36	.65
			6. Kind-of-work satisfaction	.44	.80
			7. General satisfaction	.28	.50

Z hodnot zátěží vyplývá, že všech 5 charakteristik práce má vysoké a vcelku obdobné korelace s kanonickou proměnnou $\widehat{\boldsymbol{U}}_1$, tedy tato proměnná může být interpretována jako „index charakteristiky práce“. Je tu tedy rozdíl v interpretaci pomocí kanonických vah a zátěží způsobený vcelku vysokými korelacemi mezi původními proměnnými. Zatímco interpretace kanonické proměnné $\widehat{\boldsymbol{V}}_1$ zůstává stejná jako při použití kanonických vah, protože korelace jsou nejvyšší u stejných čtyř původních proměnných.

Literatura

Dunham R.B., Reactions to Job Characteristics: Moderating Effects of the Organization, *Academy of Management Journal*, Vol. 20, No. 1, pp. 42-65, 1977.

Everitt B. & Hothorn T., An Introduction to Applied Multivariate Analysis with R, 2011.

Johnson R. A. & Wichern D. W., Applied Multivariate Statistical Analysis, 6th Edition, 2008.

Meloun M. & Militký M., Interaktivní statistická analýza dat, 2012.