Cvičení 4: Kanonická diskriminační analýza

Příklad na třídění do dvou skupin: Třídění lebek Tibeťanů (Příklad je převzat z knihy Meloun M., Militký J., Hill, M.: Počítačová analýza vícerozměrných dat v příkladech. Academia Praha 2005)
Datový soubor lebky.sta obsahuje údaje o 32 lebkách nalezených na pohřebištích v Tibetu. Sledují se tyto proměnné:
ID … identifikátor (1 pro lebky z okolí Sikkimu, 2 pro lebky z okolí Lhasy)
Ldelka … největší délka lebky (v mm)
Lsirka … největší horizontální šířka lebky (v mm)
Lvyska … výška lebky (v mm)
Ovyska … výška horní části obličeje (v mm)
Osirka … šířka obličeje mezi body lícních kostí (v mm)

	1 ID	2 Ldelka	3 Lsirka	4 Lvyska	5 Ovyska	6 Osirka
1	1	190,5	152,5	145,0	73,5	136,5
2	1	172,5	132,0	125,5	63,0	121,0
3	1	167,0	130,0	125,5	69,5	119,5
4	1	169,5	150,5	133,5	64,5	128,0
5	1	175,0	138,5	126,0	77,5	135,5
6	1	177,5	142,5	142,5	71,5	131,0
7	1	179,5	142,5	127,5	70,5	134,5
8	1	179,5	138,0	133,5	73,5	132,5
9	1	173,5	135,5	130,5	70,0	133,5
10	1	162,5	139,0	131,0	62,0	126,0
11	1	178,5	135,0	136,0	71,0	124,0
12	1	171,5	148,5	132,5	65,0	146,5
13	1	180,5	139,0	132,0	74,5	134,5
14	2	183,0	149,0	121,5	76,5	142,0
15	2	169,5	130,0	131,0	68,0	119,0
16	2	172,0	140,0	136,0	70,5	133,5
17	2	170,0	126,5	134,5	66,0	118,5
18	2	182,5	136,0	138,5	76,0	134,0
19	2	179,5	135,0	128,5	74,0	132,0
20	2	191,0	140,5	140,5	72,5	131,5
21	2	184,5	141,5	134,5	76,5	141,5
22	2	181,0	142,0	132,5	79,0	136,5
23	2	173,5	136,5	126,0	71,5	136,5
24	2	188,5	130,0	143,0	79,5	136,0
25	2	175,0	153,0	130,0	76,5	142,0
26	2	196,0	142,5	123,5	76,0	134,0
27	2	200,0	139,5	143,5	82,5	146,0
28	2	185,0	134,5	140,0	81,5	137,0
29	2	174,5	143,5	132,5	74,0	136,5
30	2	195,5	144,0	138,5	78,5	144,0
31	2	197,0	131,5	135,0	80,5	139,0
32	2	182,5	131,0	135,0	68,5	136,0

Úkolem je provést kanonickou diskriminační analýzu a následně pomocí zařazovacího pravidla založeného na průměru kanonických proměnných zařadit lebky do dvou skupin

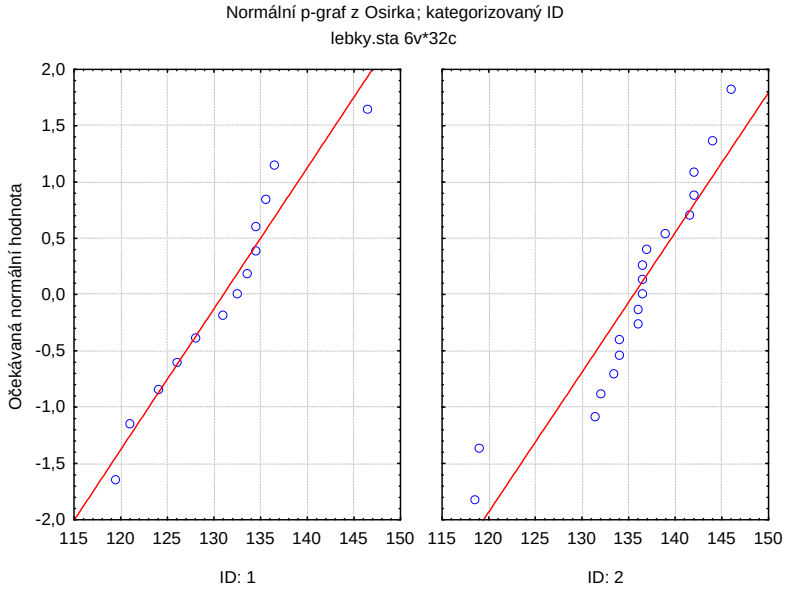
Výsledky (s částečným návodem)

