Vzorová úloha 4.5 *Ukázka pojmů a podstaty kanonické korelační analýzy*

Kanonická korelační analýza se často užívá v psychologii a pedagogice, např. k validování testu inteligence. Postup je pak takový, že dva testy, standardní test a nový test jsou aplikovány na jedny a tytéž osoby. Na dva testy, každý o 10 bodovaných otázkách (0 až 100 bodů), odpovědělo 15 studentů a byly tak získány dvě matice dat *TEST1* (rozměru 15 ×10) a *TEST2* (rozměru 15 ×10). Kanonická korelace nalezne pro 15 studentů hodnoty váženého průměru z 10 bodovaných odpovědí standardního testu $U_{1,i}$, i = 1, ..., 15. Tyto pak koreluje s 15 hodnotami váženého průměru 10 bodovaných odpovědí nového a validovaného testu $V_{1,i}$, i = 1, ..., 15. Váhy jsou konstruovány tak, že maximalizují korelaci mezi těmito dvěma průměry. Jde o korelaci mezi těmito dvěma testy, když máme k dispozici 15 dvojic průměrů $\{U, V\}$. Vyčíslená korelace se nazývá *první kanonický korelační koeficient.*

Můžeme sestrojit i jiný soubor vážených průměrů (a to třeba jen pro vybrané otázky), nesouvisející s prvním souborem, a vypočítat jejich korelaci. Proces se opakuje tolikrát, až se počet kanonických korelací rovná počtu proměnných v menší ze dvou skupin.

Budeme nadále rozlišovat *původní proměnné x, y* a *kanonické proměnné V, U.* Kanonické proměnné jsou proměnné, které byly sestaveny z vážených průměrů původních proměnných, např. z odpovědí na 10 otázek testu (původní proměnné) se vytvoří kanonická proměnná, která představuje pro každého studenta jediné číslo jako výsledek dotyčného testu. Soubor kanonických proměnných $\boldsymbol{U}$ vznikl z původních proměnných $\boldsymbol{y}$**.** Soubor kanonických proměnných $\boldsymbol{V}$ vznikl z původních proměnných $\boldsymbol{x}$. V průběhu kanonické korelace by mělo být vzato v úvahu následujících několik bodů:

1. *Určení počtu párů kanonických proměnných:* počet možných párů je roven menšímu číslu z počtu proměnných v každém souboru.
2. *Kanonické proměnné je nutno také interpretovat:* stejně jako ve faktorové analýze pracujeme i zde s matematicky umělými proměnnými, které je často obtížné fyzikálně vysvětlit.
3. *Důležitost každé proměnné musí být vyhodnocena ze dvou hledisek:* musíme určit intenzitu vztahu mezi kanonickou proměnnou U a původní proměnnou y nebo proměnnými V a x, ze které byla kanonická proměnná vytvořena. Musíme rovněž vyjádřit intenzitu vztahu mezi oběma kanonickými proměnnými V a U.
4. *Pozornost je třeba věnovat velikosti výběru*: v sociálních vědách potřebujeme obvykle 10 experimentálních hodnot na jeden neznámý parametr, v přírodních vědách trochu méně.

Normalita a odlehlé body. Kanonická korelace nemá silné požadavky na normalitu. Odlehlé hodnoty však mohou zničit průběh výpočtu či přinést velké komplikace fyzikálně, biologicky či jinak.

Linearita. Kanonická korelační analýza předpokládá pouze lineární závislost mezi proměnnými. Pečlivě je třeba vyšetřit grafy každého páru proměnných a prověřit linearitu a odlehlé body. Kanonická korelace je založena na korelaci mezi dvěma soubory proměnných. Korelační matice všech proměnných lze pak rozdělit na čtyři části:

1. $\boldsymbol{R}_{xx}$. Jde o korelaci mezi proměnnými $\boldsymbol{x}$.
2. $\boldsymbol{R}_{yy}$. Jde o korelaci mezi proměnnými $\boldsymbol{y}$.
3. $\boldsymbol{R}_{xy}$. Jde o korelaci mezi proměnnými $\boldsymbol{x}$ a $\boldsymbol{y}$.
4. $\boldsymbol{R}_{yx}$. Jde o korelaci mezi proměnnými $\boldsymbol{y}$ a $\boldsymbol{x}$.

