Odhalení vnitřní struktury
ve znacích a objektech vícerozměrnou statistickou analýzou

Prof. RNDr. Milan Meloun, DrSc,
Katedra analytické chemie, Univerzita Pardubice,
532 10 Pardubice,
milan. meloun@upce. cz

Souhrn: *Na řadě praktických úloh jsou ukázány diagnostické vlastnosti vybraných metod vícerozměrné statistické analýzy, především metody hlavních komponent, faktorové analýzy a tvorby shluků. Uvedenými metodami je odhalena vnitřní struktura v datech, skryté vnitřní vazby a řada souvislostí ukrytých mezi znaky a mezi objekty.*

Odhalení struktury ve znacích a objektech: Zdrojová matice dat Jímá rozměr *n* x *m.* Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *průzkumovou (exploratorní) analýzu dat,* která umožňuje: a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafu, b) nalézt *vybočující objekty*, resp. jejich znaky, c) stanovit, zda lze použít předpoklad *lineárních vazeb,* d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda hledají 1. *strukturu a vazby ve znacích* nebo *2. strukturu a vazby v objektech.* To znamená a) hledání struktury ve znacích v metrické škále -*faktorová analýza FA, analýza hlavních komponent PCA* a *shluková analýza,* b) hledání struktury v objektech v metrické škále - *shluková analýza,* c) hledání struktury v objektech v metrické i v nemetrické škále - vícerozměrné škálování. d) hledání struktury v objektech v nemetrické škále - *korespondenční analýza.*

Určením struktury a vzájemných vazeb mezi znaky ale i mezi objekty se zabývají techniky redukce znaků na latentní proměnné, *metoda* analýzy hlavních komponent *(PCA), metoda faktorové analýzy (FA)* a shlukování *(CLU).*

Maticový diagram. Diagram ukazuje rozptylové diagramy jednotlivých dvojic znaků. Je zřejmé, že vysoké hodnoty korelačního koeficientu vedou ke zřetelné lineární závislosti (přímka), zatímco nízké hodnoty ukazují, že osoby nejsou v grafu zobrazeny na přímce ale nacházejí se spíše v chaotickém mraku bodů.

Odhalení struktury ve znacích a objektech

Zdrojová matice dat $\boldsymbol{X}$ má rozměr $n \times m$.

Nejprve je třeba provést ***průzkumovou (exploratorní) analýzu dat***:

a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů,

b) nalézt *vybočující objekty*, resp. jejich znaky,

c) stanovit, zda lze použít předpoklad *lineárních vazeb,*

d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Techniky k určení vzájemných vazeb se dále dělí zda hledají

1. *strukturu a vazby ve znacích* nebo
2. *strukturu a vazby v objektech*, což je:

a) hledání struktury ve znacích v metrické škále
faktorová analýza FA,
analýza hlavních komponent PCA a
shluková analýza.

b) hledání struktury v objektech v metrické škále
shluková analýza.

c) hledání struktury v objektech v metrické i v nemetrické škále: *vícerozměrné škálování.*

d) hledání struktury v objektech v nemetrické škále:
korespondenční analýza.

Metody vícerozměrné statistické analýzy umožňuje zpracování ***lineárních vícerozměrných modelů:***

závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných, resp. vazby mezi proměnnými jsou lineární.

Zdrojová matice dat demografického souboru *Lidé*

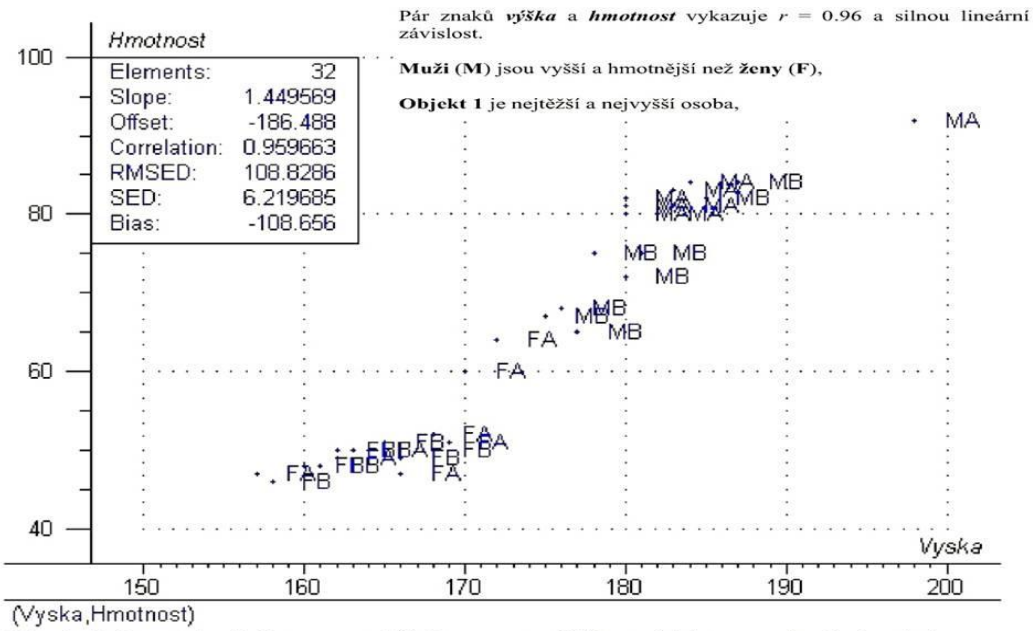
	1 Osoba	2 Vyska	3 Hmotnost	4 Vlasy	5 Boty	6 Vek	7 Prijem	8 Pivo	9 Vino	10 Sex	11 Plavani	12 Puvod	13 IQ
1MA	1MA	198	92	-1	48	48	45000	420	115	-1	98	-1	100
2MA	2MA	184	84	-1	44	33	33000	350	102	-1	92	-1	130
3MA	3MA	183	83	-1	44	37	34000	320	98	-1	91	-1	127
1FA	1FA	166	47	-1	36	32	28000	270	78	1	75	-1	112
2FA	2FA	170	60	1	38	23	20000	312	99	1	81	-1	110
3FA	3FA	172	64	1	39	24	22000	308	91	1	82	-1	102
4MA	4MA	182	80	-1	42	35	30000	398	65	-1	85	-1	140
5MA	5MA	180	80	-1	43	36	30000	388	63	-1	84	-1	129
4FA	4FA	169	51	1	36	24	23000	250	89	1	78	-1	98
5FA	5FA	168	52	1	37	27	23500	260	86	1	78	-1	100
6MA	6MA	183	81	-1	42	37	35000	345	45	-1	90	-1	105
6FA	6FA	157	47	1	36	32	32000	235	92	1	70	-1	127
7FA	7FA	164	50	1	38	41	34000	255	134	1	76	-1	102
8FA	8FA	162	49	1	37	40	34000	265	124	1	75	-1	108
7MA	7MA	180	82	-1	44	43	37000	355	82	-1	88	-1	109

PŘÍKLAD 3.1. *Zobrazování znaků zdrojové matice dat Lidé*

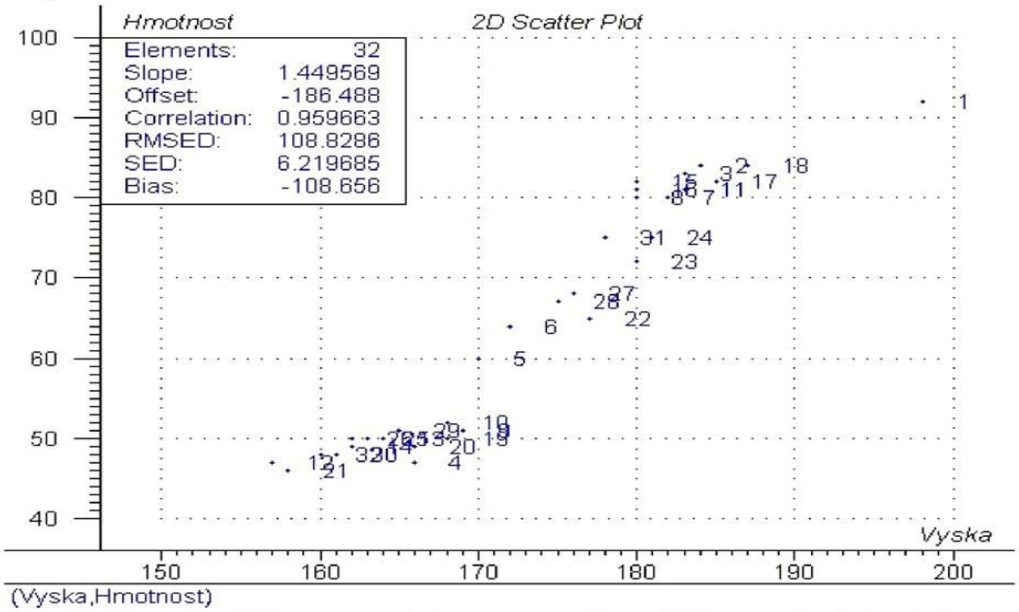
Rozptylovým diagramem je třeba nalézt trendy a vnitřní párové vztahy mezi znaky zdrojové matice *Lidé* a odhalit korelaci dvou znaků.