Testování hypotézy o normalitě sledovaných proměnných v daných dvou skupinách pomocí S-W testu:

Proměnná	Souhrnné výsledky Testy normality (lebky.sta) ID	N	W	p
Ldelka	Sikkim	13	0,971258	0,908822
Lsirka	Sikkim	13	0,946284	0,543198
Lvyska	Sikkim	13	0,900168	0,134439
Ovyska	Sikkim	13	0,944919	0,523649
Osirka	Sikkim	13	0,954446	0,666891
Ldelka	Lhasa	19	0,946640	0,345905
Lsirka	Lhasa	19	0,973572	0,844925
Lvyska	Lhasa	19	0,969669	0,769812
Ovyska	Lhasa	19	0,965452	0,683230
Osirka	Lhasa	19	0,873328	0,016463

Vidíme, že ve 2. skupině zamítá S-W test hypotézu o normalitě proměnné Osirka na hladině významnosti 0,05.

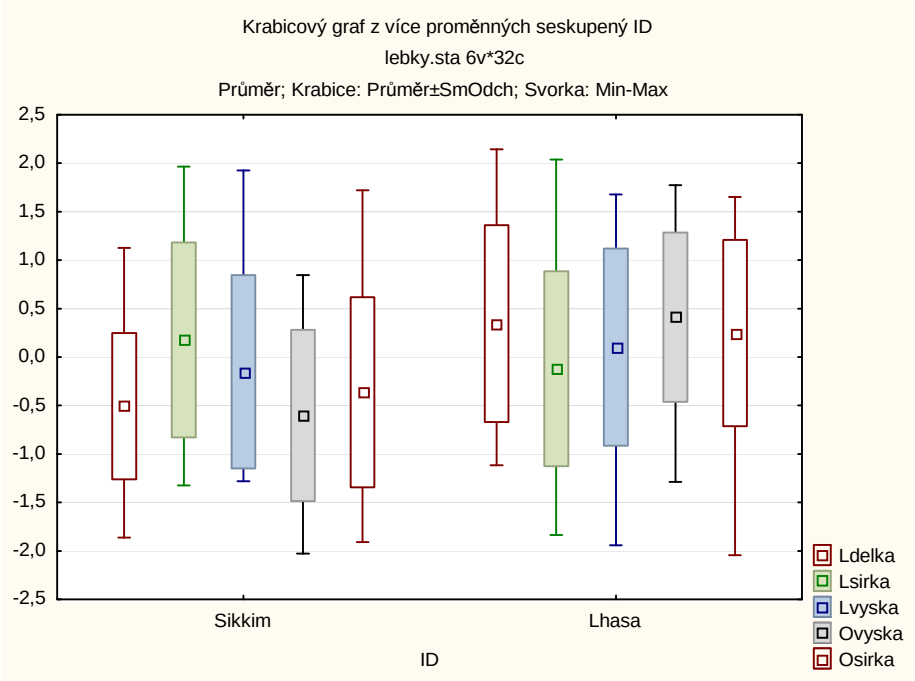
N-P plot pro proměnnou Osirka v 1. a 2. skupině



Odhad vektorů středních hodnot v 1. a 2. skupině:

Proměnná	Souhrnné výsledky Popisné statistiky (lebky.sta) ID	Průměr
Ldelka	Sikkim	175,19
Lsirka	Sikkim	140,27
Lvyska	Sikkim	132,38
Ovyska	Sikkim	69,69
Osirka	Sikkim	131,00
Ldelka	Lhasa	183,18
Lsirka	Lhasa	138,24
Lvyska	Lhasa	133,92
Ovyska	Lhasa	75,16
Osirka	Lhasa	135,55

Krabicové grafy všech (standardizovaných) proměnných v 1. a 2. skupině:



Odhad varianční matice v 1. skupině:

	Kovariance (lebky.sta) Zhrnout podmínku: ID=1				
Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	50,02244	17,52724	25,02404	22,85577	20,04167
Lsirka	17,52724	47,31731	25,57532	-0,76442	32,35417
Lvyska	25,02404	25,57532	36,83974	5,89904	12,27083
Ovyska	22,85577	-0,76442	5,89904	22,64744	10,29167
Osirka	20,04167	32,35417	12,27083	10,29167	53,25000

Odhad varianční matice ve 2. skupině:

	Kovariance (lebky.sta) Zhrnout podmínku: ID=2				
Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	90,31140	7,3012	20,1126	31,64985	39,07310
Lsirka	7,30117	47,2880	-14,6747	10,68275	30,51462
Lvyska	20,11257	-14,6747	38,1462	8,66594	4,42105
Ovyska	31,64985	10,6827	8,6659	22,14035	25,13012
Osirka	39,07310	30,5146	4,4211	25,13012	51,05263

