Kdy kanonická korelace a kdy vícerozměrná lineární regrese?

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice,
532 10 Pardubice, email: milan.meloun@upce.cz
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec,
461 17 Liberec, email> jiri.militky@vslib.cz

Souhrn: *Kanonická korelační analýza je vícerozměrná metoda, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných. První ze dvou skupin se považuje za soubor závisle proměnných* ***y*** *a druhá za soubor nezávisle proměnných* ***x****. Toto rozdělení je ale čistě účelové z důvodu výkladu a nemá žádný vliv na řešení problému. Jde v podstatě o rozšíření metody vícenásobné lineární regrese a korelační analýzy. Zatímco ve vícenásobné lineární regresi hledáme nejlepší kombinaci m nezávisle proměnných* $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}$ *k výpočtu jediné závisle proměnné y, v kanonické korelační analýze hledáme lineární vztah* $\boldsymbol{U_1} = a_1 \boldsymbol{y_1} + a_2 \boldsymbol{y_2} + \ldots + a_p \boldsymbol{y_p}$ *mezi skupinou p čili více jak jediné závisle proměnných* $\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_p}$ *a dále linearní vztah* $\boldsymbol{V_1} = b_1 \boldsymbol{x_1} + b_2 \boldsymbol{x_2} + \ldots + b_p \boldsymbol{x_p}$ *mezi skupinou m nezávisle proměnných* $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}$*. Podstata metody spočívá v tom, že se v každé skupině proměnných vyhledávají koeficienty a a b tak, aby pro všech n objektů vyčíslené kanonické proměnné* U_{1i} *a* V_{1i}*, i = 1, ..., n, vykazovaly maximální párový korelační koeficient. Po jejich nalezení se pak hledají další lineární kombinace čili kanonické proměnné* $\boldsymbol{U_2}$ *a* $\boldsymbol{V_2}$*, které mají druhý největší korelační koeficient za podmínky, že* $\boldsymbol{U_2}$ *a* $\boldsymbol{V_2}$ *jsou nekorelované s prvními kanonickými proměnnými* $\boldsymbol{U_1}$ *a* $\boldsymbol{V_1}$*.*

V kanonické korelační analýze jsou kanonické koeficienty *a* a *b* ve vztazích $\boldsymbol{U_1} = a_1 \boldsymbol{y_1} + a_2 \boldsymbol{y_2} + \ldots + a_p \boldsymbol{y_p}$ a $\boldsymbol{V_1} = b_1 \boldsymbol{x_1} + b_2 \boldsymbol{x_2} + \ldots + b_p \boldsymbol{x_p}$ hledány tak, aby maximalizovaly korelaci mezi proměnnými U_1 a V_1. Po nalezení nejlepších odhadů *a* a *b* se U_1 nazývá *první kanonická proměnná* závisle proměnných ***y*** a V_1 *první kanonická proměnná* nezávisle proměnných ***x***. Obě kanonické proměnné mají průměr roven nule. Korelace mezi U_1 a V_1 se nazývá *první kanonická korelace* a čtverec této korelace je nazýván *vlastní číslo.*

První kanonická korelace je tudíž největší možná korelace mezi lineárními kombinacemi závisle proměnných ***y*** a lineárními kombinacemi nezávisle proměnných ***x***. První kanonická korelace představuje analogii vícenásobnému korelačnímu

koeficientu ve vícenásobné lineární regresi mezi *jedinou* závisle proměnnou ***y*** a souborem nezávisle proměnných ***x***. Rozdíl proti vícenásobné lineární regresi je pouze v tom, že u kanonické korelace je *několik* závisle proměnných ***y*** a dále je nutno navíc hledat lineární kombinaci mezi nimi.

Vyčíslená korelace se nazývá *první kanonický korelační koeficient.* Můžeme sestrojit i jiný soubor vážených průměrů 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. Soubor kanonických proměnných ***U*** vznikl z původních proměnných ***y*.** Soubor kanonických proměnných ***V*** vznikl z původních proměnných ***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 ***x***.
2. $\boldsymbol{R}_{yy}$. Jde o korelaci mezi proměnnými ***y***.
3. $\boldsymbol{R}_{xy}$. Jde o korelaci mezi proměnnými ***x*** a ***y***.
4. $\boldsymbol{R}_{yx}$. Jde o korelaci mezi proměnnými ***y*** a ***x***.