Kanonická korelace může být vyjádřena s využitím metody SVD (Singlular Value Decomposition) matice $\boldsymbol{C}$, kde $\boldsymbol{C} = \boldsymbol{R}_{yy}^{-1/2}\,\boldsymbol{R}_{yx}\,\boldsymbol{R}_{xx}^{-1}\,\boldsymbol{R}_{xy}$. V SVD rozkladu matice $\boldsymbol{C}$ vztahem $\boldsymbol{C} = \hat{a}_y^T\,\boldsymbol{\Lambda}\,\hat{a}_y$ je diagonální matice $\boldsymbol{\lambda}$ vlastních čísel vytvořena z vlastních čísel matice $\boldsymbol{C}$. Pak j-té vlastní číslo λ_j matice $\boldsymbol{C}$ je rovno čtverci j-té kanonické korelace, která se nazývá r^2_j. Odtud j-tá kanonická korelace je druhou odmocninou z j-tého vlastního čísla matice $\boldsymbol{C}$.

Dva soubory kanonických koeficientů (podobně jako regresních koeficientů) se užívají pro každou kanonickou korelaci: jeden pro proměnné $\boldsymbol{x}$ a druhý pro proměnné $\boldsymbol{y}$. Tyto kanonické koeficienty jsou definovány

$$\boldsymbol{a} = (\boldsymbol{R}_{yy}^{-1/2})^T\,\hat{\boldsymbol{a}}_y \quad , \qquad \boldsymbol{b} = \boldsymbol{R}_{xx}^{-1}\,\boldsymbol{R}_{xy}\,\boldsymbol{a}\,\lambda_j^{-1/2},$$

kde $\hat{\boldsymbol{a}}_y$ je normovaná matice vlastních vektorů pro y. Kanonické skóre pro $\boldsymbol{V}$ a $\boldsymbol{U}$ vzniklo vynásobením standardizovaných dat (od prvků se odečte průměr a výsledek se vydělí směrodatnou odchylkou) maticí kanonických koeficientů $\boldsymbol{V} = \boldsymbol{Z}_x\,\boldsymbol{b}$ a $\boldsymbol{U} = \boldsymbol{Z}_y\,\boldsymbol{a}$, kde $\boldsymbol{Z}_x$ a $\boldsymbol{Z}_y$ představují standardizovaná data $\boldsymbol{X}$ a $\boldsymbol{Y}$.

Abychom pomohli interpretaci kanonických proměnných, vyčíslíme také *matice zátěží* dle vztahů:

$$\boldsymbol{L}_x = \boldsymbol{R}_{xx}\,\boldsymbol{b} \quad \text{a} \quad \boldsymbol{L}_y = \boldsymbol{R}_{yy}\,\boldsymbol{a} \quad .$$

Jsou to vlastně korelace mezi původními proměnnými a kanonickými proměnnými.

Postup kanonické korelační analýzy

1. *Bodové odhady parametrů polohy a rozptýlení všech proměnných:* vyčíslí se aritmetický průměr a směrodatná odchylka pro všechny proměnné.
2. *Korelační koeficienty všech původních proměnných:* vyčíslí se párové korelační koeficienty mezi všemi proměnnými.
3. *Kanonické korelace:* vedle kanonických korelačních koeficientů obsahuje řadu pomocných statistik k interpretaci kanonické korelace.
4. *Objasněná proměnlivost v datech:* obsahuje procento proměnlivosti v každém souboru proměnných, vysvětlovaných jiným souborem proměnných.
5. *Standardizované kanonické parametry pro kanonické proměnné Y a X:* koeficienty slouží k interpretaci proměnných v hodnotě váhy u každé proměnné.
6. *Korelace párů původní proměnné vs. kanonická proměnná:* napomůže snadnější interpretaci kanonických proměnných. Je-li kanonická proměnná silně korelovaná s původní proměnnou, má pak i stejnou či podobnou interpretaci.
7. *Tabulka kanonického skóre pro všechny objekty:* obsahuje kanonické skóre každého souboru proměnných pro každý řádek úplných dat. Hodnoty lze také vynést do grafu.
8. *Grafy kanonického skóre pro všechny objekty:* grafy ukazují na vztah mezi každým párem kanonických proměnných. Korelační koeficient v prvním grafu je *první kanonický korelační koeficient.*

Vzorová úloha 4.6: *Postup kanonické korelační analýzy*

V úloze **S4.18** *Testy IQ* bylo vyšetřeno 15 respondentů (čili 15 objektů) pěti rozličnými testy a vyčíslena hodnota *IQ* (čili dohromady šesti původními proměnnými) za účelem zjištění objektivní hodnoty výsledného inteligenčního kvocientu. Každý z testů obsahoval 10 bodovaných otázek (0 až 100 bodů), na které odpovědělo 15 studentů, matice *TEST1* až *TEST5* a *IQ* byly velikosti (15 ×10). Kanonická korelace nalezne 15 hodnot váženého průměru z 10 bodovaných odpovědí každého testu a koreluje je s 15 hodnotami váženého průměru 10 bodovaných odpovědí jiného testu. Jde o korelaci vždy mezi dvojicí testů, když je k dispozici 15 dvojic vážených průměrů $\{X, Y\}$. Pokuste se vyšetřit tři vybrané testy v závislosti na prvních třech testech čili pokuste se popsat závislostí (*TEST4, TEST5, IQ*) = *f(TEST1, TEST2, TEST3)*.