Který pár vykazuje nejsilnější korelaci tj. nejvyšší hodnotu *r*.

7FB	7FB	161	48	1	35	41	31500	116	196	1	75	1	120
8MB	8MB	178	75	-1	42	30	24000	203	208	-1	81	1	118
8FB	8FB	160	48	1	35	40	31000	118	198	1	74	1	129



Obr. 3.1b Rozptylový diagram znaků *Hmotnost* a *Výška* s objekty označenými svými jmény dat *Lidé*.



Obr. 3.1a Rozptylový diagram znaků *Hmotnost* a *Výška* s objekty označenými svými pořadovými čísly, dat *Lidé*.

○ *Řešení:*

a) ***2D- a 3D-rozptylové diagramy:*** Objekty jsou indikovány pořadovými čísly (obr. 3.1a) a nebo svými jmény (obr. 3.1b):

Pár znaků ***výška*** a ***hmotnost*** vykazuje $r = 0.96$ a silnou lineární závislost.

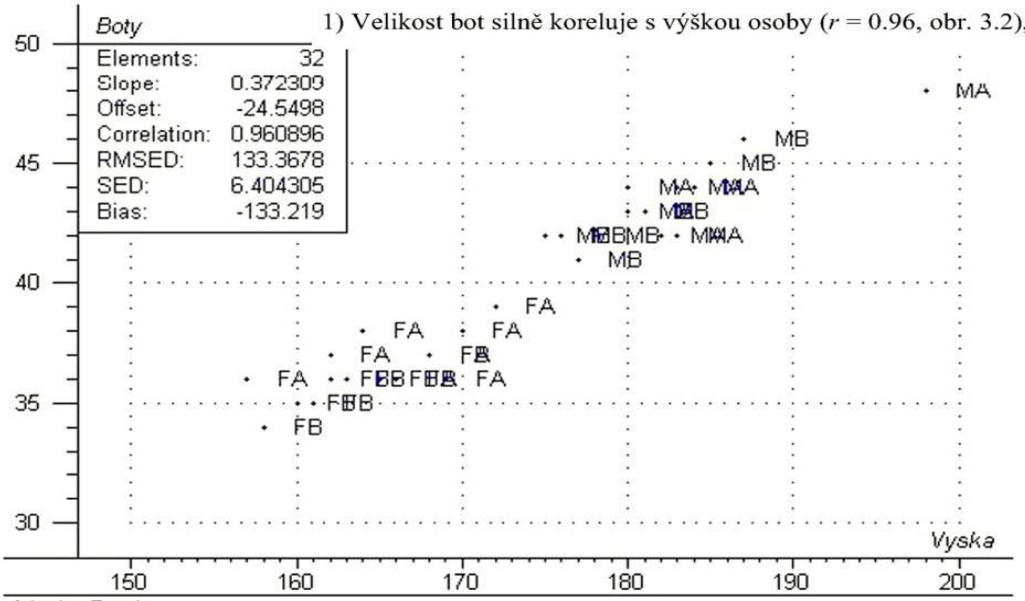
Muži (M) jsou vyšší a hmotnější než **ženy (F)**,

Objekt 1 je nejtěžší a nejvyšší osoba,

Objekty 12 a **21** jsou umístěny na opačném konci proložené regresní přímky a týkají se nejmenších a nejlehčích osob.

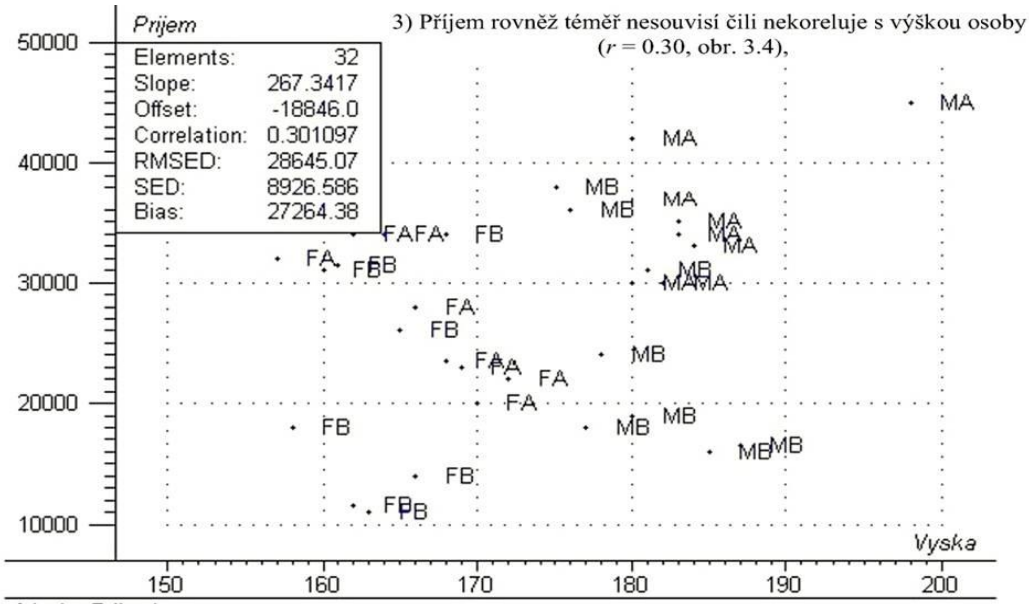
Vedle grafu jsou i **popisné statistiky** v rámečku jako jsou směrnice (**Slope**), úsek (**Offset**), *r* (**Correlation**), atd.

1) Velikost bot silně koreluje s výškou osoby (r = 0.96, obr. 3.2),



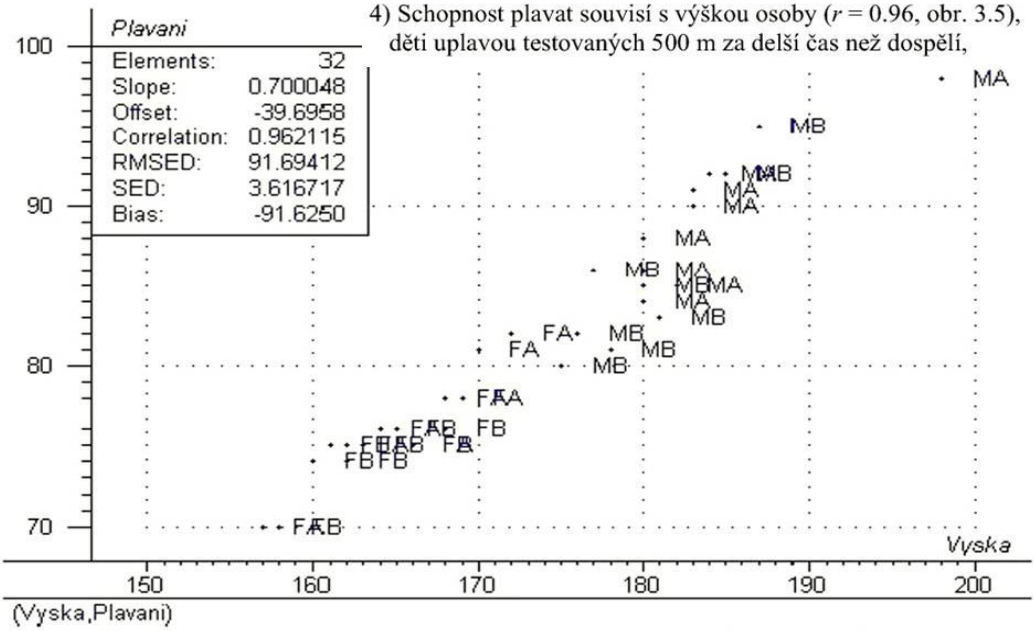
Obr. 3.2 Rozptylový diagram znaků *Boty* a *Výška* dat *Lidé*, (UNSCRAMBLER).