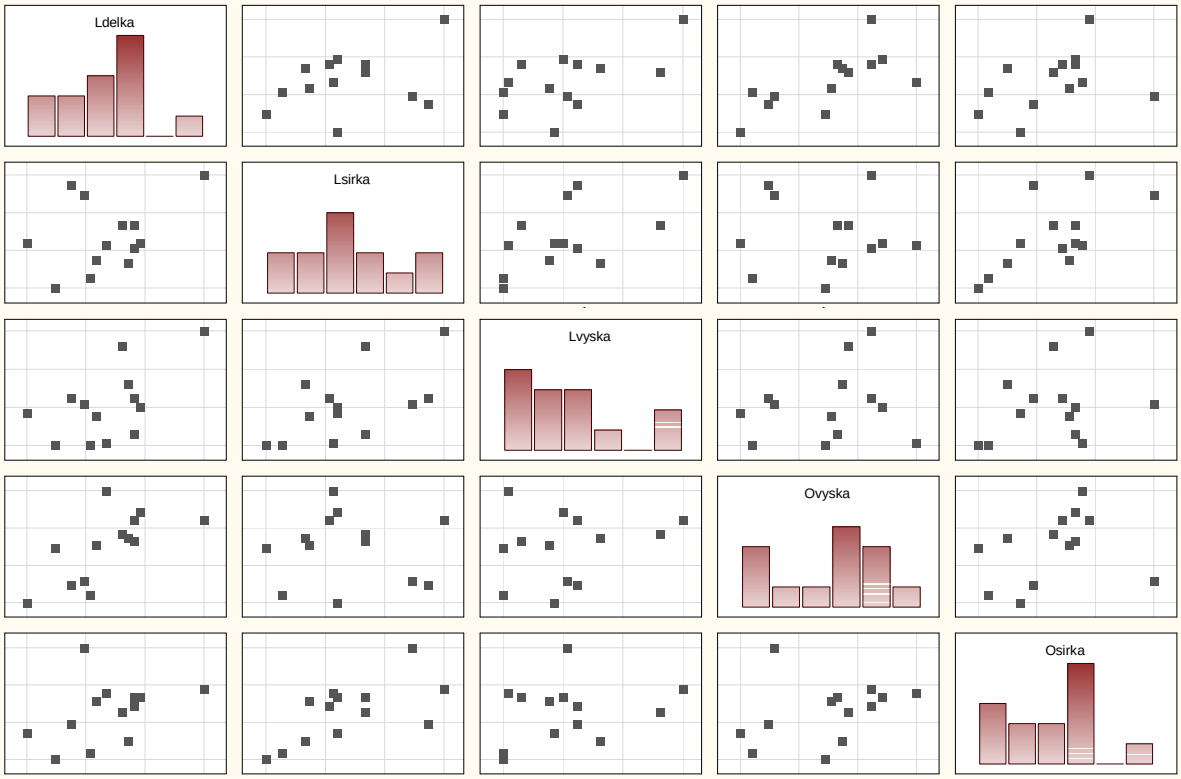
Boxův test shody variančních matic:

	Boxův M test (lebky.sta) Efekt: ID (Vypočteno pro všechny proměnné)			
	Boxovo M	Chí-kv.	sv	p
Boxovo M	22,65281	18,40191	15	0,242126