Řešení: výstup Canonical correlation (NCSS2000) pro nestandardizovaná data

1. Popisné statistiky polohy a rozptýlení:

Typ	Proměnná	Průměr	Směrodatná odchylka	Úplné řádky bez chybějících hodnot
U	*Test4*	65.53333	13.95332	15
U	*Test5*	69.93333	16.15314	15
U	*IQ*	104.3333	11.0173	15
V	*Test1*	67.93333	17.39239	15
V	*Test2*	61.4	19.39735	15
V	*Test3*	72.33334	14.73415	15

Obsahuje popisné statistiky pro všechny proměnné. Kontroluje, zda průměry dosahují "přijatelných" hodnot a zda počet úplných "neděravých" řádků je správný.

2. Korelační koeficienty párů všech původních proměnných:

	Test4	*Test5*	*IQ*	*Test1*	*Test2*	*Test3*
Test4	1.000000	-0.172864	0.371404	0.753937	0.719623	-0.140941
Test5	-0.172864	1.000000	-0.058064	0.013967	-0.281449	0.347335
IQ	0.371404	-0.058064	1.000000	0.225648	0.240651	0.074070
Test1	0.753937	0.013967	0.225648	1.000000	0.100018	-0.260801
Test2	0.719623	-0.281449	0.240651	0.100018	1.000000	0.057232
Test3	-0.140941	0.347335	0.074070	-0.260801	0.057232	1.000000

Obsahuje jednoduché korelace čili Pearsonovy korelační koeficienty mezi všemi proměnnými.

3. Kanonické korelace:

Index proměnné	Kanonická korelace	*D*	F-test	Čitatel SV	Jmen. SV	Spočtená hlad. významnosti	Wilkovo Lambda
1	0.995600	0.991219	16.58	9	22	0.000000	0.006819
2	0.467461	0.218519	0.67	4	20	0.617695	0.776503

3	0.079810	0.006370	0.07	1	11	0.795498	0.993630

F-test testuje, zda tato kanonická korelace a všechny následné jsou nulové.

Obsahuje kanonické korelace a veškeré podpůrné informace, potřebné k interpretaci. **Index proměnné** je pořadové číslo kanonické korelace. Je třeba si uvědomit, že první korelace bude vždy největší. **Kanonická korelace** je hodnota kanonického korelačního koeficientu. Koeficient má stejné vlastnosti jako jiné korelace. Rozsah je od -1 do +1, přičemž 0 značí nízkou korelaci a absolutní hodnota blízká jedné pak perfektní korelaci. ***D*** značí čtverec kanonického korelačního koeficientu (čili koeficient determinace) a udává hodnotu těsnosti proložení lineárního modelu kanonické proměnné *Y* na odpovídající *X* kanonické proměnné. ***F*-test:** hodnota *F*-testu při testování statistické významnosti Wilkova lambda, odpovídajícího řádku a všech hodnot pod tímto řádkem. V tomto případě první *F*-hodnota testuje významnost první, druhé a třetí kanonické korelace, zatímco druhá *F*-hodnota testuje významnost pouze druhé a třetí. **Čitatel SV:** počet stupňů volnosti v čitateli. **Jmenovatel SV:** počet stupňů volnosti ve jmenovateli. **Spočtená hladina významnosti:** hodnota spočtené hladiny významnosti čili pravděpodobnosti pro výše vyčíslené *F*-testační kritérium. Hodnota blízko nule ukazuje na významnou kanonickou korelaci. Hranice $\alpha = 0.05$ bývá často užívána k určení statistické významnosti, tj. hodnoty pravděpodobnosti větší než 0.05 ukazující na statistickou nevýznamnost. **Wilkovo lambda:** hodnota Wilkova lambda pro kanonickou korelaci tohoto řádku představuje vlastně vícerozměrné zobecnění *D*. Wilkovo lambda je interpretováno opačně než *D*, tedy hodnota blízká nule ukazuje na vysokou korelaci a hodnota blízká 1 na nízkou korelaci.