3) Příjem rovněž téměř nesouvisí čili nekoreluje s výškou osoby (r = 0.30, obr. 3.4),

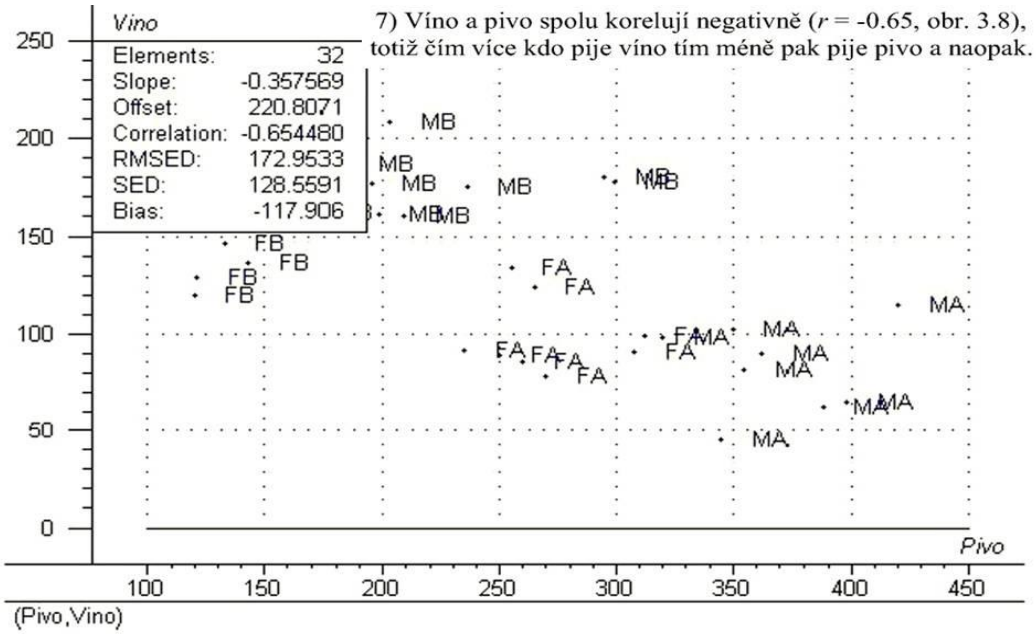


Obr. 3.4 Rozptylový diagram znaků *Příjem* a *Výška* dat *Lidé*, (UNSCRAMBLER).

4) Schopnost plavat souvisí s výškou osoby (r = 0.96, obr. 3.5), děti uplavou testovaných 500 m za delší čas než dospělí,



Obr. 3.5 Rozptylový diagram znaků *Plavání* a *Výška* dat *Lidé*, (UNSCRAMBLER).



7) Víno a pivo spolu korelují negativně (r = -0.65, obr. 3.8), totiž čím více kdo pije víno tím méně pak pije pivo a naopak.

Obr. 3.8 Rozptylový diagram znaků *Víno* a *Pivo* dat *Lidé*, (UNSCRAMBLER).

Interpretace a komentáře 2D-diagramů:

1) Velikost bot silně koreluje s výškou osoby (r = 0.96, obr. 3.2),

2) Věk téměř nekoreluje s výškou osoby (r = 0.24, obr. 3.3),

3) Příjem rovněž téměř nesouvisí čili nekoreluje s výškou osoby (r = 0.30, obr. 3.4),

4) Schopnost plavat souvisí s výškou osoby (r = 0.96, obr. 3.5), děti uplavou testovaných 500 m za delší čas než dospělí,

5) Pivo středně silně koreluje s výškou osoby (r = 0.71, obr. 3.6),

6) Víno už vůbec nekoreluje s výškou osoby (r = -0.13, obr. 3.7),

7) Víno a pivo spolu korelují negativně (r = -0.65, obr. 3.8), totiž čím více kdo pije víno tím méně pak pije pivo a naopak.

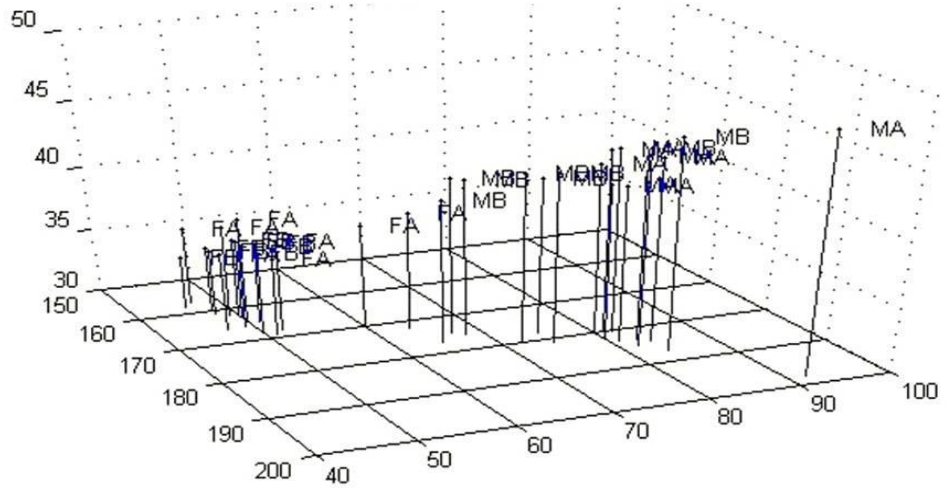
K zobrazení všech kombinací *m* znaků by se jednalo o *m*(*m* - 1)/2 kombinací.

To by bylo pro nás značně nepohodlné, a proto budeme hledat nějaký schůdnější způsob.

Pak snadno odpovíme na otázky:

1) Je znak schopnost plavat korelovaný se znakem **výška** nebo se znakem **hmotnost**?
2) Liší se **muži** od **žen** z hlediska **schopnosti plavání**?
3) Existuje v těchto diagramech nějaký *shluk* osob?
4) Které osoby se nejvíce *odlišují* od ostatních?
5) Které osoby jsou si *podobné* a které jsou si naopak zcela *nepodobné*?
6) Dají se najít i jiné kombinace tří znaků, mezi kterými existuje obdobně *silná korelace*?

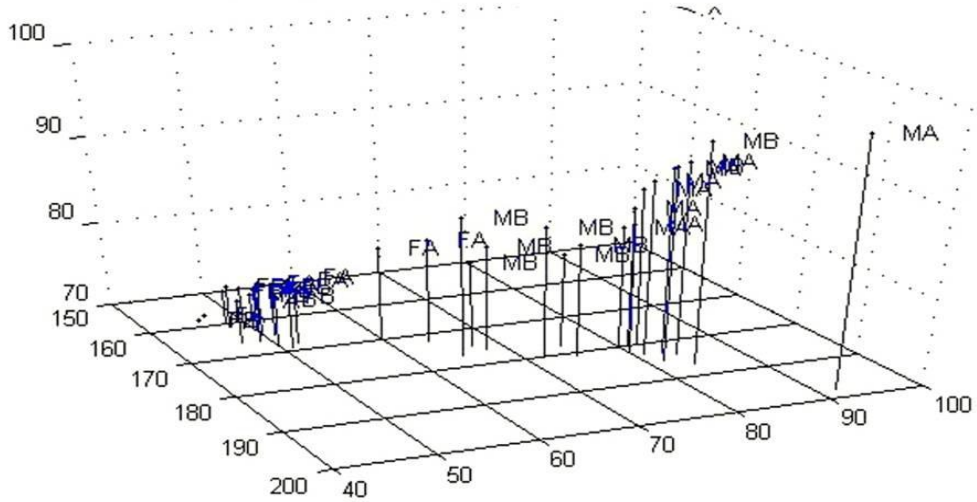
3D Scatter Plot 1) U trojice znaků **výška-hmotnost-velikost boty** (obr. 3.9) je zřejmá *silná pozitivní korelace* všech tří znaků.



(Vyska,Hmotnost,Boty)

Obr. 3.9 Rozptylový 3D-diagram znaků *Výška-Hmotnost-Boty* dat *Lidé*

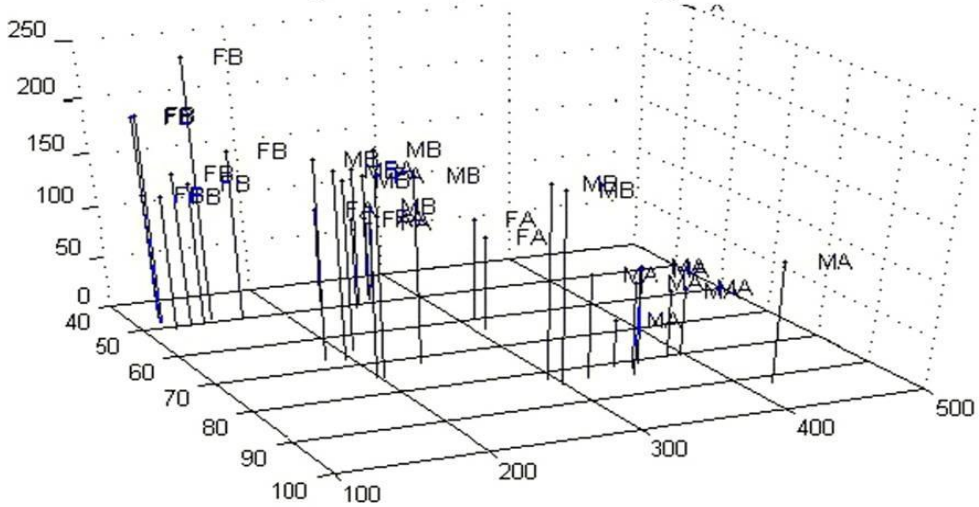
3D Scatter Plot 2) U trojice znaků **výška-hmotnost-schopnost plavat** (obr. 3.10) je zřejmá *silná pozitivní korelace* všech tří znaků.