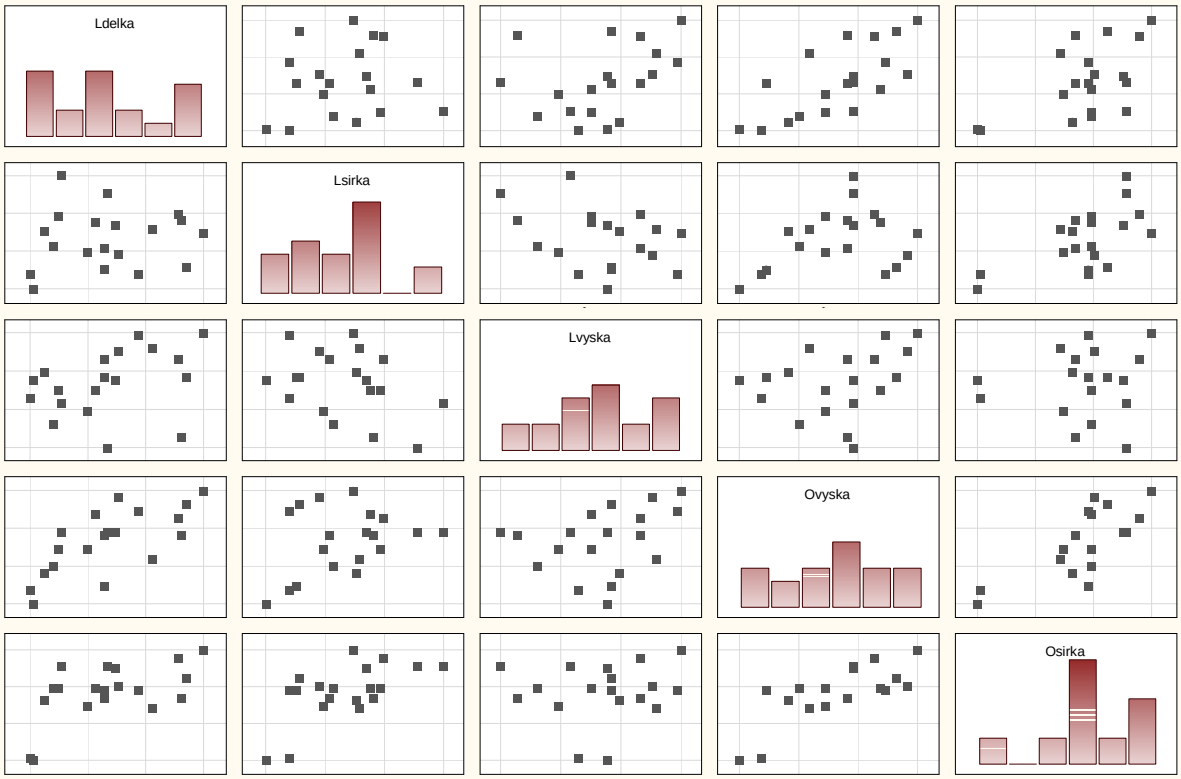
Hypotézu o shodě variančních matic nezamítáme na asymptotické hladině významnosti 0,05, protože p-hodnota = 0,242 je větší než 0,05.

Prověření linearity vztahů sledovaných proměnných v daných dvou skupinách

Maticový graf
lebky.sta 6v*32c
Zahrnout jestliže: ID=1



Maticový graf
lebky.sta 6v*32c
Zahrnout jestliže: ID=2



Odhad korelační matice $\mathbf{R}_1$

Proměnná	Korelace (lebky.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=13 (Celé případy vynechány u ChD) Zhrnout podmínku: ID=1				
	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	1,000000	0,360264	0,582930	0,679053	0,388322
Lsirka	0,360264	1,000000	0,612566	-0,023351	0,644556
Lvyska	0,582930	0,612566	1,000000	0,204227	0,277049
Ovyska	0,679053	-0,023351	0,204227	1,000000	0,296358
Osirka	0,388322	0,644556	0,277049	0,296358	1,000000

Odhad korelační matice $\mathbf{R}_2$

Proměnná	Korelace (lebky.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=19 (Celé případy vynechány u ChD) Zhrnout podmínku: ID=2				
	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	1,000000	0,111724	0,342666	0,707796	0,575437
Lsirka	0,111724	1,000000	-0,345516	0,330153	0,621045
Lvyska	0,342666	-0,345516	1,000000	0,298193	0,100182
Ovyska	0,707796	0,330153	0,298193	1,000000	0,747469
Osirka	0,575437	0,621045	0,100182	0,747469	1,000000

Test shody vektorů středních hodnot:

Proměnná	t-testy; grupováno: ID (lebky.sta) Skup. 1: 1; Skup. 2: 2 T2(celé případy 14,0638 F(5,26)=2,4377 p<,06127										
	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr Rozptyly	p Rozptyly
Ldelka	175,1923	183,1842	-2,57771	30	0,015102	13	19	7,072654	9,503231	1,805418	0,298973
Lsirka	140,2692	138,2368	0,82101	30	0,418115	13	19	6,878758	6,876628	1,000620	0,970788
Lvyska	132,3846	133,9211	-0,69592	30	0,491837	13	19	6,069575	6,176261	1,035463	0,976491
Ovyska	69,6923	75,1579	-3,21246	30	0,003136	13	19	4,758932	4,705353	1,022903	0,938035
Osirka	131,0000	135,5526	-1,75517	30	0,089438	13	19	7,297260	7,145112	1,043041	0,909092

Testová statistika se realizuje hodnotou 2,4377, odpovídající p.hodnota je menší než 0,06127, tedy na hladině významnosti 0,056 nezamítáme hypotézu o shodě vektorů středních hodnot.

