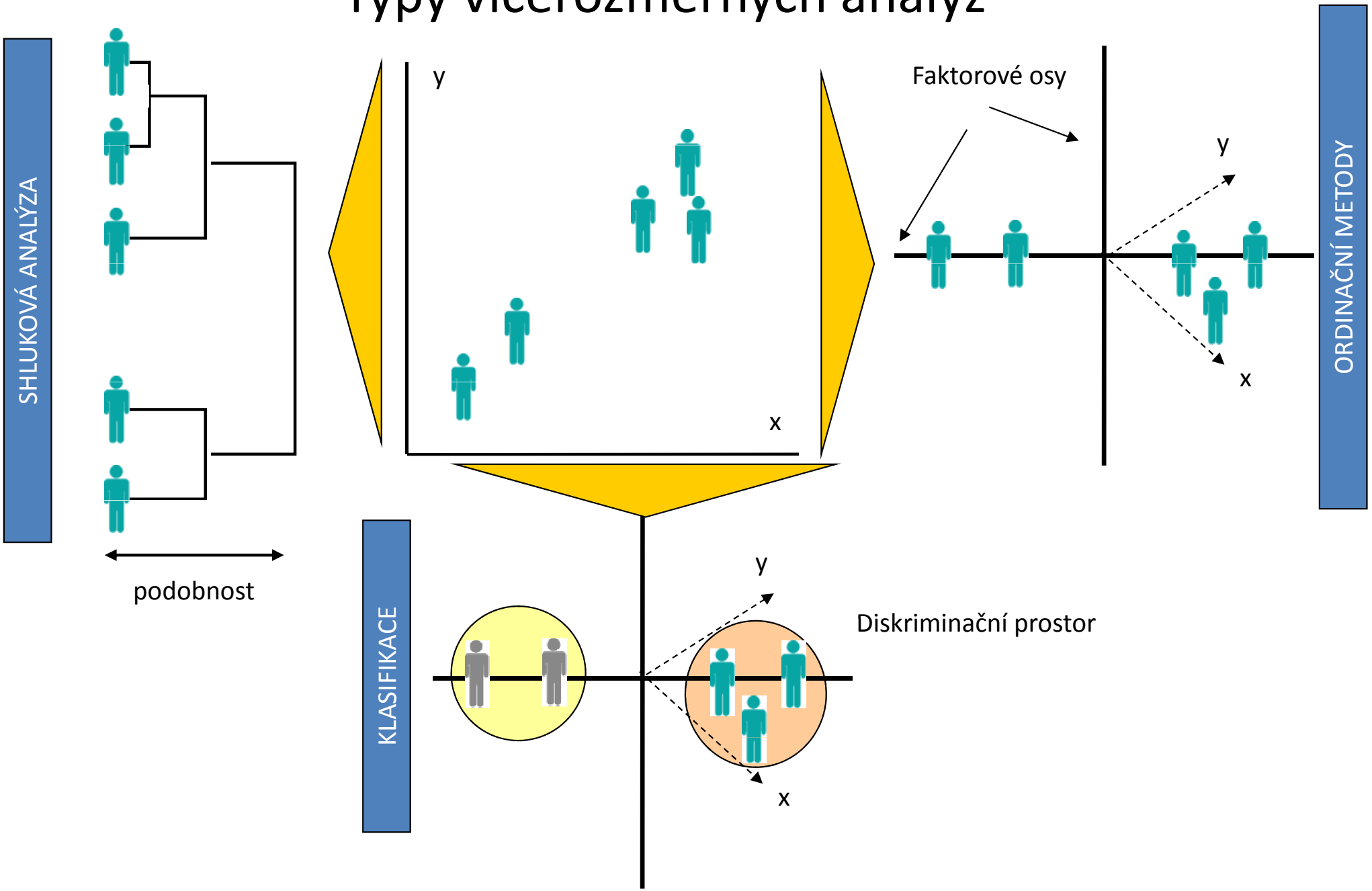
Vícerozměrné statistické metody

Diskriminační analýza

Jiří Jarkovský, Simona Littnerová

Typy vícerozměrných analýz

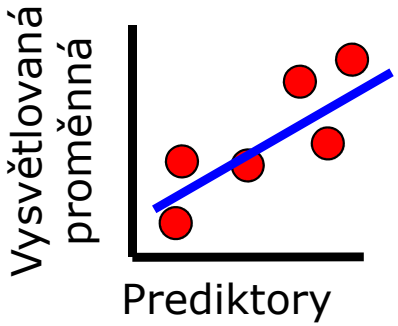


Obecné zásady tvorby predikčních modelů

- Požadavky na kvalitní predikční model
  - Maximální predikční síla
  - Maximální interpretovatelnost
  - Minimální složitost
- Tvorba modelů
  - Neobsahuje redundantní proměnné
  - Je otestován na nezávislých datech
- Výběr proměnných
  - Algoritmy typu dopředné a zpětné eliminace jsou pouze pomocným ukazatelem při výběru proměnných finálního modelu
  - Při výběru proměnných se uplatní jak klasické statistické metody (ANOVA), tak expertní znalost významu proměnných a jejich zastupitelnosti

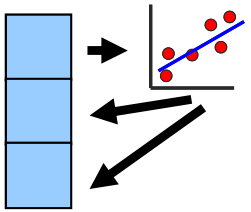
Vytváření modelů

1.Tvorba modelu



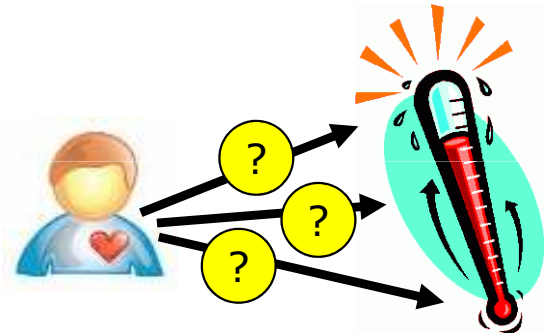
- Parametry ovlivňující vysvětlovanou charakteristiku pacienta
- Rovnice umožňující predikci
- Platnost modelu pouze v rozsahu prediktorů

2.Validace modelu



- Nebezpečí „přeučení" modelu
- Testování modelu na známých datech
- Krosvalidace