Kanonická korelace může být vyjádřena s využitím metody SVD (Singlular Value Decomposition) matice ***C***, kde $\boldsymbol{C} = \boldsymbol{R}_{yy}^{-1}\ \boldsymbol{R}_{yx}\ \boldsymbol{R}_{xx}^{-1}\ \boldsymbol{R}_{xy}$. V SVD rozkladu matice ***C*** vztahem $\boldsymbol{C} = \hat{a}_y^T\ \boldsymbol{\lambda}\ \hat{a}_y$ je diagonální matice $\boldsymbol{\lambda}$ vlastních čísel je vytvořena z vlastních čísel matice ***C***. Pak *j*-té vlastní číslo λ_j matice ***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

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

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 podělí směrodatnou odchylkou) maticí kanonických koeficientů $V = \boldsymbol{Z}_x \boldsymbol{b}$ a $U = \boldsymbol{Z}_y \boldsymbol{a}$, 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:

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

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 U a V:* 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.*

Úloha 1: *Postup kanonické korelační analýzy*

V úloze 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**
Y	*Test4*	65.53333	13.95332	15
Y	*Test5*	69.93333	16.15314	15
Y	*IQ*	104.3333	11.0173	15
X	*Test1*	67.93333	17.39239	15
X	*Test2*	61.4	19.39735	15
X	*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 prom.	**Kanonická korelace**	***D***	***F*-test**	**Čit. SV**	**Jmen. SV**	**Spočtená α**	**Wilks. λ**
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 prom.** 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 *U* na odpovídající *V* 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. **Čit. SV:** počet stupňů volnosti v čitateli. **Jmen. SV:** počet stupňů volnosti ve jmenovateli. **Spočtená α:** 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. **Wilks λ:** hodnota Wilkova lambda pro kanonickou korelaci tohoto řádku představuje vlastně vícerozměrné zobecnění *D*. Wilkovo lambda je interpretováno opačně než *D*: hodnota blízká nule ukazuje na vysokou korelaci a hodnota blízká 1 na nízkou korelaci.

4. Objasněná proměnlivost v datech:

Index kanonické proměnné	Proměnl. v těchto proměn.	Objasněno těmito proměn.	% objasnění jednotlivě	% objasnění kumulativně	Kanonický koeficient determinace
1	*Y*	*Y*	37.6	37.6	0.9912
2	*Y*	*Y*	32.1	69.7	0.2185
3	*Y*	*Y*	30.3	100.0	0.0064
1	*Y*	*X*	37.2	37.2	0.9912
2	*Y*	*X*	7.0	44.3	0.2185
3	*Y*	*X*	0.2	44.5	0.0064
1	*X*	*Y*	37.1	37.1	0.9912
2	*X*	*Y*	5.4	42.5	0.2185
3	*X*	*Y*	0.2	42.8	0.0064
1	*X*	*X*	37.4	37.4	0.9912
2	*X*	*X*	24.8	62.2	0.2185
3	*X*	*X*	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. **% objasnění jednotlivě:** tento sloupec ukazuje procento změny v označeném souboru proměnných, které je vysvětleno touto kanonickou proměnnou. **% 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*:

	$a_{1,i}$	$a_{2,i}$	$a_{3,i}$
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 *X*:

	$b_{1,i}$	$b_{2,i}$	$b_{3,i}$
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 parame-trům 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

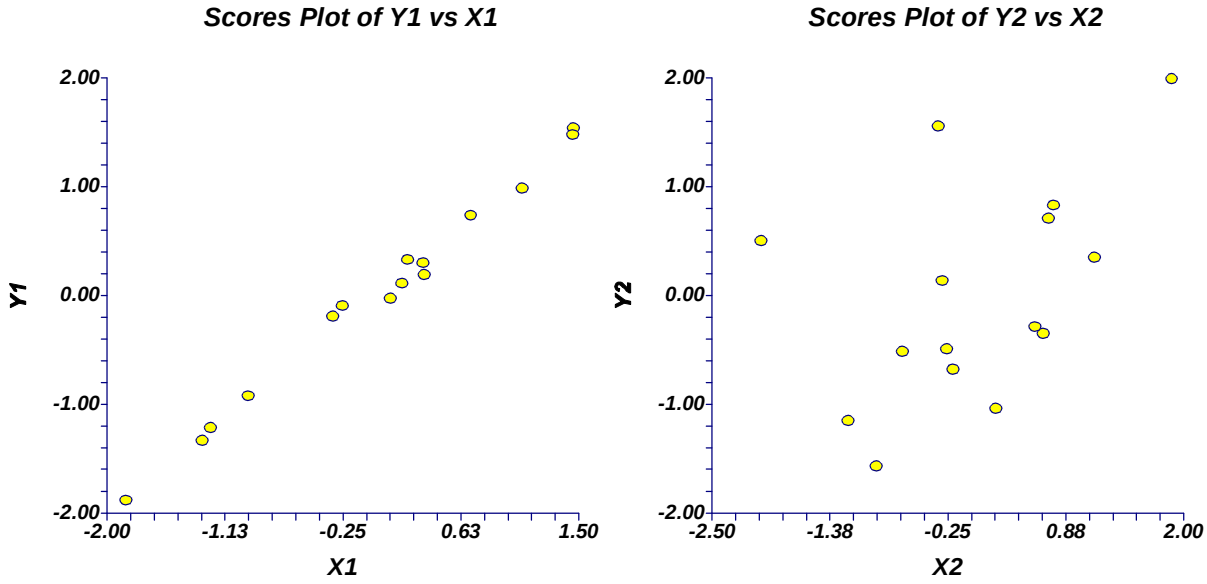
Ukazuje korelace párů mezi původní proměnnou a kanonickou proměnnou. Určení, které proměnné jsou vysoce korelované s odpovídající kanonickou proměnnou, napomůže snadnější interpretaci kanonických proměnných. Např. U_1 je vysoce korelovaná s *TEST4.* Proto předpokládáme, že U_1 má stejnou interpretaci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