(Vyska,Hmotnost,Plavani)

Obr. 3.10 Rozptylový 3D-diagram znaků *Výška-Hmotnost-Plavání* dat *Lidé*

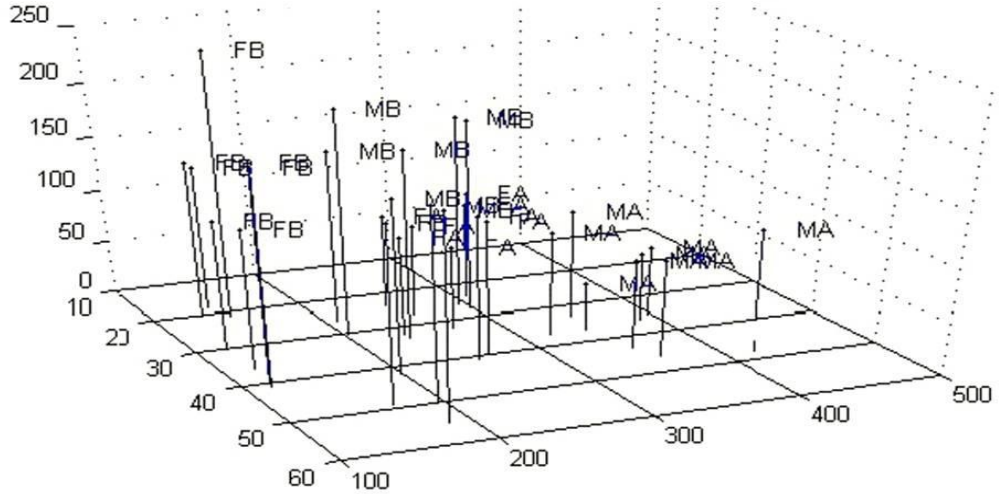
3D Scatter Plot 3) Naopak u trojice znaků **hmotnost-pivo-víno** (obr. 3.11) je vidět spíše *slabou negativní korelaci.*



(Hmotnost,Pivo,Vino)

Obr. 3.11 Rozptylový 3D-diagram znaků *Hmotnost-Pivo-Víno* dat *Lidé*

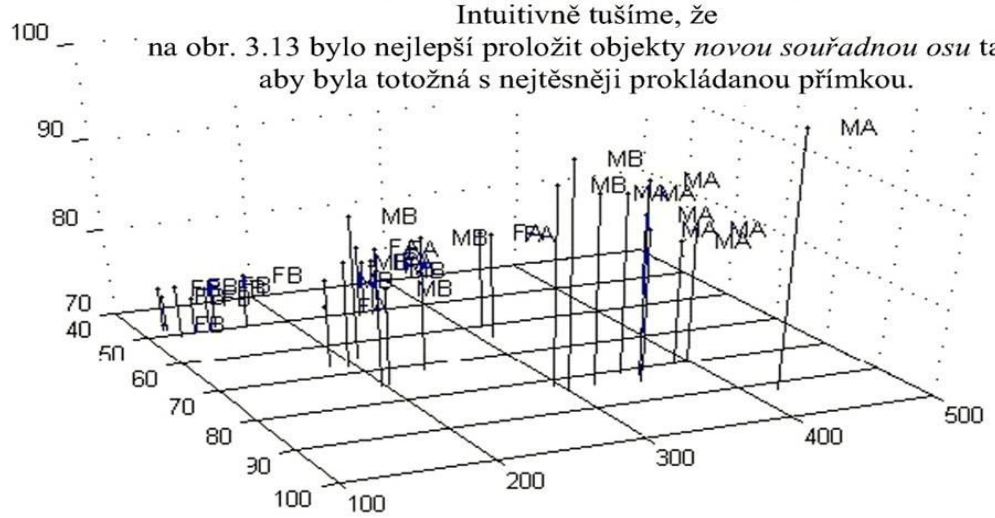
3D Scatter Plot 4) U trojice **věk-pivo-víno** (obr. 3.12)
je zřejmá spíše *slabou negativní korelaci.*



(Vek,Pivo,Vino)

Obr. 3.12 Rozptylový 3D-diagram znaků *Věk-Pivo-Víno* dat *Lidé*

3D Scatter Plot 5) U trojice znaků **hmotnost-pivo-schopnost plavání** (obr. 3.13)
je vidět středně *silná pozitivní korelace.*
Intuitivně tušíme, že
na obr. 3.13 bylo nejlepší proložit objekty *novou souřadnou osu* tak,
aby byla totožná s nejtěsněji prokládanou přímkou.



(Hmotnost,Pivo,Plavani)

Obr. 3.13 Rozptylový 3D-diagram znaků *Hmotnost-Pivo-Plavání* dat *Lidé*

Interpretace a komentáře 3D-diagramů:

1) U trojice znaků **výška-hmotnost-velikost boty** (obr. 3.9)
je zřejmá *silná pozitivní korelace* všech tří znaků.

2) U trojice znaků **výška-hmotnost-schopnost plavat** (obr. 3.10)
je zřejmá *silná pozitivní korelace* všech tří znaků.

3) Naopak u trojice znaků **hmotnost-pivo-víno** (obr. 3.11)
je vidět spíše *slabou negativní korelaci.*

4) U trojice **věk-pivo-víno** (obr. 3.12)
je zřejmá spíše *slabou negativní korelaci.*

5) U trojice znaků **hmotnost-pivo-schopnost plavání** (obr. 3.13)
je vidět středně *silná pozitivní korelace.*

6) Zatímco trojice **věk-víno-schopnost plavání** ukazuje,
že znaky spolu téměř nekorelují.