3. Aplikace modelu



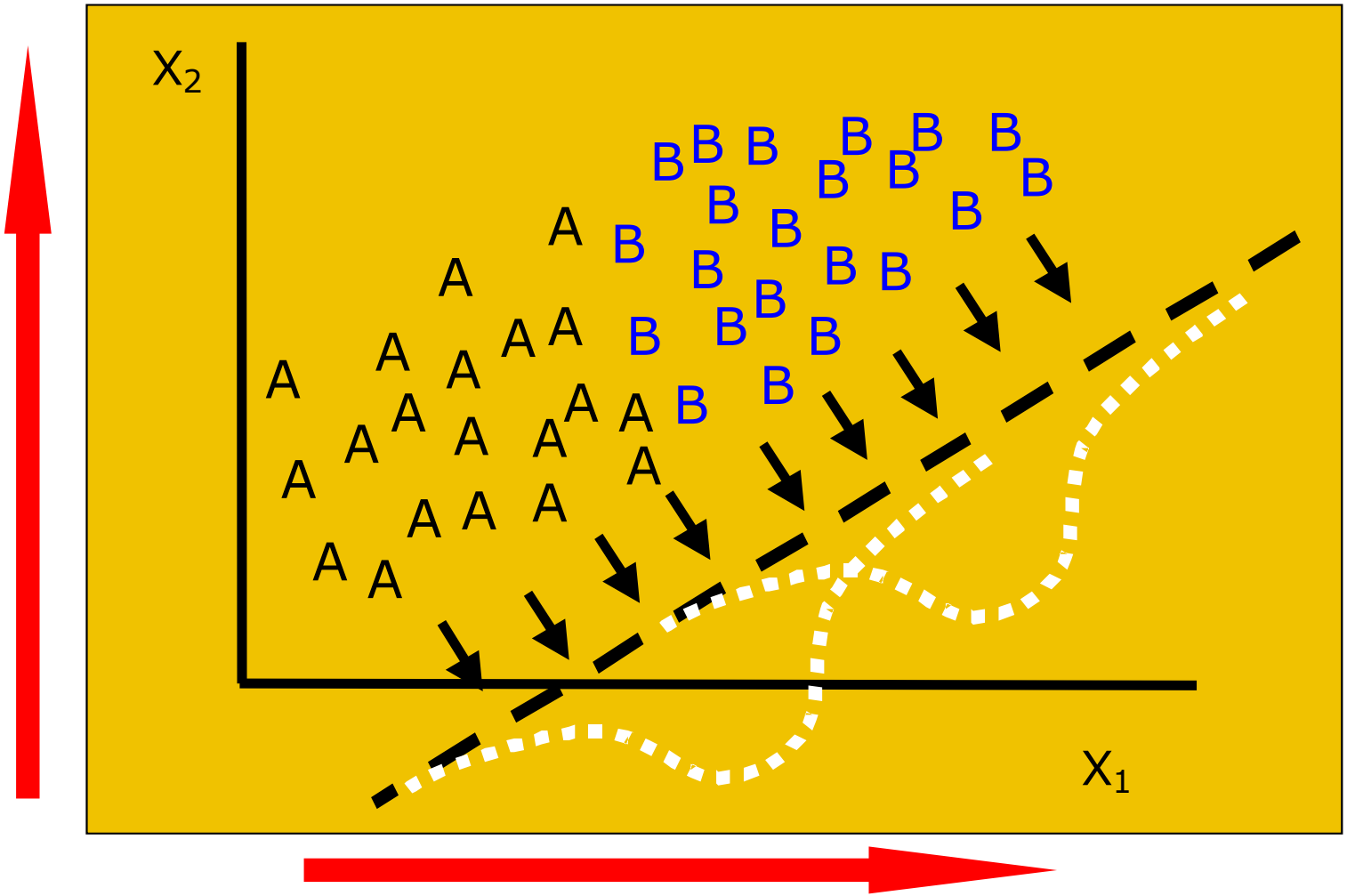
- Individuální predikce stavu nenámých pacientů
- Model musí být podložen korektní statistikou a rozsáhlými daty

Diskriminační analýza

- Cíle diskriminační analýzy
  - Identifikace proměnných (prediktorů) diskriminujících mezi předem danými skupinami objektů
  - Klasifikace objektů do skupin
- Předpoklady diskriminační analýza
  - Obdoba lineární regrese
  - Oddělení objektů podél přímky ve vícerozměrném prostoru (lineární vztah); existuje nicméně kvadratická diskriminační analýza
  - Předpoklad vícerozměrného normálního rozdělení prediktorů v každé ze skupin
  - Citlivá na přítomnost odlehlých hodnot
  - Citlivá na redundantní proměnné v modelu
- Typy diskriminační analýzy
  - Podle typu vztahu
    - Lineární
    - Kvadratická
  - Podle účelu
    - Kanonická diskriminační analýza – identifikace proměnných významných pro diskrminaci
    - Klasifikační diskriminační analýza – klasifikace neznámých objektů do skupin

Princip diskriminační analýzy

- Kombinací několika proměnných získáme nový pohled odlišující existující skupiny objektů, které není možné odlišit žádnou z proměnných samostatně



Kroky diskriminační analýzy

- Metoda lineárního modelování obdobná analýze rozptylu, regresi nebo kanonické korelační analýze (nejsnáze pochopitelná je její analogie k ANOVA)
- Výpočet probíhá v následujících základních krocích:
  - Testování významnosti rozdílů v hodnocených proměnných mezi existujícími skupinami objektů; tato část výpočtu je vlastně MANOVA (multivariate analysis of variance, vícerozměrná ANOVA)
    - Pokud je potvrzena alternativní hypotéza rozdílů mezi skupinami objektů následuje tvorba vlastního modelu
  - Nalezení lineární kombinace proměnných, která nejlépe odlišuje mezi skupinami objektů (diskriminační funkce)

Historie diskriminační analýzy

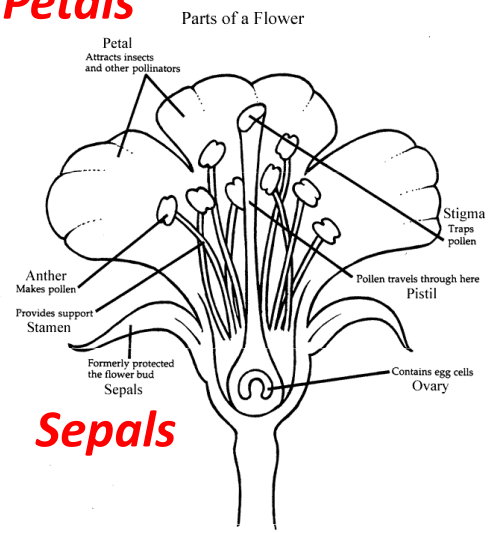
- Popsána pod názvem canonical variate analysis (CVA) Fisherem v roce 1936 pro dvě skupiny; Rao (1948, 1952) ji rozšířil pro více než 2 skupiny
- Je spjata se slavnými „Fisherovými kosatci“ na nichž ji Fisher v roce 1936 popsal
- Fisherovy kosatce
  - Shromážděny na poolostrově Gaspé (Quebec v Kanadě) botanikem Edgarem Andersonem