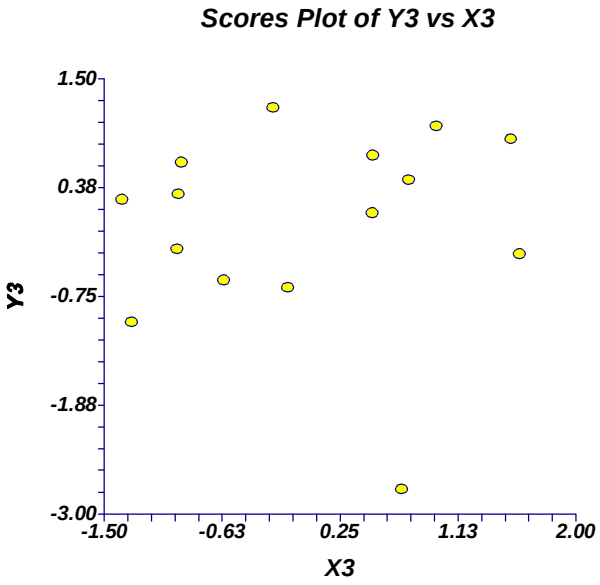
6. Tabulka kanonického skóre pro všechny objekty:

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.





Úloha 2. *Korelace vlastností u italských vín*

Pro 90 vzorků italských vín bylo naměřeno 8 fyzikálně-chemických vlastností. *Data: i* index vzorku vína, *j* jméno vzorku vína, *k* kategorie vzorku vína. Data obsahují 90 druhů vín (objekty) tří kategorií 1. Barolo, 2. Grignolino a 3. Barbera, popsaných 8 následujícími vlastnostmi (proměnné): x_1 obsah alkoholu, x_2 necukerný extrakt, x_3 fosfáty, x_4 celkové fenoly, x_5 flavanoidy, x_6 poměr absorbancí při 280 a 315 nm pro naředěné víno, x_7 poměr absorbancí při 280 a 315 nm pro flavanoidy, x_8 obsah prolinu.

i	j	*k*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	Olo0171	1	14.23	24.82	320	2.80	3.06	3.92	4.77	1065
..	..	..	..	..	..	..	..	..	..	..
90	Era2878	3	13.17	23.45	534	1.65	0.68	1.62	2.05	840

Canonical Correlation Report

Descriptive Statistics Section			**Standard**	**Non-Missing**
Type	**Variable**	**Mean**	**Deviation**	**Rows**
V	fosfaty	372.1	94.1845	90
V	fenoly	2.274444	0.630951	90
V	flavanoidy	1.989556	1.014599	90
V	pomerA1	2.595333	0.752854	90
V	pomerA2	2.951889	0.9338658	90
V	prolin	754.1778	310.4893	90
U	alkohol	13.1	0.8215004	90
U	extrakt	25.30611	1.96962	90

Correlation Section

	fosfaty	fenoly	flavanoidy	pomerA1	pomerA2
fosfaty	1.000000	0.460003	0.437309	0.196172	0.116061
fenoly	0.460003	1.000000	0.900560	0.727798	0.630202
flavanoidy	0.437309	0.900560	1.000000	0.790604	0.685836
pomerA1	0.196172	0.727798	0.790604	1.000000	0.930309
pomerA2	0.116061	0.630202	0.685836	0.930309	1.000000
prolin	0.518397	0.566570	0.590651	0.337522	0.199983
alkohol	0.341996	0.293228	0.228402	0.014661	-0.129063
extrakt	0.330043	0.463001	0.413509	0.241733	0.152094

	prolin	alkohol	extrakt
fosfaty	0.518397	0.341996	0.330043
fenoly	0.566570	0.293228	0.463001
flavanoidy	0.590651	0.228402	0.413509
pomerA1	0.337522	0.014661	0.241733
pomerA2	0.199983	-0.129063	0.152094
prolin	1.000000	0.599580	0.418690
alkohol	0.599580	1.000000	0.468140
extrakt	0.418690	0.468140	1.000000

Variation Explained Section

Canonical Variate Number	Variation in these Variables	Explained by these Variates	Individual Percent Explained	Cumulative Percent Explained	Canonical Correlation Squared
1	*V*	*V*	26.3	26.3	0.4751
2	*V*	*V*	41.1	67.5	0.1325
1	*V*	*U*	12.5	12.5	0.4751
2	*V*	*U*	5.4	18.0	0.1325
1	*U*	*V*	32.4	32.4	0.47512
2	*U*	*V*	4.2	36.6	0.1325
1	*U*	*U*	68.3	68.3	0.4751
2	*U*	*U*	31.7	100.0	0.1325