Zdrojová matice dat demografického souboru *Lidé*

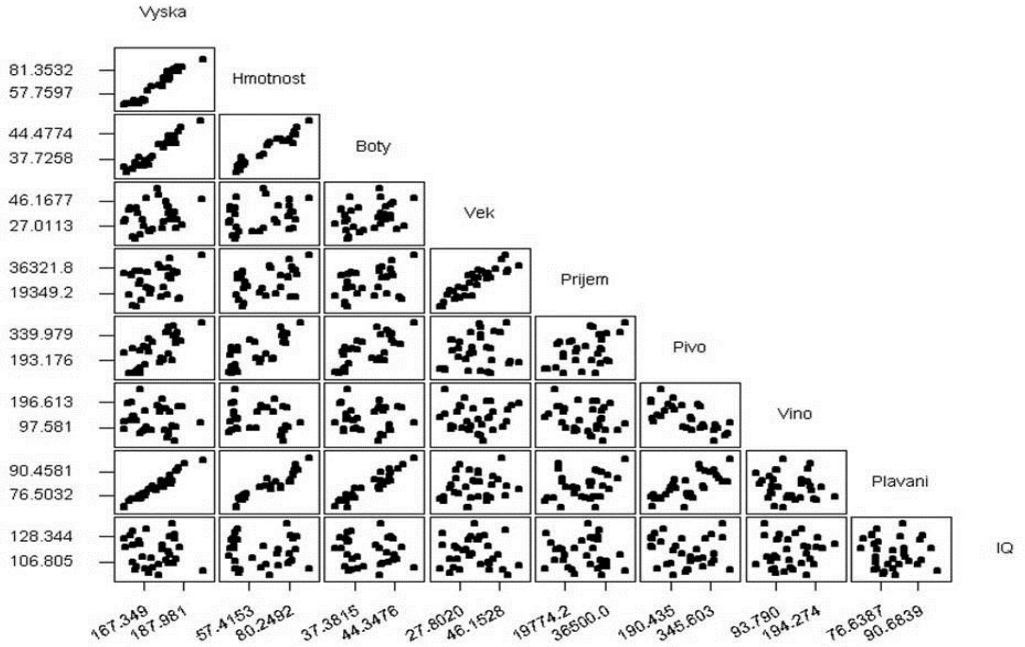
	1 Osoba	2 Vyska	3 Hmotnost	4 Vlasy	5 Boty	6 Vek	7 Prijem	8 Pivo	9 Vino	10 Sex	11 Plavani	12 Puvod	13 IQ
1MA	1MA	198	92	-1	48	48	45000	420	115	-1	98	-1	100
2MA	2MA	184	84	-1	44	33	33000	350	102	-1	92	-1	130
3MA	3MA	183	83	-1	44	37	34000	320	98	-1	91	-1	127
1FA	1FA	166	47	-1	36	32	28000	270	78	1	75	-1	112
2FA	2FA	170	60	1	38	23	20000	312	99	1	81	-1	110
3FA	3FA	172	64	1	39	24	22000	308	91	1	82	-1	102
4MA	4MA	182	80	-1	42	35	30000	398	65	-1	85	-1	140
5MA	5MA	180	80	-1	43	36	30000	388	63	-1	84	-1	129
4FA	4FA	169	51	1	36	24	23000	250	89	1	78	-1	98
5FA	5FA	168	52	1	37	27	23500	260	86	1	78	-1	100
6MA	6MA	183	81	-1	42	37	35000	345	45	-1	90	-1	105
6FA	6FA	157	47	1	36	32	32000	235	92	1	70	-1	127
7FA	7FA	164	50	1	38	41	34000	255	134	1	76	-1	102
8FA	8FA	162	49	1	37	40	34000	265	124	1	75	-1	108
7MA	7MA	180	82	-1	44	43	37000	355	82	-1	88	-1	109
8MA	8MA	180	81	-1	44	46	42000	362	90	-1	86	-1	113
1MB	1MB	185	82	-1	45	26	16000	295	180	-1	92	1	109
2MB	2MB	187	84	-1	46	27	16500	299	178	-1	95	1	119
1FB	1FB	168	50	1	37	49	34000	170	162	1	76	1	135
2FB	2FB	166	49	1	36	21	14000	150	245	1	75	1	123
3FB	3FB	158	46	1	34	30	18000	120	120	1	70	1	119
3MB	3MB	177	65	-1	41	26	18000	209	160	-1	86	1	120
4MB	4MB	180	72	-1	43	33	19000	236	175	-1	85	1	115
5MB	5MB	181	75	-1	43	42	31000	198	161	-1	83	1	105
4FB	4FB	163	50	1	36	18	11000	143	136	1	75	1	102
5FB	5FB	162	50	1	36	20	11500	133	146	1	74	1	132
6MB	6MB	176	68	-1	42	50	36000	195	177	-1	82	1	96
7MB	7MB	175	67	1	42	55	38000	185	187	-1	80	1	105
6FB	6FB	165	51	1	36	36	26000	121	129	1	76	1	126
7FB	7FB	161	48	1	35	41	31500	116	196	1	75	1	120
8MB	8MB	178	75	-1	42	30	24000	203	208	-1	81	1	118
8FB	8FB	160	48	1	35	40	31000	118	198	1	74	1	129

Tabulka 3.1 Matice korelačních koeficientů *r* znaků zdrojové matice dat *Lidé*.

Pod korelačním koeficientem je vypočtená hladina významnosti p.

Je-li $p < 0.05$, je r statisticky významný a v tabulce vyznačeno **tučně**.

	Vyska	*Hmotn.*	*Boty*	*Vek*	*Prijem*	*Pivo*	*Vino*	*Plav.*
Hmotn.	**0.96**							
	0							
Boty	**0.961**	**0.969**						
	0	0						
Vek	0.241	0.254	0.323					
	0.184	0.16	0.071					
Prijem	0.301	0.335	**0.354**	**0.885**				
	0.094	0.061	0.047	0				
Pivo	**0.715**	**0.738**	**0.697**	0.128	0.417			
	0	0	0	0.484	0.018			
Vino	-0.138	-0.197	-0.09	0.027	-0.297	**-0.654**		
	0.451	0.28	0.629	0.884	0.099	0		
Plavani	**0.962**	**0.937**	**0.933**	0.158	0.252	**0.725**	-0.166	
	0	0	0	0.387	0.165	0	0.365	
IQ	-0.122	-0.03	-0.12	-0.08	-0.107	-0.107	0.068	-0.116
	0.507	0.854	0.512	0.673	0.561	0.559	0.713	0.526



Směry maximálního rozptylu v rozptylovém diagramu:

Intuitivně tušíme, že
na obr. 3.13 bylo nejlepší proložit objekty *novou souřadnou osu* tak,
aby byla totožná s nejtěsněji prokládanou přímkou.

Vhodný počet potřebných os není 3 ale pouze 1.

Je to proto, že znaky x_1, x_2 a x_3 spolu silně korelují
a objekty leží na prostorové přímce.

Nejtěsněji objekty prokládaná přímka,
a tím pádem i **nová souřadnicová osa** zvaná ***PC*1**
leží ve směru *největšího rozptylu objektů*.

Souřadnicová osa *PC*1 **není** obecně přitom paralelní
s žádnou z původních os znaků x_1, x_2 a x_3.

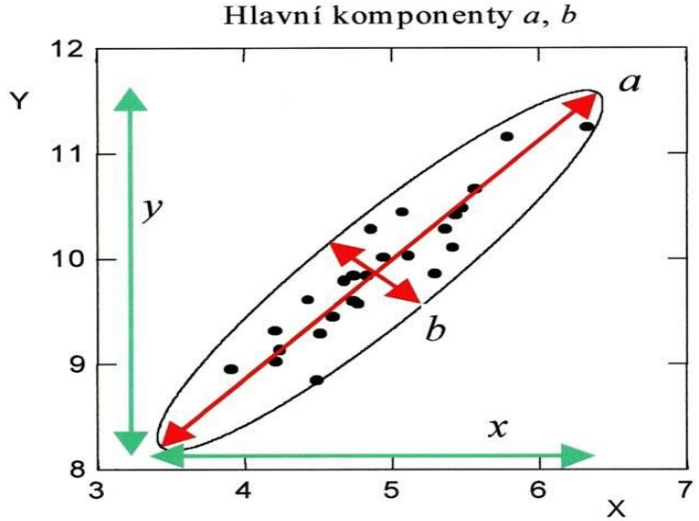
Novou souřadnou osu *PC*1 nazveme ***první hlavní komponentou.***
Bude ležet *ve směru maximálního rozptylu* zdrojové matice dat ***X***.

S touto osou je spjata jakási nová latentní proměnná.

Metoda hlavních komponent PCA poskytne nejenom
tuto první komponentu ale také další hlavní komponenty.

Obvykle jedna, dvě, maximálně však tři nové souřadnicové osy
*PC*1, *PC*2 a *PC*3 vystihnou dohromady největší podíl
proměnlivosti dat.

Princip hlavních komponent:



Pravoúhlým promítnutím dat na osu *a* získáme nejvíce informace o původních datech na jediné přímce.

První hlavní komponenta proložením nejmenšími čtverci

Je třeba proložit přímku v prostoru x_1, x_2 a x_3 umístěnými objekty.

Jak?

Z každého objektu spustíme k této přímce kolmici
a obdržíme e_i *reziduum*.

Nejtěsněji proloženou přímku dostaneme,
když suma čtverců reziduí bude dosahovat své minimální velikosti

$$RSS = \sum_{i=1}^{n} e_i^2 = \text{minimum} \qquad (3.1)$$

Existují dvě kritéria ke konstrukci *první hlavní komponenty:*
1. Směr největšího rozptylu,
2. Metoda nejmenších čtverců MNČ.

Rozšíření k dalším hlavním komponentám

Může nastat případ, že první hlavní komponenta *PC*1
nestačí dostatečně popsat rozptýlení objektů.

Druhá hlavní komponenta *PC*2 je ortogonální vůči první
(*je na ni kolmá*) a je umístěna
do směru druhého největšího rozptýlení objektů v rovině.

U elipsoidu objektů jsou do jeho prostorových os umístěny
tři osy největšího rozptýlení objektů
čili osy tří hlavních komponent *PC*1, *PC*2 a *PC*3.