Versicola

Virginic

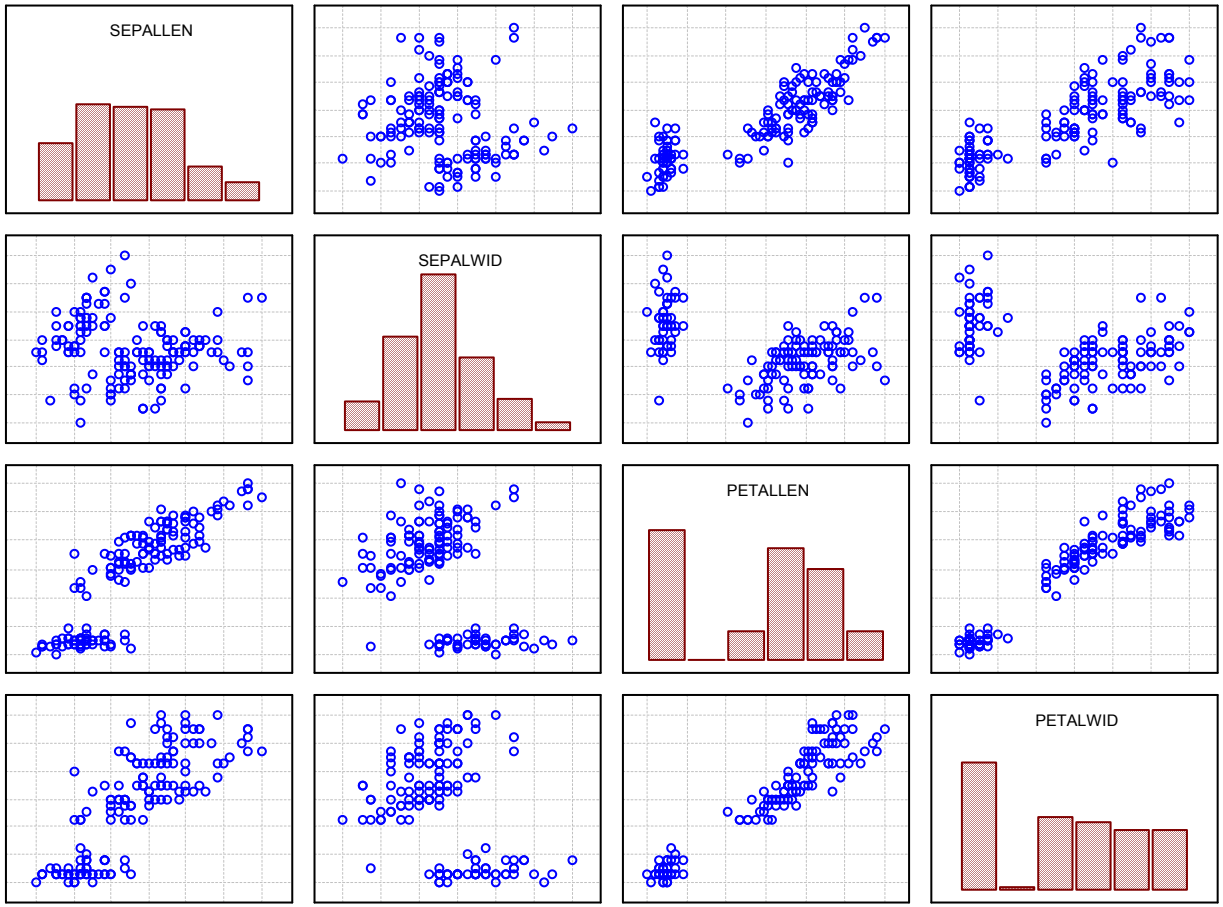
Setosa

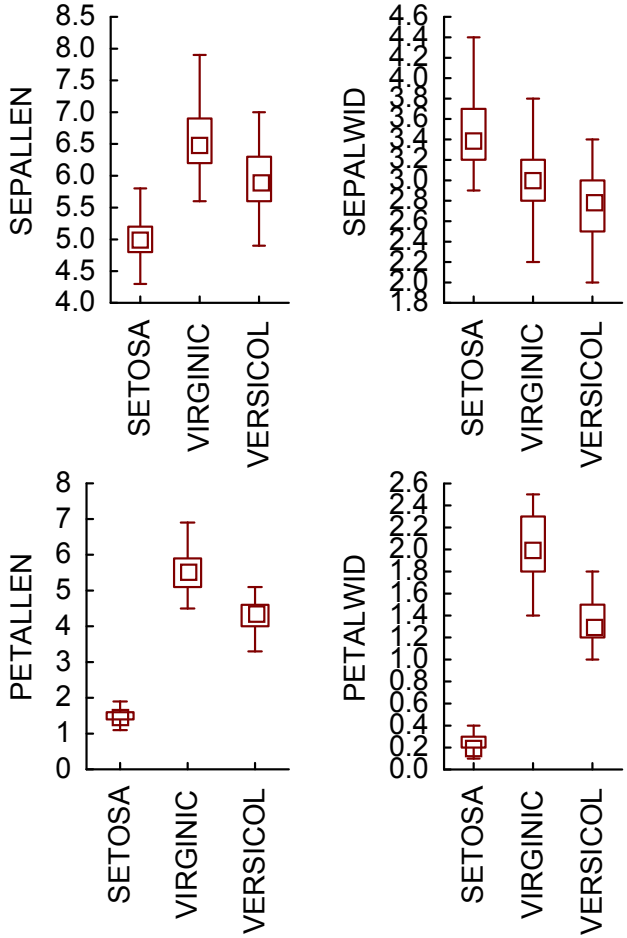
Petals



Sepals

Předstupeň diskriminační analýzy: popis vztahu prediktorů a existujících skupin objektů



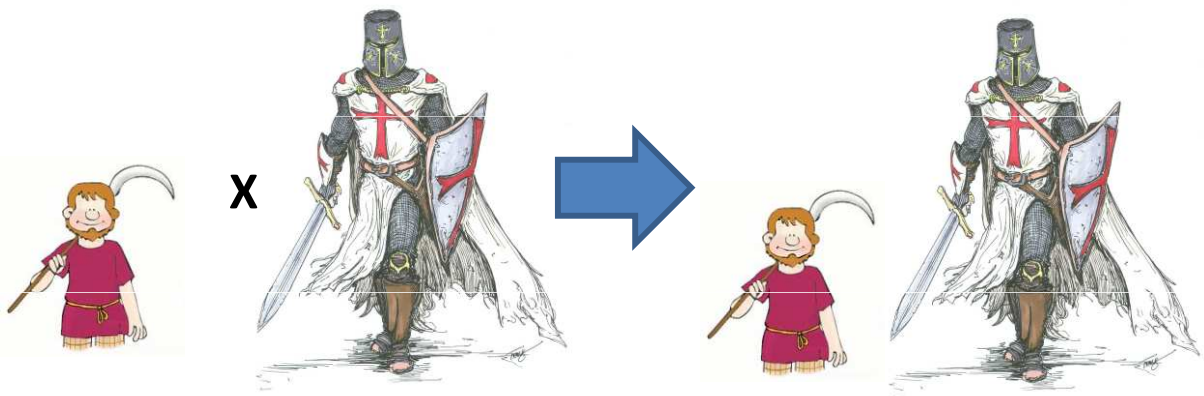


Variable	Analysis of Variance (diskriminacni) Marked effects are significant at p < .05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
SEPALLEN	63.2121	2	31.6061	38.95620	147	0.265008	119.265	0.000000
SEPALWID	11.3449	2	5.6725	16.96200	147	0.115388	49.160	0.000000
PETALLEN	437.1028	2	218.5514	27.22260	147	0.185188	1180.161	0.000000
PETALWID	80.4133	2	40.2067	6.15660	147	0.041882	960.007	0.000000

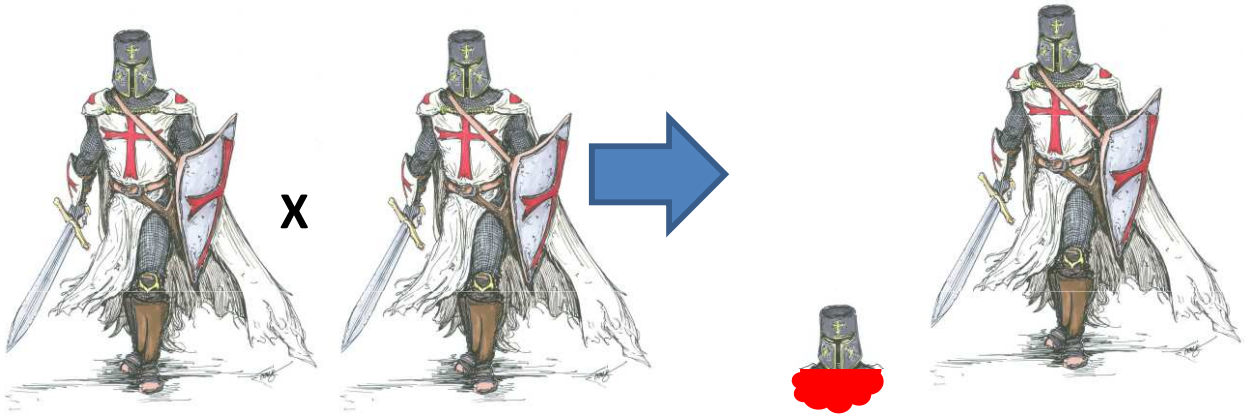
Nicméně pozor na pouze jednorozměrný výběr proměnných – diskriminace objektů může být dána pouze jejich kombinací

Význam identifikace redundantních proměnných

- Redundantní proměnné snižují stabilitu modelu a mohou vést až k nesmyslným výsledkům



Proměnná se silnější diskriminační silou a nekorelovaná s druhou proměnnou snadno vyhrává zařazení do modelu, další proměnné následují dle jejich významu



V případě dvou korelovaných proměnných s obdobnou diskriminační silou pouze jedna vyhrává zařazení do modelu (výsledek dán nepatrnými náhodnými odlišnostmi), druhá je vyřazena nebo vstupuje s do modelu s minimálním významem -> problém s interpretací a stabilitou

Identifikace redundantních proměnných

- Korelační analýza a XY grafy
  - Jednoduchý výpočet
  - Analyzuje vztahy pouze dvojic proměnných

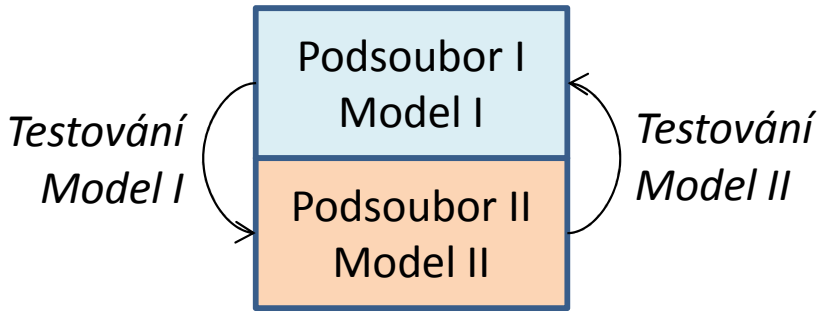
- Analýza hlavních komponent nebo faktorová analýza
  - Analyzuje vzájemné vztahy sady proměnných
  - Usnadňuje výběr neredundantních proměnných nebo nahrazení proměnných faktorovými osami

- Analýza vzájemného vysvětlení proměnných (analýza redundance)
  - Ve statistických software často součást regresní analýzy nebo diskriminační analýzy
  - R^2 a Tolerance – R^2 popisuje kolik variability dané proměnné je vysvětleno ostatními proměnnými v modelu? Tolerance je $1-R^2$, tedy kolik unikátní variability na proměnnou připadá (principem je vícerozměrná regrese, ta determinuje i předpoklady výpočtu)
  - VIF (Variance Inflation Factor) je počítán jako 1/Tolerance, při VIF>10 je kolinearita považována za velmi závažnou (nicméně nejsou dány žádné závazné hranice VIF)

- Expertní znalost proměnných
  - Vyřazovány jsou korelované proměnné s obtížným měřením, zatížené chybami, nízkou vyplněností apod.