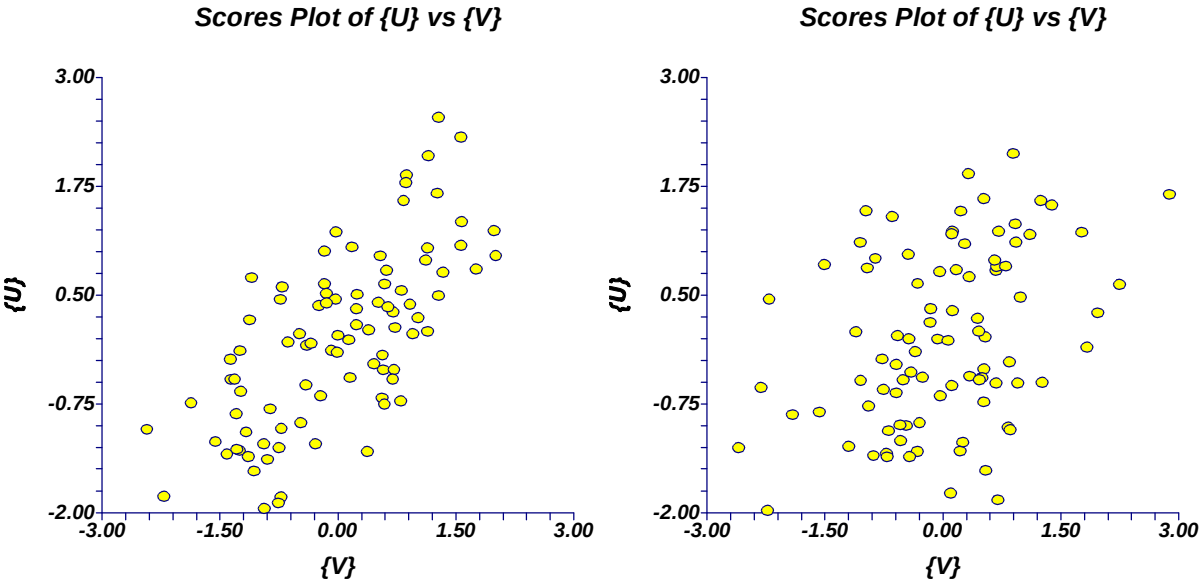
Variable - Variate Correlations Section

	V1	*V2*	*U1*	*U2*
fosfaty	0.542372	-0.327108	0.373862	-0.119054
fenoly	0.524897	-0.828603	0.361817	-0.301578
flavanoidy	0.426436	-0.806388	0.293947	-0.293493
pomerA1	0.098935	-0.706984	0.068197	-0.257313
pomerA2	-0.112681	-0.718903	-0.077672	-0.261652
prolin	0.897688	-0.086501	0.618785	-0.031483
alkohol	0.675020	0.073725	0.979269	0.202562
extrakt	0.439387	-0.280434	0.637430	-0.770508

Canonical Correlations Section: $\alpha = 0.05$

Variate Number	Canonical Correlation	R^2	F-Val.	Num DF	Den DF	Prob α	Wilks' λ
1	0.689310	0.475149	**6.59**	12	164	**0.000000**	0.455326
2	0.363960	0.132467	**2.53**	5	83	**0.034739**	0.867533

F-value tests whether this canonical correlation and those following are zero.



Úloha 3. *Korelace spotřeby proteinů v potravinách v Evropě*

Data: i značí index, *j* název země, x_1 červené maso, x_2 bílé maso, x_3 vejce, x_4 mléko, x_5 ryby, x_6 obilniny, x_7 škrob, x_8 ořechy, x_9 ovoce a zelenina,

i	*j*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
..	...	...	...	...	...	...	...	...	...	...
25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Canonical Correlation Report

Descriptive Statistics Section Type	Variable	Mean	Standard Deviation	Non-Missing Rows
V	x_5 = ryby	4.284	3.402533	25
V	x_6 = obilniny	32.25	10.97479	25
V	x_7 = škrob	4.276	1.634085	25
V	x_8 = ořechy	3.072	1.985682	25
V	x_9 = ovoce, zelenina	4.136	1.803903	25
U	x_1 = červené maso	9.828	3.347078	25
U	x_2 = bílé maso	7.896	3.694081	25
U	x_3 = vejce	2.936	1.117617	25
U	x_4 = mléko	17.112	7.105416	25