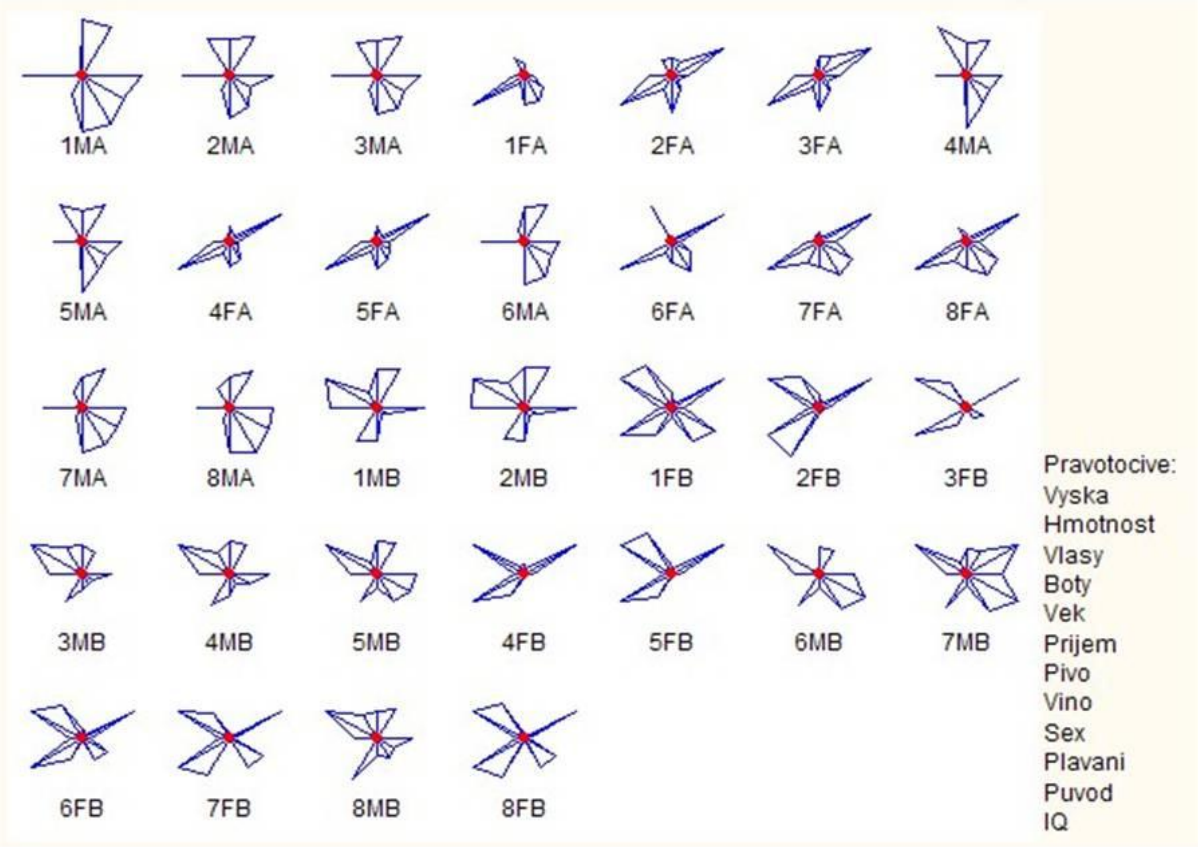
Nové tři osy nekorelují, jsou vzájemně ortogonální tj. **na sebe kolmé.**
PCA umožňuje projekci zobecnit na libovolný počet *m* znaků.

○ *Závěr:*

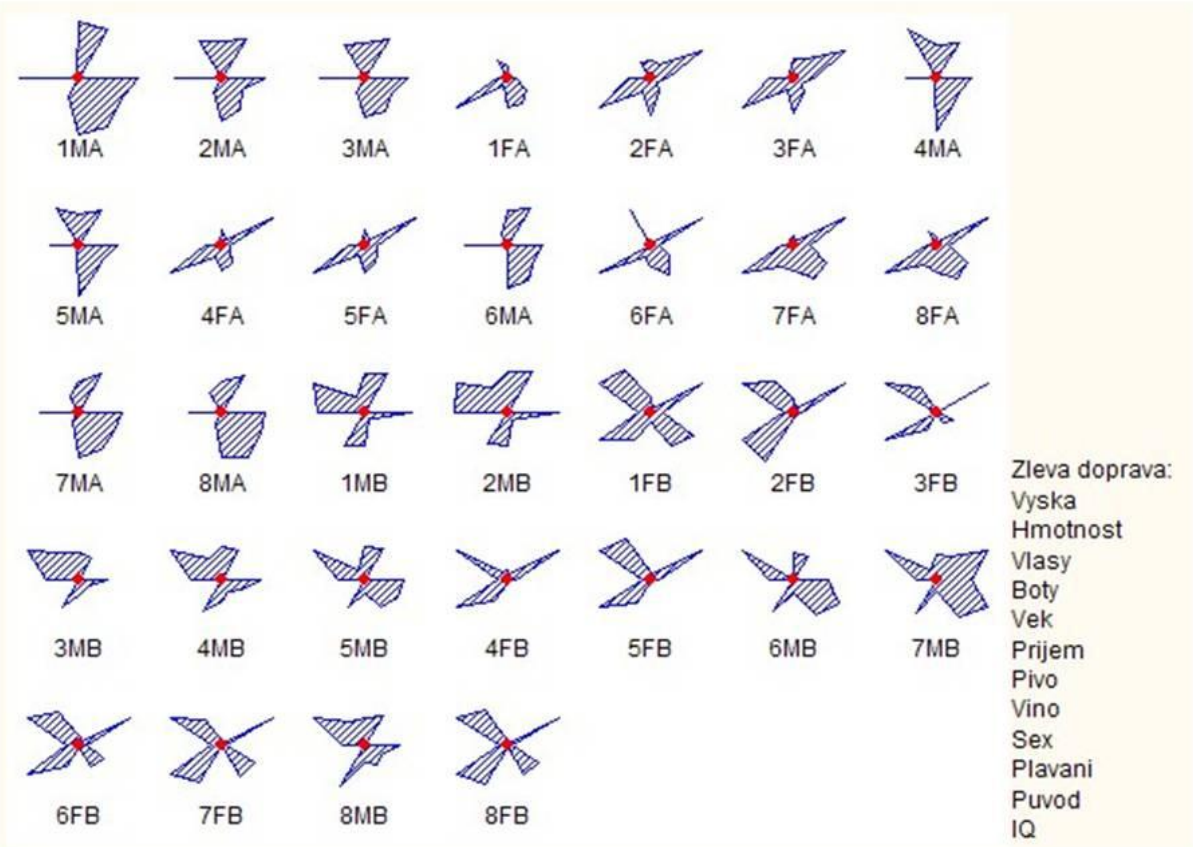
Vyšetřili jsme matici dat v 2D- a 3D-rozptylových diagramech:
a získali řadu závěrů:

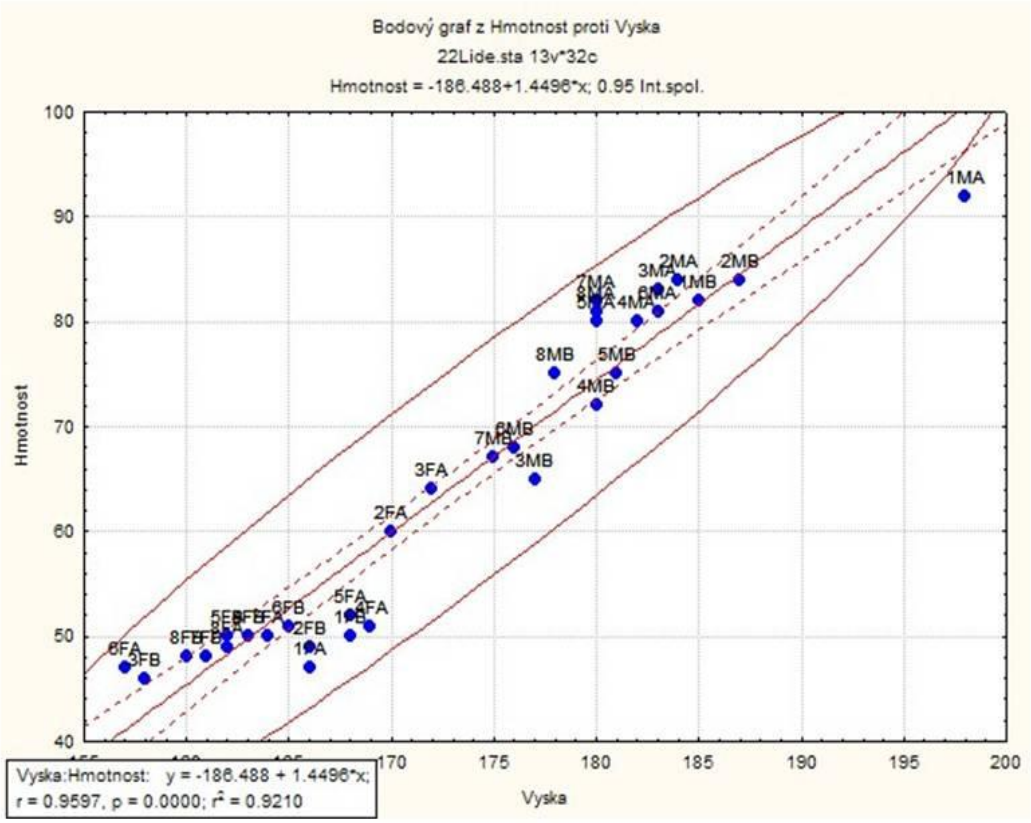
- čím vyšší jsou osoby, tím jsou těžší,
- nebo schopnost plavat koreluje s výškou a hmotností osob.
- existují minimálně dva shluky, jeden pro muže
  a jeden pro ženy.
- některé znaky ukázaly shluky odpovídající původu osob.