Ověření diskriminační funkce na nezávislém souboru

- Při tvorbě modelů může dojít k problému, kdy vytvořený model je perfektně „vycvičen“ řešit danou úlohu na datovém soubor na němž byla vytvořena
- Z tohoto důvodu je problematické testovat výsledky modelu na stejném souboru, na němž byla vytvořena -> jde o důkaz kruhem
- Řešením je testování výsledků modelu na souboru se známým výsledkem (zde známým zařazením objektů do skupin), který se nepodílel na definici modelu
  - Krosvalidace
    - datový soubor je náhodně rozdělen na několik podsouborů (2 nebo více)
    - Na jednom podsouboru je vytvořen model a jeho výsledky testovány na zbývajících podsouborech
    - Výpočet je proveden postupně na všech podsouborech
  - One out leave out
    - Model je vytvořen na celém souboru bez jednoho objektu
    - na tomto objektu je model testován
    - postup je zopakován pro všechny objekty
  - Permutační metody
    - Jackknife, bootstrap – model je postupně vytvářen na náhodných podvýběrech souboru a testován na zbytku dat



Algebra diskriminační analýzy

- Výpočet diskriminační analýzy je možné snadno popsat analogií s ANOVA a PCA
  - ANOVA – definice matice rozptylu jako rozptylu vztaženého k rozdílům mezi skupinami
  - PCA – identifikace faktorových os vysvětlujících maximum rozptylu (zde rozptylu mezi skupinami)

G- počet skupin, n – počet objektů	Suma čtverců	Matice rozptylu
Celkový rozptyl	T	$S = T/n-1$
Sloučený rozptyl uvnitř skupin	$W = W_1 + \ldots + W_g$	$V = W/n-g$
Rozptyl mezi skupinami	$B = T - W$	$A = B/g-1$

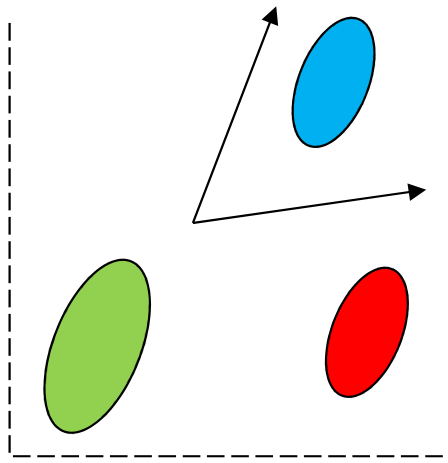
  - Pro rozptyl mezi skupinami pak hledáme pohled maximalizující vysvětlenou variabilitu; v obecném tvaru jde o stejný vzorec jako v případě PCA

$$(V^{-1}A - \lambda_k I)u_k = 0 \longrightarrow (A - \lambda_k V)u_k = 0$$

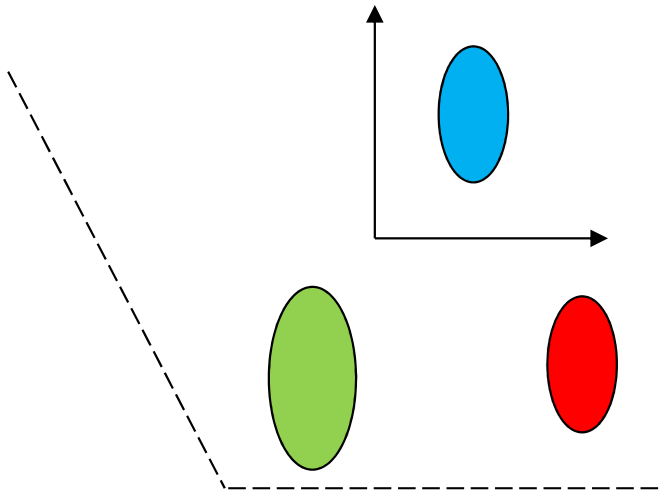
Kde λ jsou eigenvalue a u eigenvektory

  - Počet os definovaných eigenvektory je g-1
  - Eigenvektory jsou různě standardizovány
    - Normalizované eigenvektory $C = U(U'VU)^{-1/2}$ definují tzv. kanonický prostor diskriminační analýzy; transformace vede k maximalizaci variability mezi centroidy skupin a sféricitě rozptylu uvnitř skupin
    - Další metody jsou standardizace na délku 1 nebo druhou odmocninu z eigenvalue; ty nicméně nezaručují sfericitu rozptylu uvnitř skupin

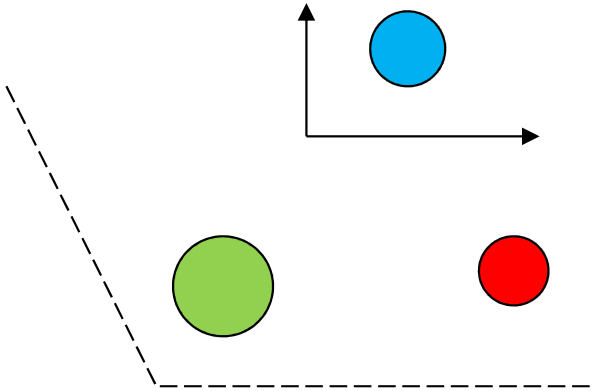
Vztah původních proměnných a kanonických os



Kanonické osy nejsou v prostoru původních proměnných ortogonální



Kanonické osy použité jako ortogonální mění rotaci skupin objektů v prostoru



Normalizace eigenvektorů pomocí $C = U(U'VU)^{-1/2}$ vede ke sféricitě variability (pouze v případě homogenity rozptylu)

Původní proměnné → Kanonické osy

Skupiny objektů

Dle Legendre a Legendre, 1998, Numerical ecology

Pojmy a výstupy diskriminační analýzy

- Popis významu proměnných v modelu
  - Wilks lambda modelu
  - Wilks lambda proměnných
  - Partial lambda
  - Tolerance
- Kanonická analýza
  - Eigenvektory
  - Eigenvalues
- Klasifikace neznámých objektů
  - Diskriminační funkce
  - A priori probability
  - Posterior probability

Wilks lambda

- Měří odlišnost v pozici centroidů skupin definovaných danými proměnnými
- Je počítána jako poměr determinantů matice sumy čtverců a vektorového součinu W a T, kde
  - W je složená matice sumy čtverců uvnitř každé analyzované skupiny (analogie k variabilitě uvnitř skupin v ANOVA)
  - T je matice skalárních produktů centrovaných proměnných pro všechny objekty bez ohledu na to, z jaké skupiny pochází (obdoba celkové variability v ANOVA)

$$\Lambda = \frac{|W|}{|T|}$$

- Tento poměr má rozsah od 0 (maximální rozdíl v pozici centroidů skupin) až 1 (žádný rozdíl mezi centroidy skupin)
- Wilks lambda může být následně převedeno na chi-square nebo F statistiku a statisticky testováno