Variation Explained Section

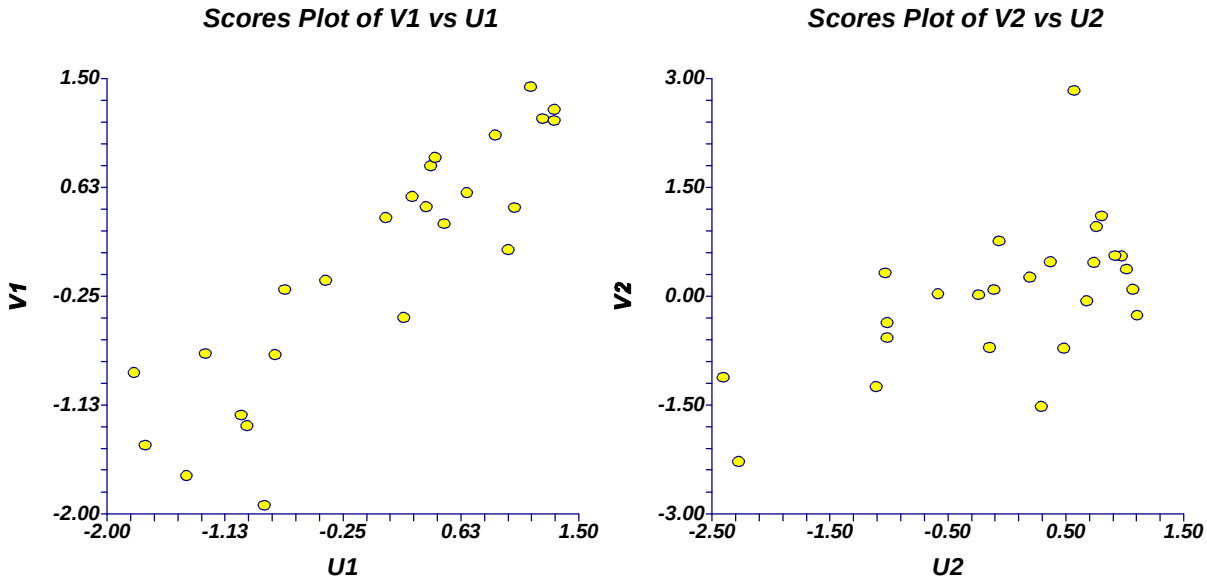
Canonical Variate Number	Variation in these Variables	Explained by these Variates	Individual Percent Explained	Cumulative Percent Explained	Canonical Correlation Squared
1	*V*	*V*	29.4	29.4	0.7972
2	*V*	*V*	16.5	45.9	0.3464
3	*V*	*V*	17.8	63.7	0.3265
4	*V*	*V*	14.7	78.4	0.0557
1	*U*	*U*	55.3	55.3	0.7972
2	*U*	*U*	10.4	65.6	0.3464
3	*U*	*U*	19.1	84.8	0.3265
4	*U*	*U*	15.2	100.0	0.0557

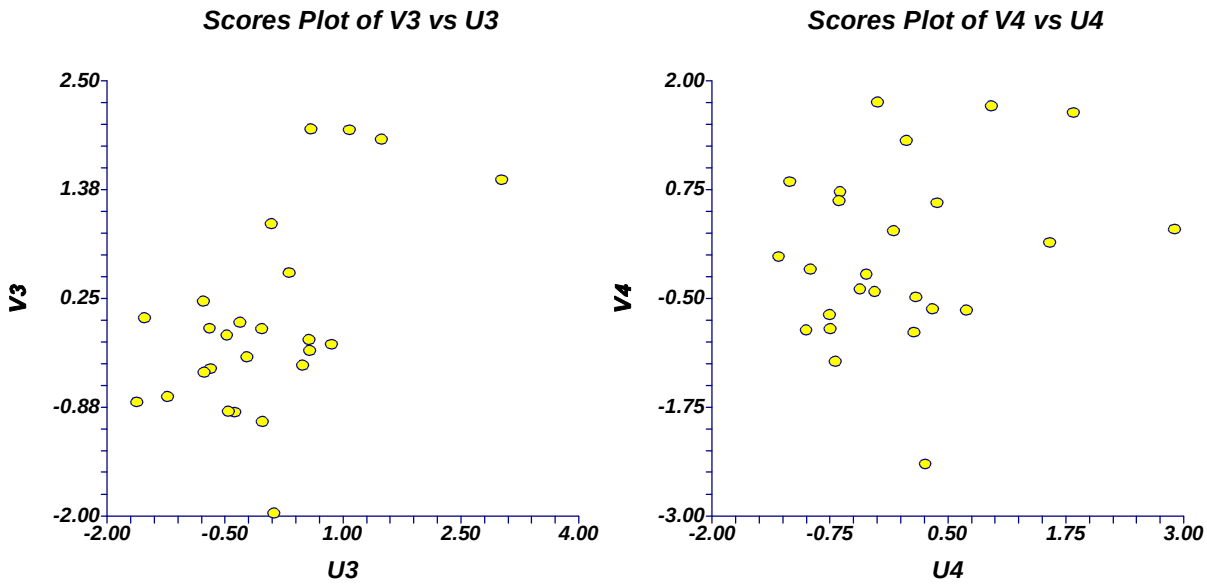
Canonical Correlations Section: $\alpha = 0.05$

Variate Number	Canonical Correlation	R^2	*F*-Value	Num DF	Den DF	Prob α	Wilks' λ
1	0.892837	0.797157	**2.99**	20	54	**0.000714**	0.084311
2	0.588592	0.346440	**1.48**	12	45	**0.165436**	0.415647
3	0.571440	0.326544	**1.52**	6	36	**0.198432**	0.635975
4	0.235915	0.055656	**0.56**	2	19	**0.580415**	0.944344

F-value tests whether this canonical correlation and those following are zero.