4. Objasněná proměnlivost v datech:

Index kanonické proměnné	Proměnlivost v těchto proměnných	Objasněno těmito proměnnými	Procento objasnění jednotlivě	Procento objasnění kumulativně	Kanonický koeficient determinace
1	*U*	*U*	37.6	37.6	0.9912
2	*U*	*U*	32.1	69.7	0.2185
3	*U*	*U*	30.3	100.0	0.0064
1	*U*	*V*	37.2	37.2	0.9912
2	*U*	*V*	7.0	44.3	0.2185
3	*U*	*V*	0.2	44.5	0.0064
1	*V*	*U*	37.1	37.1	0.9912
2	*V*	*U*	5.4	42.5	0.2185
3	*V*	*U*	0.2	42.8	0.0064
1	*V*	*V*	37.4	37.4	0.9912
2	*V*	*V*	24.8	62.2	0.2185
3	*V*	*V*	37.8	100.0	0.0064

Obsahuje procento proměnlivosti v každém souboru proměnných, vysvětlovaných jiným souborem proměnných. **Index kanonické proměnné:** pořadové číslo (index) kanonické proměnné. Nesmíme zapomenout, že maximální počet proměnných se rovná minimálnímu počtu proměnných v každém souboru. **Proměnlivost v těchto proměnných:** je stejné jako následující. **Objasněno těmito proměnnými:** každý řádek tabulky obsahuje výsledek, jak dokonale je soubor proměnných vysvětlen dotyčnou kanonickou proměnnou. Tento sloupec označuje, který soubor proměnných je právě komentován. **Procento objasnění jednotlivě:** tento sloupec ukazuje procento změny v označeném souboru proměnných, které je vysvětleno touto kanonickou proměnnou. **Procento objasnění kumulativně:** tento sloupec ukazuje kumulativní procento změny v označeném souboru proměnných, které je vysvětleno touto kanonickou proměnnou a ostatními výše. **Kanonický koeficient determinace:** čtverec kanonického korelačního koeficientu.

5. Standardizované kanonické parametry pro kanonické proměnné *U*:

	U_1	U_2	U_3
Test4	1.021375	0.104989	0.370860
Test5	-0.005995	0.990267	0.224017
IQ	-0.065358	0.229775	-1.050237

6. Standardizované kanonické parametry pro kanonické proměnné *V*:

	V_1	V_2	V_3
Test1	0.690657	0.592485	0.510311
Test2	0.655584	-0.428196	-0.636097
Test3	-0.008941	0.919574	-0.485199

Koeficienty jsou užity k určení standardních skóre pro kanonické proměnné *V* a *U* . Slouží k interpretaci proměnných v hodnotě váhy, dané u každé proměnné při konstrukci kanonické proměnné. Jsou analogické standardizovaným parametrům β ve vícenásobné lineární regresi.

7. Korelace párů původní proměnné vs. kanonická proměnná:

	U_1	U_2	U_3	V_1	V_2	V_3
Test4	0.998137	0.019146	-0.057927	0.993745	0.008950	-0.004623
Test5	-0.178759	0.958777	0.220890	-0.177972	0.448190	0.017629
IQ	0.314333	0.211270	-0.925505	0.312950	0.098760	-0.073865
Test1	0.755221	0.144834	0.045750	0.758559	0.309832	0.573230
Test2	0.720964	-0.147861	-0.048910	0.724151	-0.316308	-0.612826
Test3	-0.150877	0.346177	-0.052251	-0.151544	0.740547	-0.654694

Ukazuje korelace párů mezi původní proměnnou a kanonickou proměnnou. Určení, které proměnné jsou vysoce korelované s odpovídající kanonickou proměnnou, napomůže snadnější interpretaci kanonických proměnných. Např. U_1 je vysoce korelovaná s *TEST4*. Proto předpokládáme, že U_1 má stejnou interpretaci jako *TEST4*.