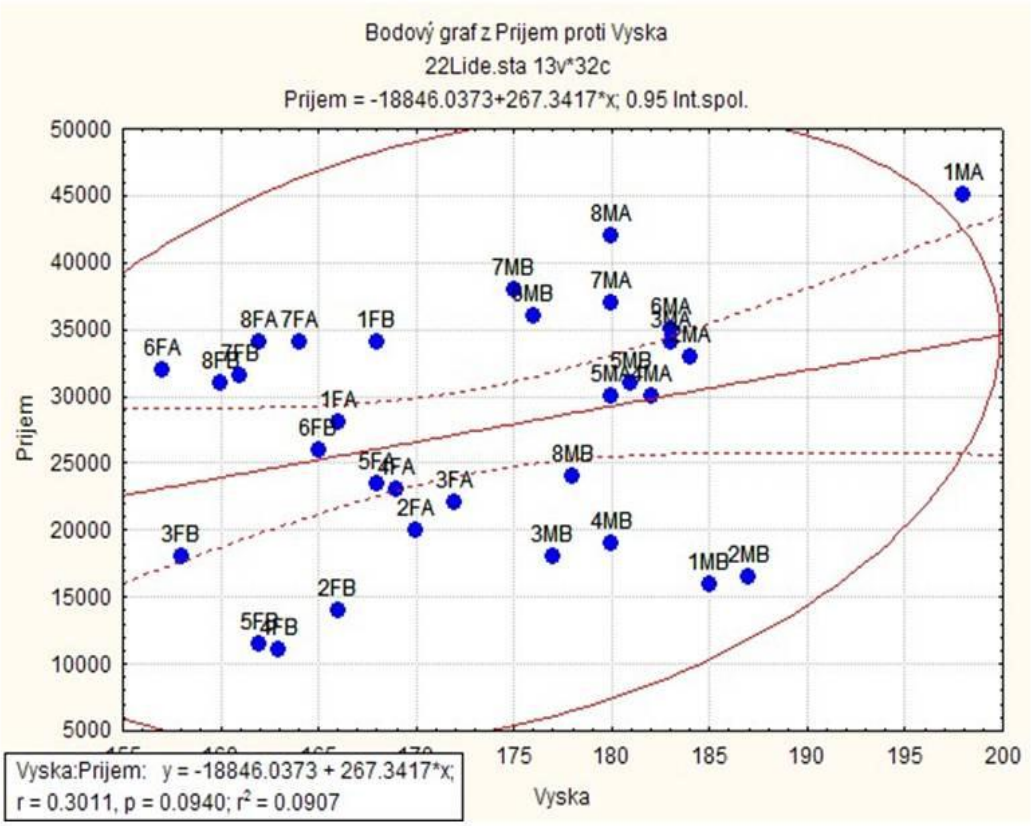
EDA: Symbolové grafy hvězdiček



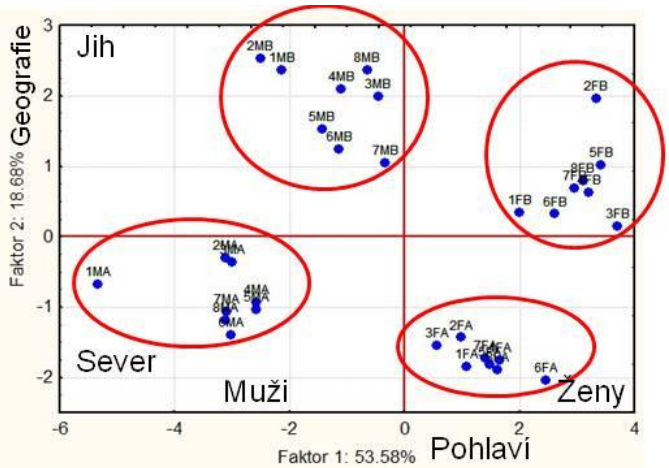
EDA: Symbolové grafy mnohoúhelníků

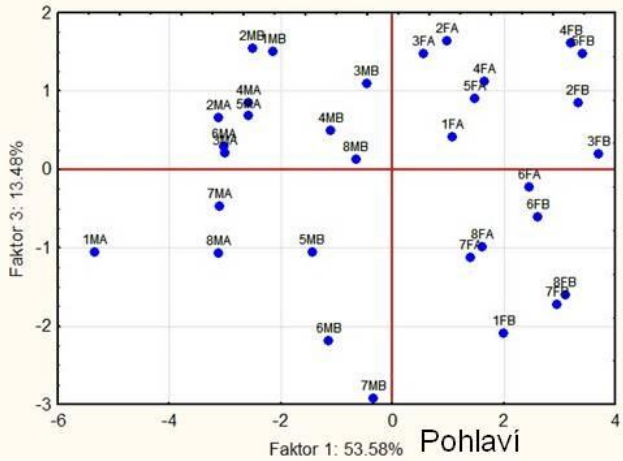


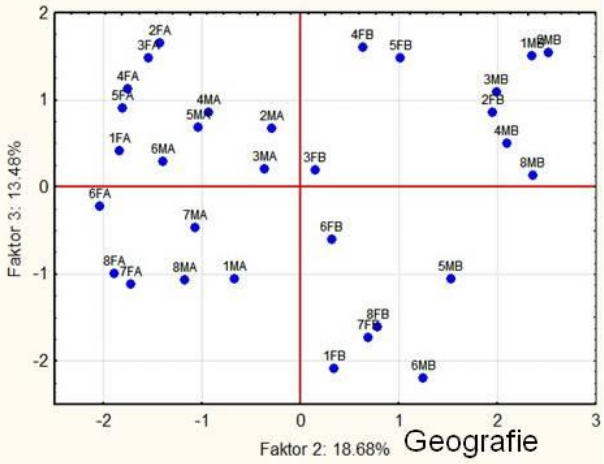

Bodový graf z Hmotnost proti Vyska
22Lide.sta 13v*32c
Hmotnost = -186.488+1.4496*x; 0.95 Int.spol.
Hmotnost
Vyska
Vyska:Hmotnost: y = -186.488 + 1.4496*x;
r = 0.9597, p = 0.0000; r² = 0.9210