Výpočet vlastních čísel matice $\mathbf{BE}^{-1}$

Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (lebky.sta)					
	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	0,468795	0,564951	0,680830	10,57216	5	0,060555

Výpočet standardizovaných a prostých koeficientů 1. kanonické proměnné

Proměnná	Prosté koeficienty (lebky.sta) pro kanonické proměnné
	Kořen1
Ldelka	0,02479
Lsirka	-0,09494
Lvyska	-0,01911
Ovyska	0,12902
Osirka	0,06216
Konstant	-6,43096
Vlastní	0,46879
KumPodíl	1,00000

Proměnná	Standardiz. koeficienty (lebky.sta) pro kanonické proměnné
	Kořen1
Ldelka	0,213491
Lsirka	-0,652968
Lvyska	-0,117197
Ovyska	0,609848
Osirka	0,447939
Vlastní	0,468795
KumPodíl	1,000000

Y_1 = 0,02479*Ldelka – 0,09494*Lsirka – 0,01911*Lvyska + 0,12902*Ovyska + 0,06216*Osirka - 6,43096

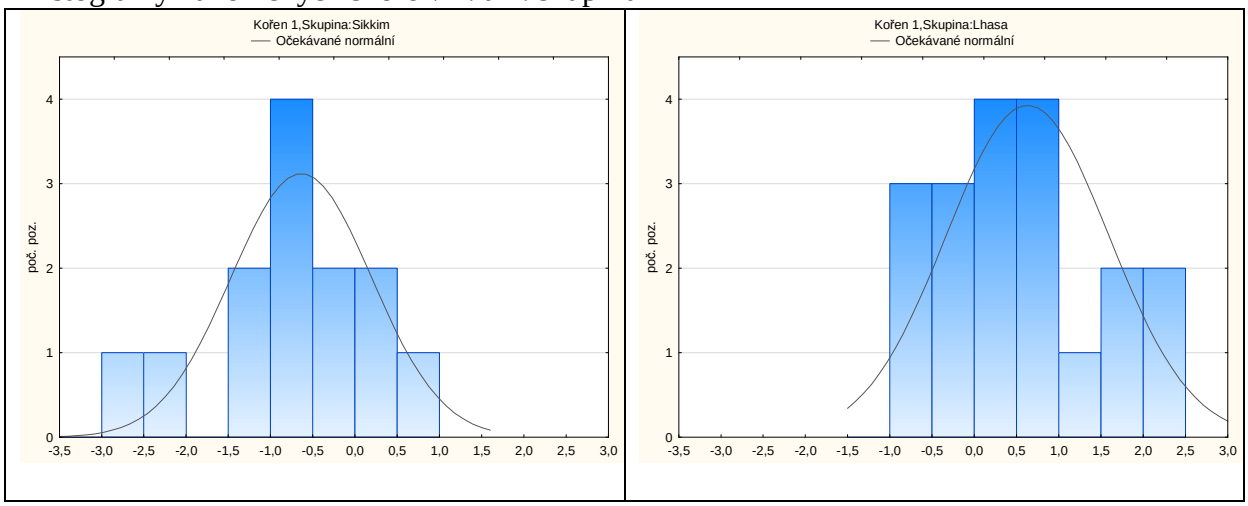
Podle absolutních hodnot standardizovaných koeficientů soudíme, ž největší vliv n a1. kanonickou proměnnou má proměnná Lsirka, poté Ovyska.

Koeficienty korelace mezi původními proměnnými a 1. kanonickou proměnnou

	Faktorová strukturní matice (lebky.sta) Korelační proměnné - Kanonické kořeny (vnitřní korelace)
Proměnná	Kořen1
Ldelka	0,687356
Lsirka	-0,218927
Lvyska	0,185569
Ovyska	0,856615
Osirka	0,468024

Největší koeficient korelace pozorujeme u proměnné Ovyska.

Histogramy kanonických skóre v 1. a 2. skupině



Skupinové centroidy 1. kanonické proměnné

	Průměry kan. proměnných (lebky.sta)
Skup.	Kořen1
Sikkim	-0,801461
Lhasa	0,548368

Výpočet dělicího bodu:

$$C = \frac{-0{,}801461 + 0{,}548368}{2} = -0{,}1265$$

V tabulce s daty vytvoříme dvě nové proměnné skore a zarazeni. Do Dlouhého jména proměnné skore napíšeme
=0,02479*Ldelka – 0,09494*Lsirka – 0,01911*Lvyska + 0,12902*Ovyska + 0,06216*Osirka - 6,43096
a do Dlouhého jména proměnné zarazeni napíšeme
=iif(skore > -0,1265;1;0)

V proměnné skore jsou uložena kanonická skóre jednotlivých objektů a v proměnné zarazeni dostaneme zařazení objektů do skupin podle jejich kanonického skóre:

	1 ID	2 Ldelka	3 Lsirka	4 Lvyska	5 Ovyska	6 Osirka	7 skore	8 zarazeni
1	Sikkim	190,5	152,5	145,0	73,5	136,5	-0,98995	0
2	Sikkim	172,5	132,0	125,5	63,0	121,0	-1,43545	0
3	Sikkim	167,0	130,0	125,5	69,5	119,5	-0,63653	0
4	Sikkim	169,5	150,5	133,5	64,5	128,0	-2,79044	0
5	Sikkim	175,0	138,5	126,0	77,5	135,5	0,77197	1
6	Sikkim	177,5	142,5	142,5	71,5	131,0	-0,91497	0
7	Sikkim	179,5	142,5	127,5	70,5	134,5	-0,4902	0
8	Sikkim	179,5	138,0	133,5	73,5	132,5	0,08511	1
9	Sikkim	173,5	135,5	130,5	70,0	133,5	-0,15836	0
10	Sikkim	162,5	139,0	131,0	62,0	126,0	-2,27126	0
11	Sikkim	178,5	135,0	136,0	71,0	124,0	-0,55354	0
12	Sikkim	171,5	148,5	132,5	65,0	146,5	-1,3174	0
13	Sikkim	180,5	139,0	132,0	74,5	134,5	0,296965	1
14	Lhasa	183,0	149,0	121,5	76,5	142,0	0,334435	1
15	Lhasa	169,5	130,0	131,0	68,0	119,0	-0,90427	0
16	Lhasa	172,0	140,0	136,0	70,5	133,5	-0,66337	0
17	Lhasa	170,0	126,5	134,5	66,0	118,5	-0,91559	0
18	Lhasa	182,5	136,0	138,5	76,0	134,0	0,6696	1
19	Lhasa	179,5	135,0	128,5	74,0	132,0	0,49891	1
20	Lhasa	191,0	140,5	140,5	72,5	131,5	-0,19211	0
21	Lhasa	184,5	141,5	134,5	76,5	141,5	0,80416	1
22	Lhasa	181,0	142,0	132,5	79,0	136,5	0,719895	1
23	Lhasa	173,5	136,5	126,0	71,5	136,5	0,212705	1
24	Lhasa	188,5	130,0	143,0	79,5	136,0	1,877875	1
25	Lhasa	175,0	153,0	130,0	76,5	142,0	-0,40608	0
26	Lhasa	196,0	142,5	123,5	76,0	134,0	0,673805	1
27	Lhasa	200,0	139,5	143,5	82,5	146,0	2,260135	1
28	Lhasa	185,0	134,5	140,0	81,5	137,0	1,74141	1
29	Lhasa	174,5	143,5	132,5	74,0	136,5	-0,22875	0
30	Lhasa	195,5	144,0	138,5	78,5	144,0	1,1765	1
31	Lhasa	197,0	131,5	135,0	80,5	139,0	2,41456	1
32	Lhasa	182,5	131,0	135,0	68,5	136,0	0,367855	1

Klasifikační matice

ID	zarazeni 0	zarazeni 1	Řádk. součty
Sikkim	10	3	13
Lhasa	6	13	19
Vš.skup.	16	16	32

Správně zařazeno je 23/32 = 71,9 % lebek, chybně pak 9/32 = 28,1 %.

Příklad na třídění do tří skupin: Pro data o 45 vzorcích rudy (viz cvičení 3) proveďte kanonickou diskriminační analýzu. Pomocí zařazovacího pravidla založeného na kvadrátu Mahalanobisovy vzdálenosti kanonických skóre jednotlivých objektů od skupinových centroidů kanonických proměnných zařaďte vzorky rudy k jednotlivým nalezištím. (Průzkumová analýza dat a test shody vektorů středních hodnot již byly provedeny ve cv. 3.)

Vlastní čísla matice $\mathbf{BE^{-1}}$

	Test chí-kvadrát po odstranění post. kořenů (ropa.sta)					
Kořeny odstraněny	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	2,539965	0,847060	0,179593	69,54102	8	0,000000
1	0,572938	0,603529	0,635753	18,34428	3	0,000373

Prosté a standardizované koeficienty 1. a 2. kanonické proměnné

	Prosté koeficienty (ropa.sta) pro kanonické proměnné	
Proměnná	Kořen1	Kořen2
X1	0,038714	0,02040
X2	-0,078466	0,01247
X3	-0,000385	0,00855
X4	-0,002482	-0,00406
Konstant	1,417310	-3,45403
Vlastní	2,539965	0,57294
KumPodíl	0,815947	1,00000