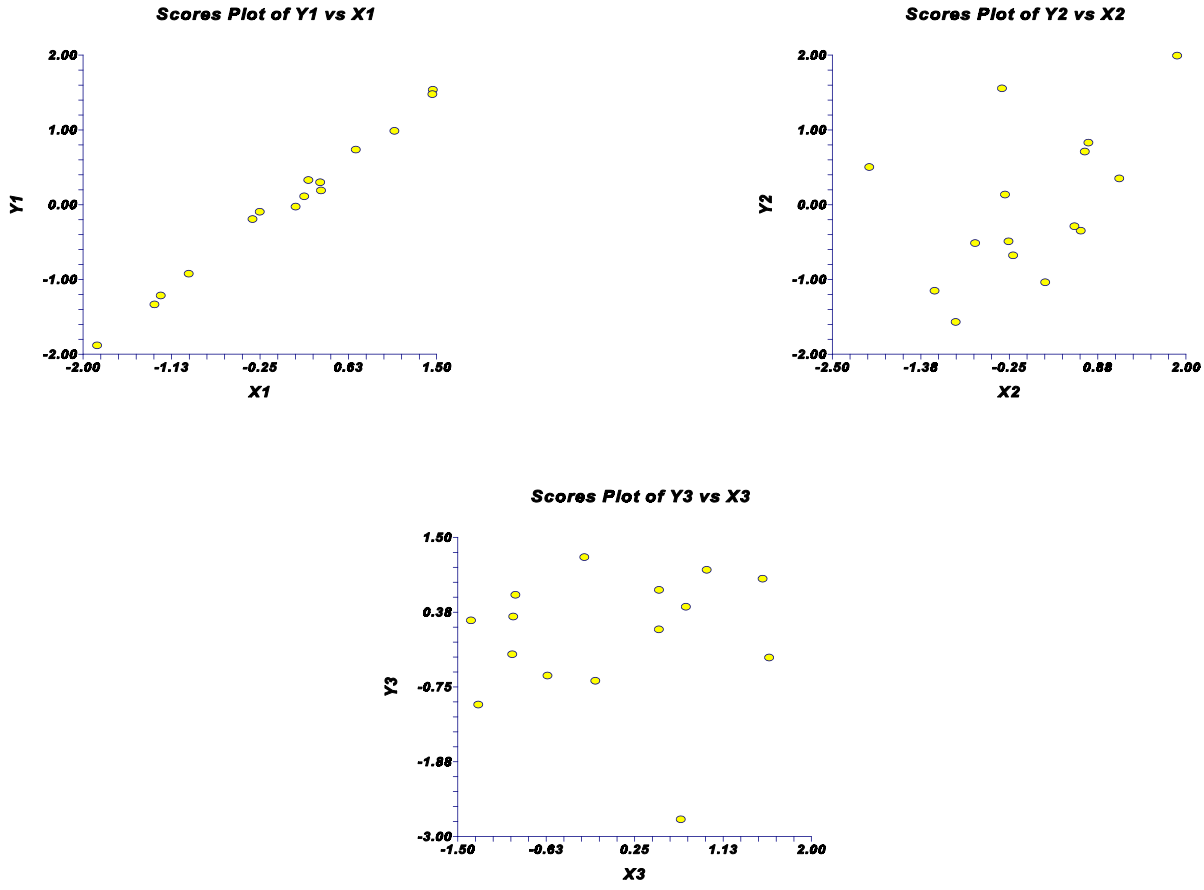
6. Tabulka kanonického skóre pro všechny objekty:

Řádek	U_1	U_2	U_3	V_1	V_2	V_3
1	-0.193124	-0.348044	-0.308495	-0.323303	0.660431	1.582089
2	-1.214743	0.350598	0.877022	-1.232224	1.150186	1.517131
3	-0.026336	0.135325	0.250782	0.103271	-0.304012	-1.369888
4	1.536744	1.992049	-0.657871	1.461462	1.887123	-0.138798
5	0.189923	0.709643	0.455333	0.354314	0.711949	0.757851
6	0.986597	-0.677646	0.115011	1.081350	-0.201044	0.489839
7	0.299464	-0.490602	0.708912	0.345665	-0.258540	0.491428
8	-0.922687	0.503305	1.011073	-0.954587	-2.031644	0.963769

9	-1.881691	-0.288458	0.308479	-1.862181	0.579830	-0.951854
10	-1.333760	0.829021	-1.015632	-1.294283	0.756978	-1.297593
11	0.111861	-1.151067	-2.741954	0.188193	-1.199877	0.707092
12	0.329061	1.555086	-0.579356	0.228934	-0.342184	-0.612825
13	0.736439	-1.037650	0.634374	0.698925	0.206974	-0.929772
14	1.477329	-0.513679	1.201759	1.456751	-0.684236	-0.247278
15	-0.095076	-1.567882	-0.259437	-0.252288	-0.931936	-0.961191

Obsahuje kanonické skóre každého souboru proměnných pro každý řádek úplných dat. Jde o hodnoty, které lze rovněž vynést do grafu.

7. Grafy kanonického skóre pro všechny objekty: grafy ukazují na vztah mezi každým párem kanonických proměnných. Korelační koeficient r_1 dat v prvním grafu (U_1 versus V_1 v grafech označený jako *Y1* versus *X1*) je první kanonický korelační koeficient.



Obr. 4.14 Grafy tří párů kanonických skóre pro všechny objekty.