Popis modelu

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,144)	p-value	Toler.	1-Toler. (R-Sqr.)
	Discriminant Function Analysis Summary (Spreadsheet1)					
	No. of vars in model: 4; Grouping: IRISTYPE (3 grps)					
	Wilks' Lambda: .02344 approx. F (8,288)=199.15 p<0.0000					
SEPALLEN	0.[illegible]	0.[illegible]	[illegible]	0.[illegible]	0.[illegible]	0.[illegible]
SEPALWID	0.[illegible]	0.[illegible]	[illegible]	0.[illegible]	0.[illegible]	0.[illegible]
PETALLEN	0.05302[illegible]	0.669520	35.59014	0.000000	0.36512[illegible]	0.63487[illegible]
PETALWID	0.03154[illegible]	0.74300[illegible]	24.90433	0.000000	0.64931[illegible]	0.35068[illegible]

1. **Celkové Wilks lambda** – na škále 0 (nejlepší diskriminace) až 1 (žádná diskriminace) popisuje celkovou kvalitu modelu všech proměnných
2. **Wilks lambda jednotlivých proměnných** – jde o wilks lambda celého modelu při vyřazení dané proměnné
3. **Partial lambda** – unikátní příspěvek dané proměnné k diskriminaci
4. **F to remove** – F statistika asociovaná s příslušnou partial lambda
5. **P value** – statistická významnost F to remove a tedy i partial lambda
6. **Tolerance** – unikátní variabilita proměnné nevysvětlená ostatními proměnnými v modelu
7. **R^2** –variabilita proměnné vysvětlená kombinací ostatních proměnných v modelu

Mahalanobisova vzdálenost (Mahalanobis 1936)

- Jde o obecné měřítko vzdálenosti beroucí v úvahu korelaci mezi parametry a je nezávislá na rozsahu hodnot parametrů. Počítá vzdálenost mezi objekty v systému souřadnic jehož osy nemusí být na sebe kolmé. V praxi se používá pro zjištění vzdálenosti mezi skupinami objektů. Jsou dány dvě skupiny objektů w1 a w2 o n1 a n2 počtu objektů a popsané p parametry:

$$D_5^2(w_1, w_2) = \overline{d_{12}} V^{-1} \overline{d_{12}}$$

- Kde $\overline{d_{12}}$ je vektor o délce p rozdílů mezi průměry p parametrů v obou skupinách. V je vážená disperzní matice (matice kovariancí parametrů) uvnitř skupin objektů.

$$V = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1 + (n - 2)S_2\right]$$

- kde S1 a S2 jsou disperzní matice jednotlivých skupin. Vektor $\overline{d_{12}}$ měří rozdíl mezi p-rozměrnými průměry skupin a V vkládá do rovnice kovarianci mezi parametry.

Mahalanobisova vzdálenost v diskriminační analýze

- Používána pro popis vzájemných vzdáleností centroidů skupin
- Používána pro popis vzdáleností objektů od centroidů skupin a následně pro výpočet posterior probabilities zařazení objektů do skupin

	Squared Mahalanobis Distances (Spreadshe		
IRISTYPE	SETOSA	VERSICOL	VIRGINIC
SETOSA	0.000(	89.8641	179.384
VERSICOL	89.864	0.0000(	17.201
VIRGINIC	179.384	17 10	0.000(

Vzdálenosti centroidů

Vzdálenosti objektů od centroidů

	Squared Mahalanobis Distances from Group Centroids (Spreadsh Incorrect classifications are marked with *			
Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
1	SETOSA	0.241	90.660	181.558
	VIRGINIC	208.571	27.318	1.894
	VERSICOL	105.266	2.232	13.072
	VIRGINIC	207.918	31.749	4.450
*	VIRGINIC	133.066	5.252	7.235
	SETOSA	1.333	84.011	170.056
	VIRGINIC	173.183	26.562	11.048
	VERSICOL	131.661	8.430	14.764
*	VERSICOL	130.862	8.669	6.506
	SETOSA	2.286	113.650	210.023
	VERSICOL	99.233	1.296	13.817
*	VERSICOL	149.030	8.439	4.864
	VIRGINIC	158.981	12.751	1.234
	VERSICOL	79.107	1.407	26.653
	VIRGINIC	161.852	12.170	1.978
	VIRGINIC	174.081	16.052	2.390
	VIRGINIC	209.029	29.514	1.939
	SETOSA	2.769(	67.471	145.700

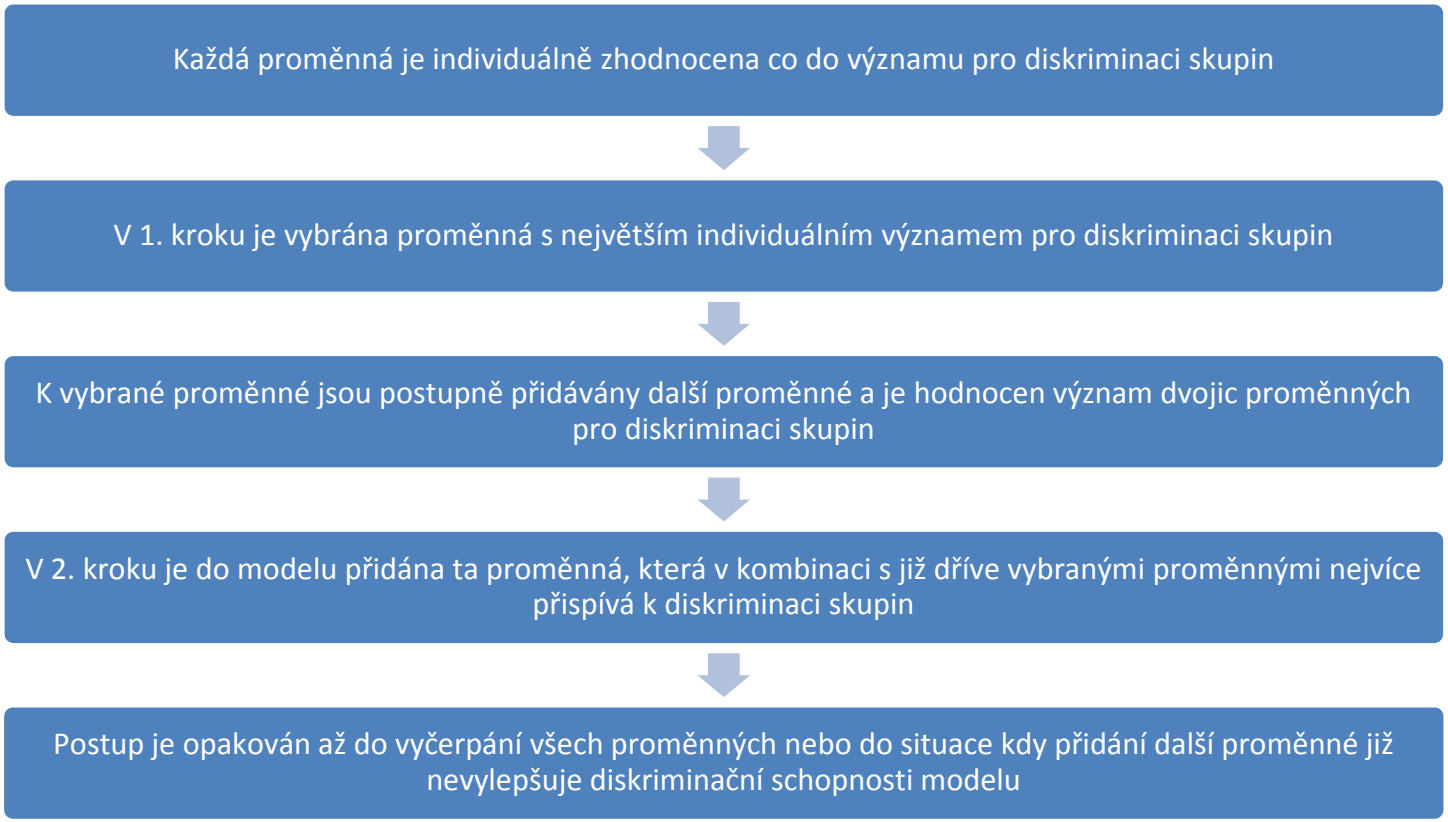
Dopředná a zpětná eliminace

- Dopředná a zpětná eliminace proměnných z modelu (forward, backward stepwise) je obecná technika používaná při tvorbě regresních, diskriminačních a jiných modelů
- Proměnné jsou do modelu postupně přidávány (ubírány) podle jejich významu v modelu