	Standardiz. koeficienty (ropa.sta) pro kanonické proměnné	
Proměnná	Kořen1	Kořen2
X1	0,603935	0,31818
X2	-0,541523	0,08603
X3	-0,039448	0,87501
X4	-0,627859	-1,02704
Vlastní	2,539965	0,57294
KumPodíl	0,815947	1,00000

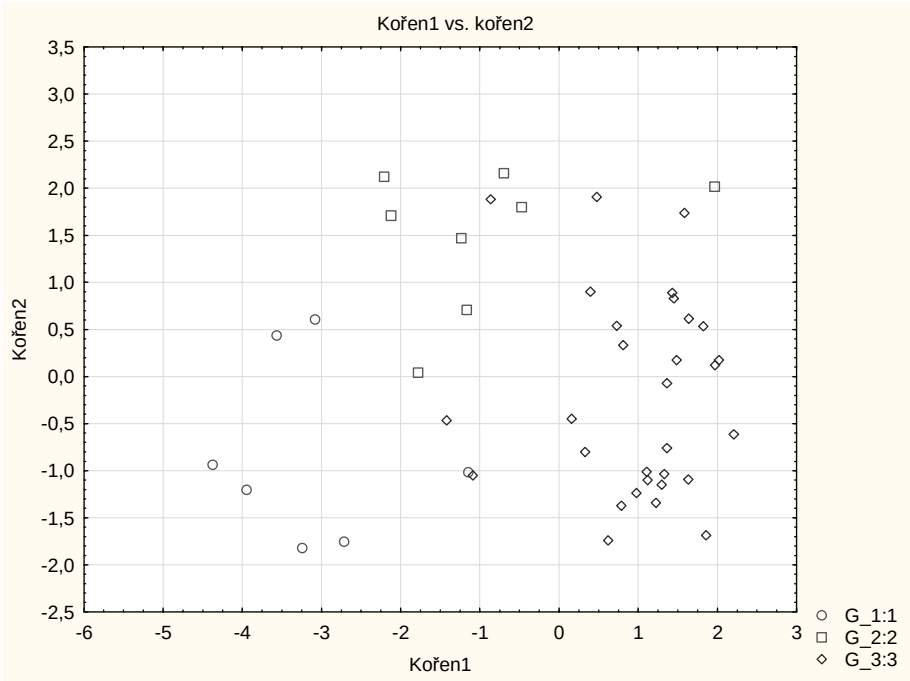
Největší vliv na 1. kanonickou proměnnou má X4 (obsah aromatických uhlovodíků) a na 2. kanonickou proměnnou má největší vliv X3 (obsah nasycených uhlovodíků).