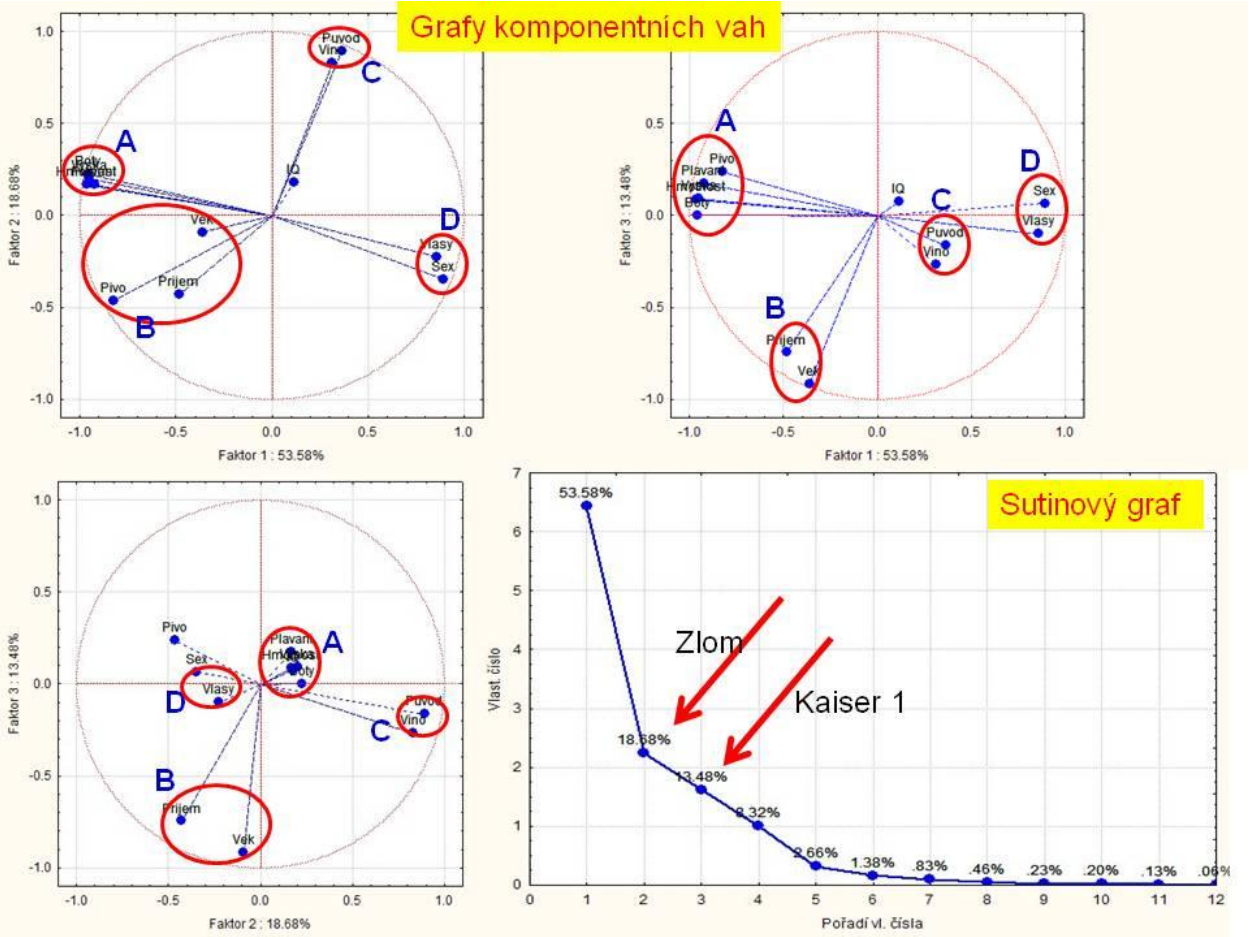

Bodový graf z Prijem proti Vyska
22Lide.sta 13v*32c
Prijem = -18846.0373+267.3417*x; 0.95 Int.spol.
Prijem
Vyska
Vyska:Prijem: y = -18846.0373 + 267.3417*x;
r = 0.3011, p = 0.0940; r² = 0.0907

3 rozptylové diagramy komponentního skóre , tj. grafy 1-2, 1-3, 2-3

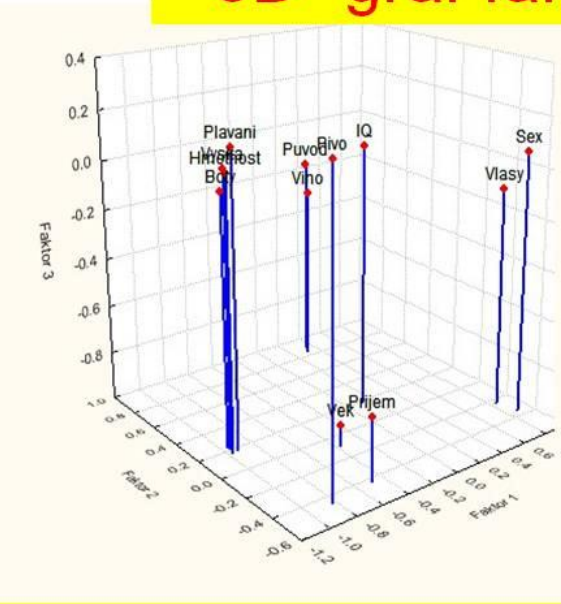


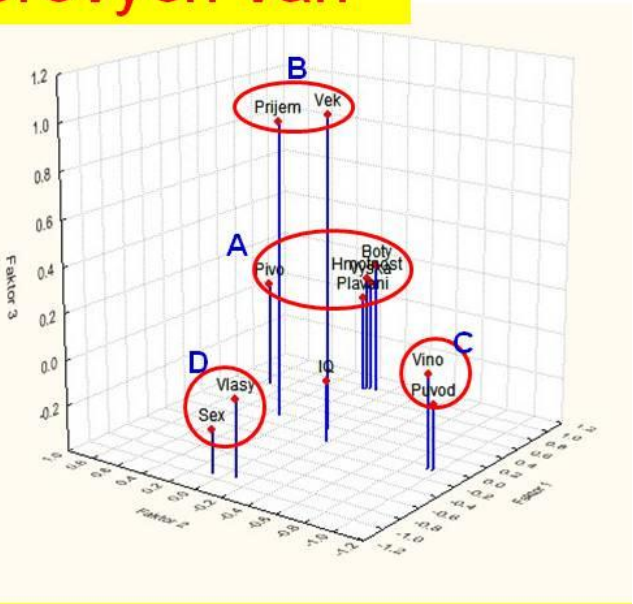






3D- graf faktorových vah

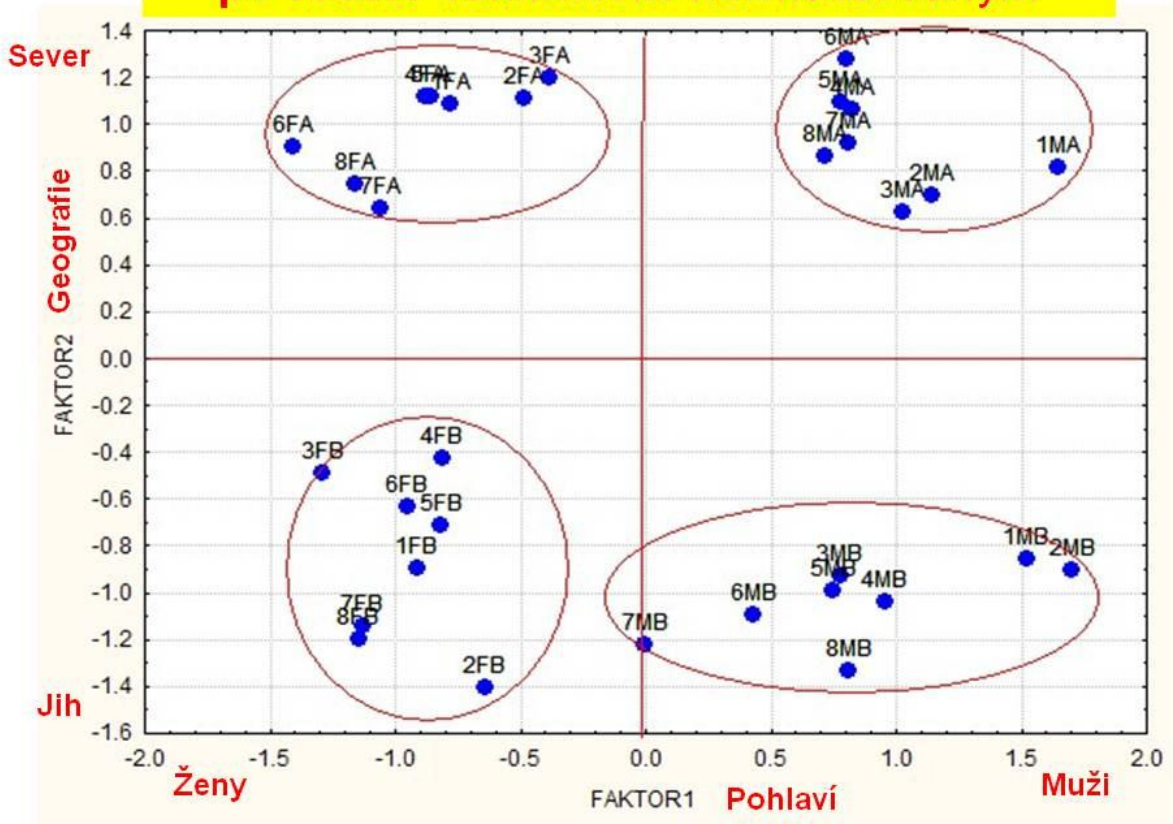




Faktorová analýza:
Graf faktorových vah **bez rotace**

Faktorová analýza:
Graf faktorových vah **po rotaci Varimaxem normalizovaným**

Rozptylový diagram faktorového skóre **po rotaci** Varimaxem normalizovaným



Závěr:

Určení struktury a vazeb

Zdrojová matice má rozměr $n \times m$.

1. **Hledání struktury *v proměnných* v metrické škále:** *faktorová analýza* a *analýza hlavních komponent*.

2. **Hledání struktury *v objektech* v metrické škále**: *shluková analýza*.

3. **Hledání struktury *v objektech* v metrické i v nemetrické škále:** *vícerozměrné škálování.*

4. **Hledání struktury *v objektech* v nemetrické škále:** *korespondenční analýza.*

Poděkování: Práce vznikla za podpory grantu Ministerstva zdravotnictví NS9831-4/2008 a vědeckých záměrů Ministerstva školství, kultury a mládeže MSMT0021627502.

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.
[3] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.
[4] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[5] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.
[6] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[7] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982
[9] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester 1992.
[10] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[11] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[12] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004..
[13] Meloun M. , Militký J., Hill M., *Počítačová analýza vícerozměrných dat v příkladech,* Academia Praha 2005.
[15] Meloun M. , Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002, 2006.