Schéma dopředné eliminace proměnných v modelu

V případě zpětné eliminace začíná proces od modelu se všemi proměnnými a postupně jsou vyřazovány proměnné s nejmenším příspěvkem k diskriminační síle modelu

Proces je třeba expertně kontrolovat, riziková je např. přítomnost redundantních proměnných



Definice modelu prostřednictvím stepwise analýzy

- Před zahájením výpočtu je třeba nastavit Toleranci přidání proměnné (=hodnota při které nebude proměnná do modelu zařazena z důvodu redundance), F to enter a F to remove jsou hodnoty F spjaté s danou proměnnou, při které je daná proměnná zařazena/ vyřazena z modelu

Forward stepwise

Discriminant Function Analysis Summary (Spreadsheet1)
Step 1, N of vars in model: 1; Grouping: IRISTYPE (3 grp
Wilks' Lambda: .05863 approx. F (2,147)=1180.2 p<0.00(

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,147)	p-value	Toler.	1-Toler (R-Sqr.)
PETALLEN	1.00000	0.05862	1180.16	0.0(	1.00000	0.0(

Discriminant Function Analysis Summary (Spreadsheet1)
Step 2, N of vars in model: 2; Grouping: IRISTYPE (3 grps)
Wilks' Lambda: .03688 approx. F (4,292)=307.10 p<0.0000

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,146)	p-value	Toler.	1-Toler. (R-Sqr.)
PETALLEN	0.59921	0.06155	1112.95	0.00000	0.85717	0.14282
SEPALWID	0.05862	0.62911	43.035	0.00000	0.85717	0.14282

Discriminant Function Analysis Summary (Spreadsheet1)
Step 3, N of vars in model: 3; Grouping: IRISTYPE (3 grps)
Wilks' Lambda: .02498 approx. F (6,290)=257.50 p<0.0000

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,145)	p-value	Toler.	1-Toler. (R-Sqr.)
PETALLEN	0.03831	0.65183	38.7244	0.00000	0.73641	0.26358
SEPALWID	0.04377	0.57052	54.5769	0.00000	0.74921	0.25078
PETALWID	0.03688	0.67713	34.5686	0.00000	0.66890	0.33109

Discriminant Function Analysis Summary (Spreadsheet1)
Step 4, N of vars in model: 4; Grouping: IRISTYPE (3 grps)
Wilks' Lambda: .02344 approx. F (8,288)=199.15 p<0.0000

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,144)	p-value	Toler.	1-Toler. (R-Sqr.)
PETALLEN	0.03502	0.66920	35.5901	0.00000	0.36512	0.63487
SEPALWID	0.03058	0.76648	21.9359	0.00000	0.60885	0.39114
PETALWID	0.03154	0.74300	24.9043	0.00000	0.64931	0.35068
SEPALLEN	0.02497	0.93846	4.7211	0.01032	0.34799	0.65200

Backward stepwise

Discriminant Function Analysis Summary (Spreadsheet1)
Step 0, N of vars in model: 4; Grouping: IRISTYPE (3 grps)
Wilks' Lambda: .02344 approx. F (8,288)=199.15 p<0.0000

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,144)	p-value	Toler.	1-Toler. (R-Sqr.)
SEPALLEN	0.02497	0.93846	4.7211	0.01032	0.34799	0.65200
SEPALWID	0.03058	0.76648	21.9359	0.00000	0.60885	0.39114
PETALLEN	0.03502	0.66920	35.5901	0.00000	0.36512	0.63487
PETALWID	0.03154	0.74300	24.9043	0.00000	0.64931	0.35068

Discriminant Function Analysis Summary (Spreadsheet1)
Step 1, N of vars in model: 3; Grouping: IRISTYPE (3 grps)
Wilks' Lambda: .02498 approx. F (6,290)=257.50 p<0.0000

N=150	Wilks' Lambda	Partial Lambda	F-remove (2,145)	p-value	Toler.	1-Toler. (R-Sqr.)
SEPALWID	0.04377	0.57052	54.5769	0.00000	0.74921	0.25078
PETALLEN	0.03831	0.65183	38.7244	0.00000	0.73641	0.26358
PETALWID	0.03688	0.67713	34.5686	0.00000	0.66890	0.33109

Kanonická analýza

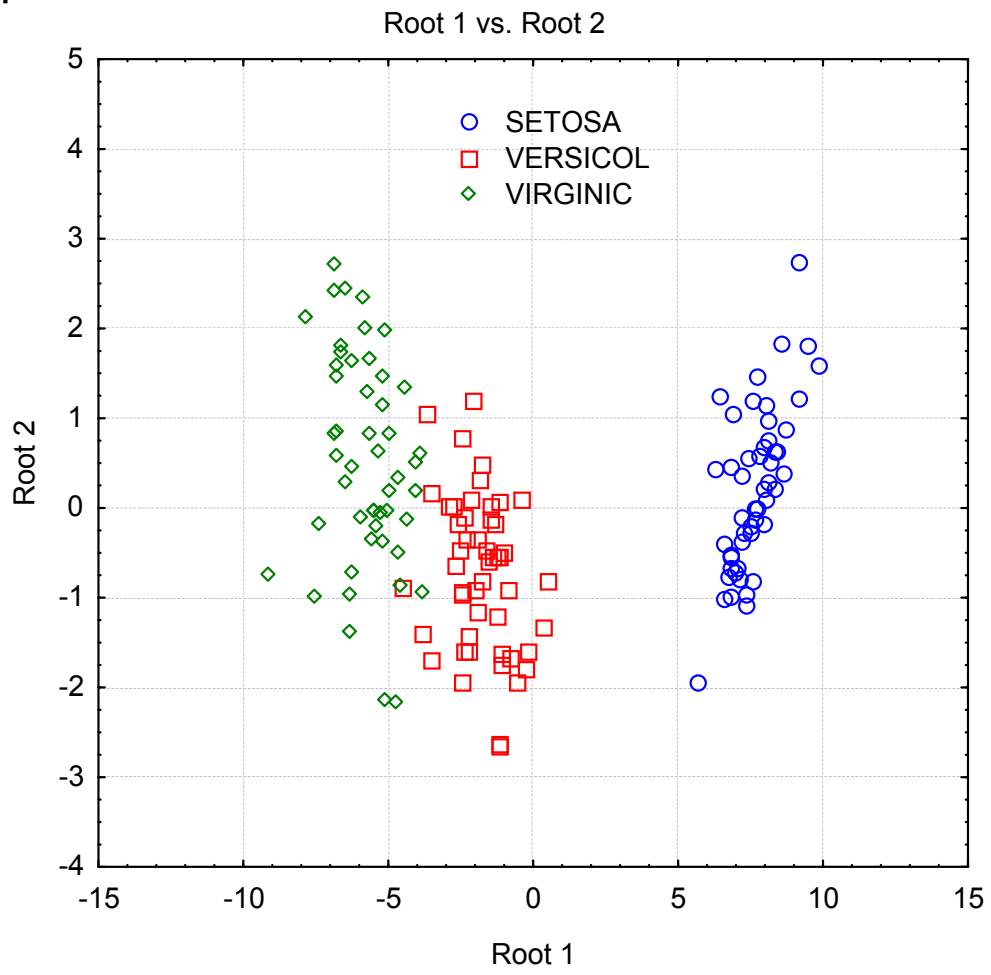
- Analogická k výpočtu analýzy hlavních komponent, liší se významem vytvořených os (kanonických kořenů; eigenvektorů)
- Na rozdíl od PCA, kde význam osy je spjat s vyčerpanou variabilitou dat u diskriminační analýzy je význam os určen následovně:
  - 1. osa – největší diskriminace mezi centroidy skupin objektů
  - 2. osa – druhá největší diskriminace mezi centroidy skupin objektů
  - Atd.
- Počet kanonických kořenů je dán jako počet skupin objektů-1
- Na rozdíl od PCA nemusí být kanonické kořeny ortogonální

Kanonická analýza - výsledky

- Eigenvektory popisují příspěvek jednotlivých proměnných k definici kanonických kořenů
- Eigenvalues popisují variabilitu spjatou s kanonickými osami (tedy s rozdílem mezi centroidy skupin)