Koeficienty korelace mezi jednotlivými proměnnými a dvěma kanonickými proměnnými

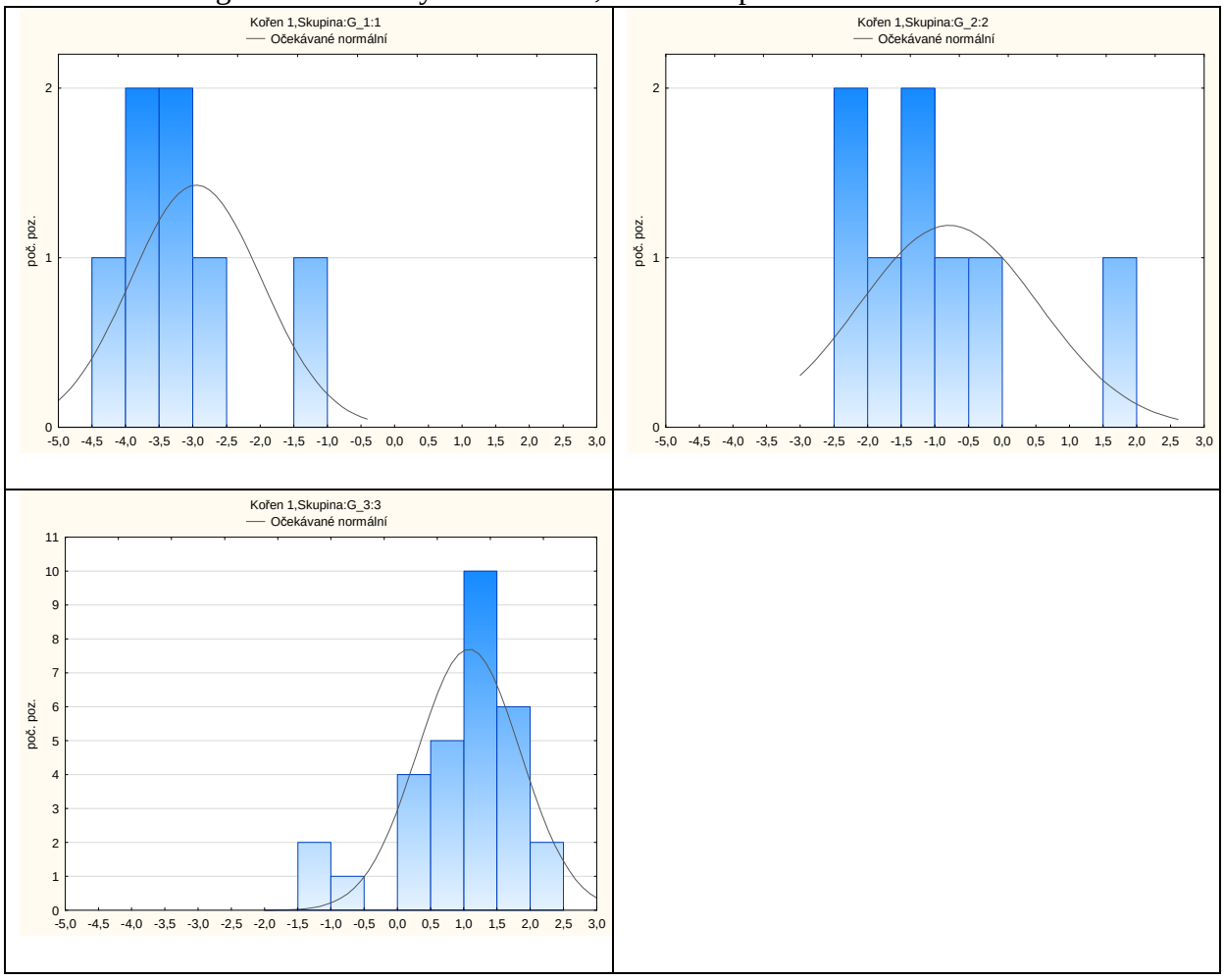
	Faktorová strukturní matice (ropa.sta) Korelační proměnné - Kanonické kořeny (vnitřní korelace)	
Proměnná	Kořen1	Kořen2
X1	0,650362	-0,225347
X2	-0,667040	0,444579
X3	-0,587476	0,443857
X4	-0,354906	-0,628094

Pro 1. kanonickou proměnnou jsou charakteristické proměnné X2 a X1, pro 2. kanonickou proměnnou pak X4.

Znázornění rozmístění objektů na ploše prvních dvou kanonických proměnných



Zobrazení histogramů kanonických skóre v 1., 2. a 3. skupině



Výpočet skupinových centroidů 1. a 2. kanonické proměnné

	Průměry kan. proměnných (ropa.sta)	
Skup.	Kořen1	Kořen2
G_1:1	-3,15356	-0,812170
G_2:2	-0,96574	1,504262
G_3:3	0,99336	-0,211630

Zařazení objektů do skupin podle kvadrátů Mahalanobisových vzdáleností

	ropa.sta													
	1 ID	2 kořen1	3 kořen2	4 centroid11	5 centroid12	6 centroid21	7 centroid22	8 centroid31	**9 centroid32**	10 d1	11 d2	12 d3	13 minimum	14 zarazeni
1	1	-4,37	-0,94	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	1,50119338	17,5673012	29,3106594	1,50119338	1
2	1	-1,15	-1,02	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	4,06128395	6,38224266	5,22594887	4,06128395	1
3	1	-3,08	0,61	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	2,02307861	5,28423696	17,2953052	2,02307861	1
4	1	-3,57	0,44	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	1,73172291	7,90114134	21,2182046	1,73172291	1
5	1	-3,95	-1,20	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	0,78114226	16,214537	25,3709166	0,78114226	1
6	1	-3,24	-1,82	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	1,02611937	16,2502383	20,5331348	1,02611937	1
7	1	-2,71	-1,75	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	1,07848886	13,6700818	16,109897	1,07848886	1
8	2	-2,12	1,71	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	7,4254849	1,37691853	13,4099973	1,37691853	2
9	2	-1,17	0,71	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	6,26505577	0,67264826	5,51816731	0,67264826	2
10	2	1,96	2,02	-3,15356	-0,81217	-0,96574	1,504262	0,99336	-0,21663	34,1837776	8,83984563	5,93059793	5,93059793	3

Klasifikační matice

ID	zarazeni 1	zarazeni 2	zarazeni 3	Řádk. součty
1	7	0	0	7
2	1	6	1	8
3	2	2	26	30
Vš.skup.	10	8	27	45

Relativní četnost správně zařazených případů: (7+6+26)/45 = 39/45 = 86,7 %