Plots Section





Úloha 4. *Korelace spotřeby a ceny s ostatními technickými vlastnostmi aut*
Databáze 155 aut byla tříděna dle 11 důležitých vlastností.
Data: i index, x_1 spotřeba benzinu v počtu ujetých mil na 1 gallon (mpg), x_2 počet válců (cylinders), x_3 vrtání (displace), x_4 výkon v koňských silách (horsepower), x_5 zrychlení (accel), x_6 poslední dvojčíslí roku výroby (year), x_7 hmotnost vozu (weight), x_8 země původu (origin: 1 USA, 2 Evropa, 3 Japonsko), x_9 výrobce (make), x_{10} model, x_{11} cena vozu v US$ v roce 1978 (price).

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	43.1	4	90	48	21.5	78	1985	2	Volkswagen	Rabbit Dl	2400
..	..	..	..	..	..	..	..	..	..	..	..
155	31.0	4	119	82	19.4	82	2720	1	Chevrolet	S-10	4500

Canonical Correlation Report

Descriptive Statistics Section			**Standard**	**Non-Missing**
Type	**Variable**	**Mean**	**Deviation**	**Rows**
V	cylinders	4.84106	1.376434	151
V	displace	154.1722	73.15579	151
V	horsepower	89.42384	24.20095	151
V	accel	16.1894	2.405415	151
V	weight	2677.139	606.1441	151
U	mpg	28.5351	7.216594	151
U	price	4618.212	2029.606	151

Correlation Section

	cylinders	displace	horsepower	accel	weight
cylinders	1.000000	0.933000	0.761545	-0.194820	0.821521
displace	0.933000	1.000000	0.801880	-0.175235	0.910983
horsep.	0.761545	0.801880	1.000000	-0.452144	0.813215
accel	-0.194820	-0.175235	-0.452144	1.000000	-0.026554
weight	0.821521	0.910983	0.813215	-0.026554	1.000000
mpg	-0.679312	-0.749998	-0.774144	0.161410	-0.842125
price	0.033617	0.074110	0.079361	0.085936	0.209855

	mpg	price
cylinders	-0.679312	0.033617
displace	-0.749998	0.074110
horsepower	-0.774144	0.079361
accel	0.161410	0.085936
weight	-0.842125	0.209855
mpg	1.000000	-0.010216
price	-0.010216	1.000000

Variation Explained Section

Canonical Variate Number	Variation in these Variables	Explained by these Variates	Individual Percent Explained	Cumulative Percent Explained	Canonical Correlation Squared
1	V	V	59.4	59.4	0.7898
2	V	V	15.3	74.6	0.0947
1	V	U	46.9	46.9	0.7898
2	V	U	1.4	48.3	0.0947
1	U	V	39.7	39.7	0.7898
2	U	V	4.7	44.4	0.0947
1	U	U	50.3	50.3	0.7898
2	U	U	49.7	100.0	0.0947

Variable - Variate Correlations Section

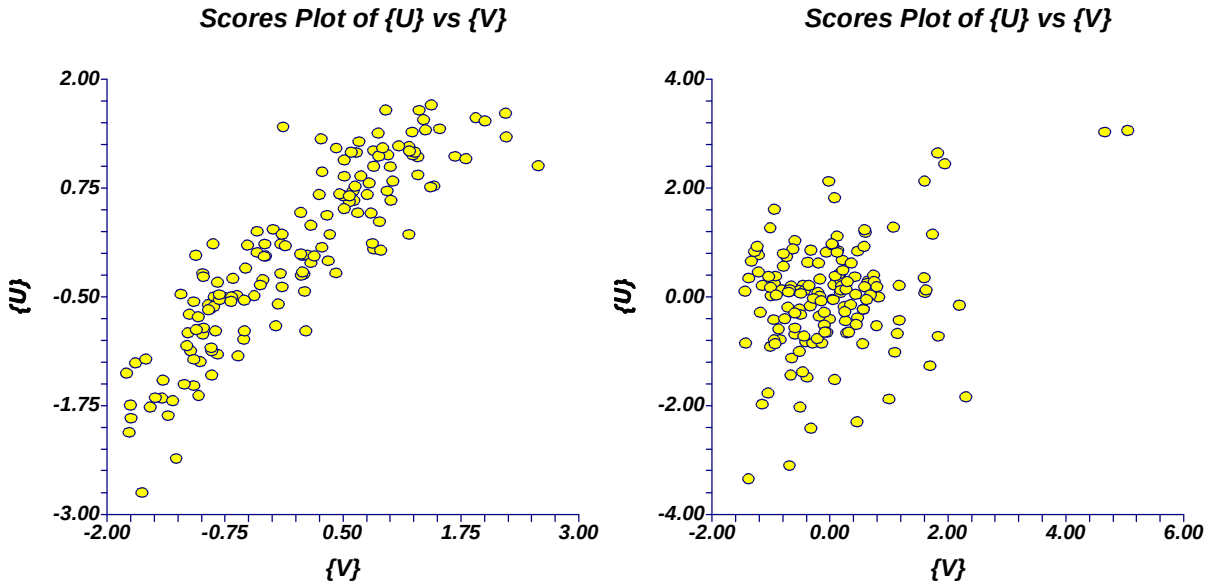
	V1	V2	U1	U2
cylinders	-0.746314	-0.484916	-0.663257	-0.149220
displace	-0.834696	-0.419205	-0.741803	-0.129000
horsep.	-0.862399	-0.423706	-0.766423	-0.130385
accel	0.150103	0.410155	0.133398	0.126215
weight	-0.973938	-0.073027	-0.865549	-0.022472
mpg	0.858715	0.079273	0.966248	0.257612
price	-0.237703	0.296513	-0.267470	0.963566