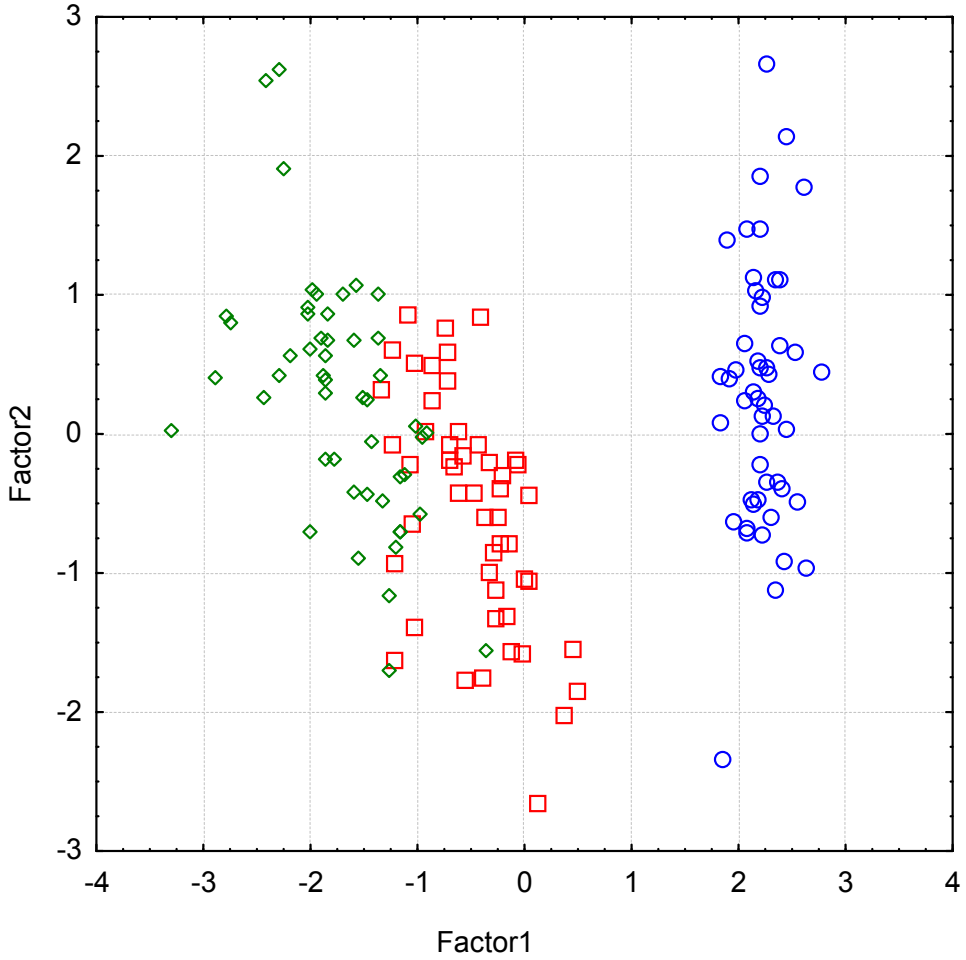
	Chi-Square Tests with Successive Roots Removed (Spreadshe					
Roots Removed	Eigen-value	Canonical R	Wilks' Lambda	Chi-Sqr.	df	p-value
0	32.1919	0.98482	0.02343	546.115	8	0.00000
1	0.2853	0.47119	0.77797	36.529	3	0.00000

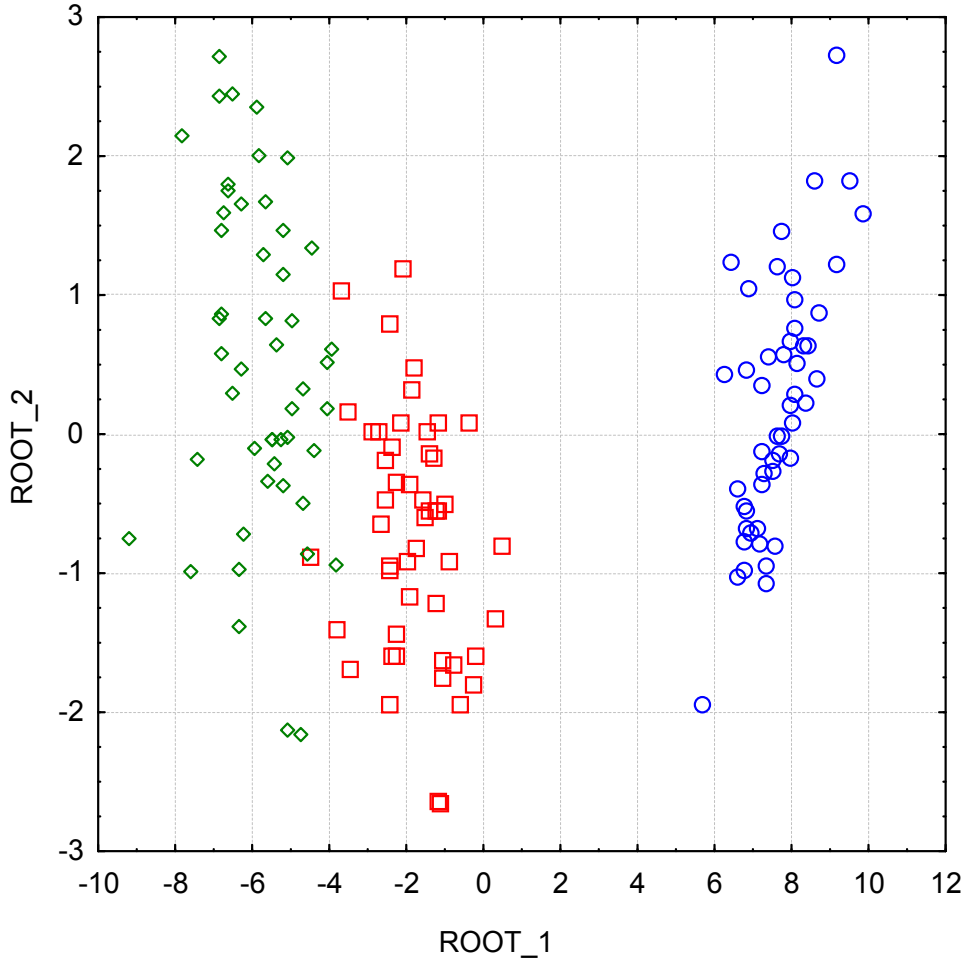
	Standardized Coefficients for Canonical Variables	
Variable	Root 1	Root 2
SEPALLEN	0.4269	0.01240
SEPALWID	0.5212	0.73526
PETALLEN	-0.9472	-0.40103
PETALWID	-0.5751	0.58104
Eigenval	32.1919	0.28539
Cum.Prop	0.9912	1.00000



PCA vs. Diskriminační analýza

- Maximální vyčerpaná variabilita (PCA) vs. Maximální diskriminace (DA)





Klasifikace neznámých objektů pomocí diskriminační analýzy

- Využívá tzv. klasifikační funkci
- Jde o sadu rovnic (pro každou skupinu jedna rovnice)
- Objekt je zařazen do skupiny, jejíž klasifikační funkce nabývá nejvyšší hodnoty
- V kombinaci s apriori a posterior probabilities je určena finální pravděpodobnost zařazení objektu do skupiny

Variable	Classification Functions; grouping: IRISTYPE (Spreadsh		
	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
SEPALLEN	23.5442	15.6982	12.446
SEPALWID	23.5879	7.0725	3.685
PETALLEN	-16.4306	5.2115	12.767
PETALWID	-17.3984	6.4342	21.079
Constant	-86.3085	-72.8526	-104.368

$$SETOSA = 23.5 * SEPALLEN + 23.6 * SEPALWID + \ldots atd.$$

Group	Classification Matrix (Spreadsheet1) Rows: Observed classifications Columns: Predicted classifications			
	Percent Correct	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
SETOSA	100.000	50	0	0
VERSICOL	96.000	0	48	2
VIRGINIC	98.000	0	1	49
Total	98.000	50	49	51