Canonical Correlations Section: $\alpha = 0.05$

Variate Number	Canonical Correlation	R^2	F-Value	Num DF	Den DF	Prob Level	Wilks' λ
1	0.888711	0.789806	**37.22**	10	288	**0.000000**	0.190289
2	0.307724	0.094694	**3.79**	4	145	**0.005799**	0.905306

F-value tests whether this canonical correlation and those following are zero.

Plots Section



Úloha 5: *Korelace výšek a stáří muže a ženy u 169 dvojic*

i	x_1	x_2	x_3	x_4	x_5
1	49	1809	43	1590	25
...	...	...	...	...	...
169	59	1720	56	1530	24

Canonical Correlation Report

Descriptive Statistics Section Type	Variable	Mean	Standard Deviation	Non-Missing Rows
U	x_1 = věk manžela [roky]	42.85207	11.76033	169
U	x_2 = výška manžela [mm]	1732.959	66.53914	169
U	x_5 = věk manžela 1. svatby	25.21894	5.396703	169
V	x_3 = věk manželky [roky]	40.60947	11.4086	169
V	x_4 = výška manželky [mm]	1603.136	62.20348	169

Correlation Section

	x_1	x_2	x_5	x_3	x_4
x_1	1.000000	-0.225150	0.362437	0.938238	-0.193101
x_2	-0.225150	1.000000	-0.103211	-0.156320	0.304125
x_5	0.362437	-0.103211	1.000000	0.236422	-0.036811
x_3	0.938238	-0.156320	0.236422	1.000000	-0.206279
x_4	-0.193101	0.304125	-0.036811	-0.206279	1.000000

Variable - Variate Correlations Section

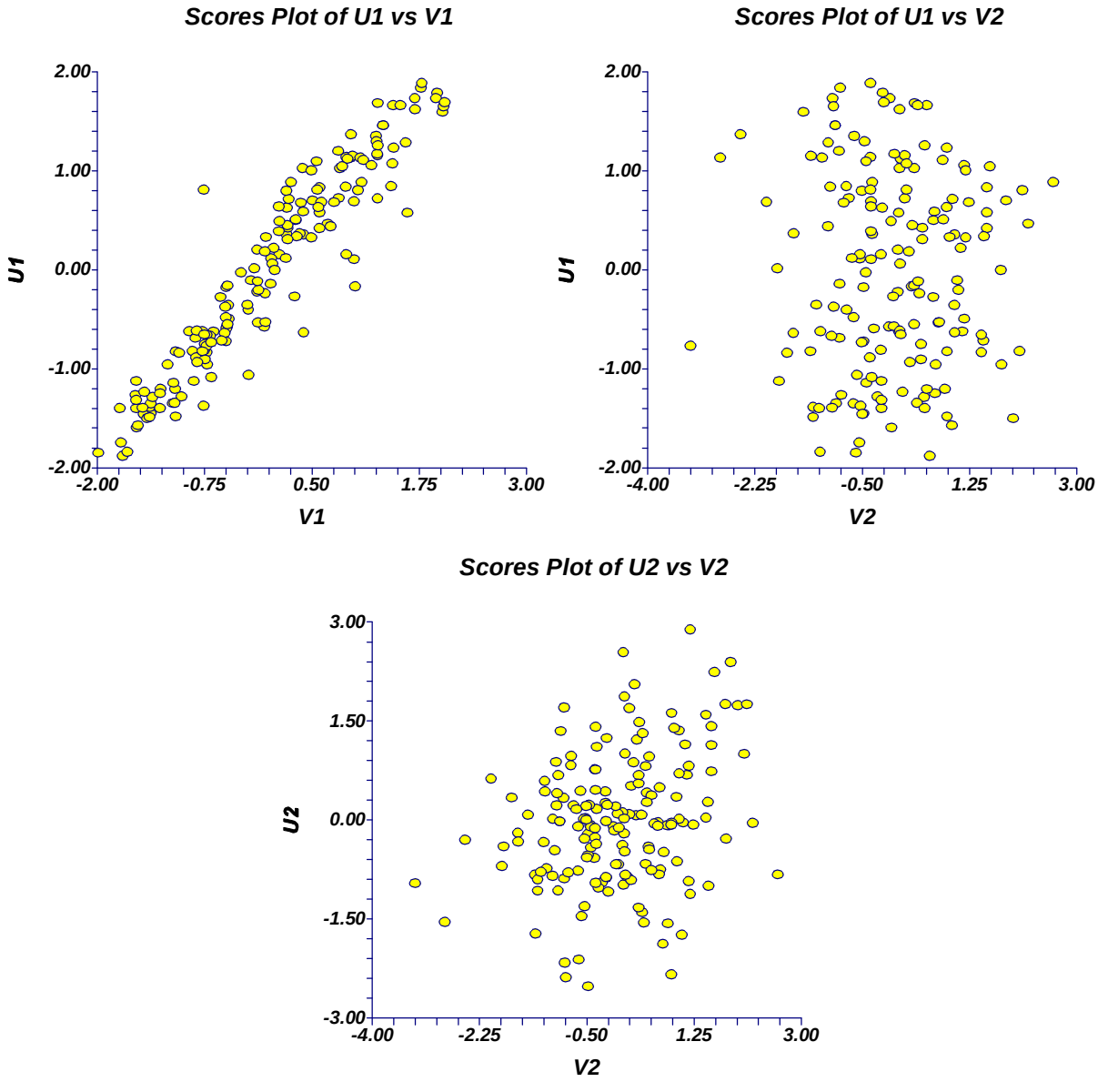
	U_1	U_2	V_1	V_2
x_1	0.991163	-0.056286	0.938101	-0.016071
x_2	-0.159967	0.982639	-0.151403	0.280564
x_5	0.249984	0.028215	0.236601	0.008056
x_3	0.946318	-0.005027	0.999845	-0.017606
x_4	-0.178900	0.280374	-0.189020	0.981973

Canonical Correlations Section: α = 0.05

Variate Number	Canonical Correlation	R^2	F-Value	Num DF	Den DF	Prob Level	Wilks' λ
1	0.946464	0.895795	122.0	63	28	0.000000	0.095710
2	0.285521	0.081522	7.32	2	165	0.000898	0.918478

F-value tests whether this canonical correlation and those following are zero.

Plots Section



Poděkování: Předložený prpojekt byl vypracován za finační podpory Vědeckého záměru MŠMT č. MSM253100002 a grantu Grantové agentury České republiky 303/00/1559.

Doporučená literatura:

[1] Siotani M., Hayakawa T., Fujikoshi Y.: Modern Multivariate Statistical Analysis, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Kendall M. G., Stuart A.: The Advanced Theory of Statistics, Vol. III. New York 1966.
[3] James W., Stein C.: Estimation with Quadratic Loss, Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.
[4] Guanadeskian R., Kettenring J. R.: Biometrics **28**, 80 (1972).
[5] Campbell N. A.: Appl. Statist., 29, 231 (1980).
[6] Hu J., Skrabal P., Zollinger H.: Dyes and Pigments, **8**, 189 (1987).
[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: Graphical Methods for Data Analysis. Duxburg Press, Belmont, California 1983.
[8] Barnett V., (Edit.): Interpreting Multivariate Data. Wiley, Chichester 1981, kap. 6.
[9] Jolliffe I. T.: Principal Component Analysis. Springer Verlag, New York 1986.
[10] Barnett V.,B. S.: Graphical Techniques for Multivariate Data. London 1978.
[12] Andrews D. F.: Biometrics, **28**, 125 (1972).
[13] Kulkarni S. R., Paranjape S. R.: Commun. Statist., **13**, 2511 (1984).
[14] Guanadeskian R.: Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York 1977.
[15] Kleiner B., Hartigan J. A., J. Amer. Statist. Assoc., **76**, 260 (1981).
[16] Kres H.: Statistical Tables for Multivariate Analysis. Springer, New York 1983.
[17] Seber G. A. F.: Multivariate Observations. Wiley, New York 1984.
[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: Z. Anal. Chem., **327**, 679 (1987).
[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: Technometrics, **24**, 139 (1982).
[20] Johnson R.A., Wichern D.W.: Applied Multivariate Statistical Analysis, Prentice Hall, 1998.
[21] Ajvjazin S., Bežajeva Z., Staroverov O.: Metody vícerozměrné analýzy, SNTL Praha 1981
[22] Meloun M., Militký J. , Forina M.: Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester 1992.
[23] Brereton R. G. Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies, Elsevier 1992,
[24] Krzanowski W. J.: Principles of Multivariate Analysis, A User's Perspective, Oxford Science Publications 1988,
[25] Jeffers J. N. R., Applied Statistician, **16**, 225 (1967).
[26] Meloun M. , Militký J., Statistické zpracování experimentálních dat, Plus Praha 1994.
[27] Martens H., Naes T., Multivariate calibration, Wiley (1989) Chichester.
[28] Thomas E. V., Anal. Chem., **66** (1994) 795A-804A.
[29] Malinowski F., Howery D., Factor Analysis in Chemistry, Wiley (1980) New York.
[30] Everitt B. S., Dunn G., Applied Multivariate Data Analysis, Arnold, London 2001.