A priori a posterior probabilities

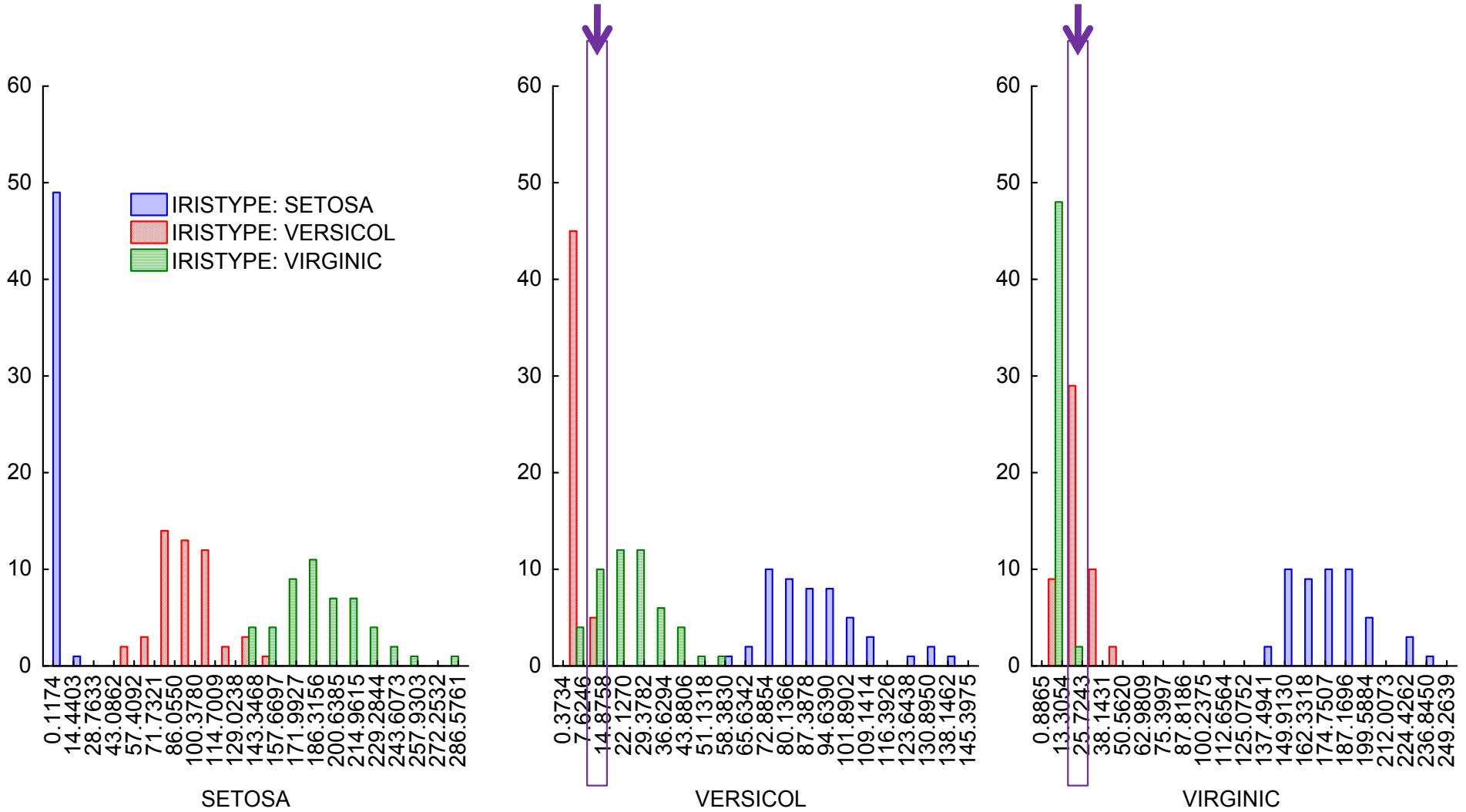
- A priori probabilities – přirozeně daná pravděpodobnost výskytu skupiny objektů
  - Proporcionální – předpokládáme, že struktura souboru odpovídá realitě a tedy i poměr skupin objektů v souboru odpovídá realitě
  - Rovnoměrná – každá skupina má pravděpodobnost dánu jako 100 / počet skupin
  - Uživatelské – dány expertní znalostí a nastaveny analytikem

- Posterior probabilities
  - Vznikají jako kombinace apriori pravděpodobností a Mahalanobisových vzdáleností objektu od centroidů skupin

	Classification of Cases (Spreadsheet1) Incorrect classifications are marked with *					Squared Mahalanobis Distances from Group Incorrect classifications are marked with *					Posterior Probabilities (Spreadsheet1) Incorrect classifications are marked with *				
Case	Observed Classif.	1 p=.33333	2 p=.33333	3 p=.33333	Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333	Case	Observed Classif.	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
1	SETOSA	SETOSA	VERSICOL	VIRGINIC	1	SETOSA	0.241	90.660	181.558	1	SETOSA	1.00000	0.00000	0.00000
	VIRGINIC	VIRGINIC	VERSICOL	SETOSA		VIRGINIC	208.571	27.318	1.894		VIRGINIC	0.00000	0.00000	0.99999
	VERSICOL	VERSICOL	VIRGINIC	SETOSA		VERSICOL	105.266	2.232	13.072		VERSICOL	0.00000	0.99559	0.00441
	VIRGINIC	VIRGINIC	VERSICOL	SETOSA		VIRGINIC	207.918	31.749	4.450		VIRGINIC	0.00000	0.00000	0.99999
*	VIRGINIC	VERSICOL	VIRGINIC	SETOSA	*	VIRGINIC	133.066	5.252	7.235	*	VIRGINIC	0.00000	0.72938	0.27061
	SETOSA	SETOSA	VERSICOL	VIRGINIC		SETOSA	1.333	84.011	170.056		SETOSA	1.00000	0.00000	0.00000
	VIRGINIC	VIRGINIC	VERSICOL	SETOSA		VIRGINIC	173.183	26.562	11.048		VIRGINIC	0.00000	0.00042	0.99957
	VERSICOL	VERSICOL	VIRGINIC	SETOSA		VERSICOL	131.661	8.430	14.764		VERSICOL	0.00000	0.95957	0.04042
*	VERSICOL	VIRGINIC	VERSICOL	SETOSA	*	VERSICOL	130.862	8.669	6.506	*	VERSICOL	0.00000	0.25322	0.74677

Klasifikace objektů dle vzdálenosti

Inconclusive area – nejednoznačné zařazení, nízké p vzhledem ke všem skupinám



Mahalanobisova vzdálenost od daného centroidu

Celkové vyhodnocení výsledků diskriminační analýzy

- Popis výsledků klasifikace vůči známému zařazení objektů do skupin
- Pro validní výsledky a hodnocení kvality modelu by mělo být provedeno na souboru, který se nepodílel na definici modelu (viz. crossvalidace apod.)
- Kromě vlastní klasifikační funkce a Mahalanobisových vzdáleností ovlivňuje zařazení objektů do skupin i apriori pravděpodobnost zařazení

Classification Matrix (Spreadsheet1)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	SETOSA p=.33333	VERSICOL p=.33333	VIRGINIC p=.33333
SETOSA	100.000	50	0	0
VERSICOL	96.000	0	48	2
VIRGINIC	98.000	0	1	49
Total	98.000	50	49	51

Classification Matrix (Spreadsheet1)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	SETOSA p=.20000	VERSICOL p=.70000	VIRGINIC p=.10000
SETOSA	100.000	50	0	0
VERSICOL	100.000	0	50	0
VIRGINIC	90.000	0	5	45
Total	96.666	50	55	45

Výsledky při různé apriori pravděpodobnosti

Diskriminační analýza - shrnutí

- Cílem analýzy je:
  - Identifikace proměnných odlišujících vícerozměrně skupiny objektů
  - Vytvoření modelu pro klasifikaci neznámých objektů
- Omezení analýzy
  - Vícerozměrné normální rozdělení v každé skupině
  - Pozor na odlehlé hodnoty
  - Pozor na redundantní proměnné
  - Rovnice modelu je v základní verzi lineární a tedy i hodnocený problém musí mít lineární řešení
  - Testování modelu provádět na souboru, který se nepodílel na definici modelu
- Výstupy
  - Klasifikační funkce pro zařazení objektů do skupin
  - Pravděpodobnost zařazení jednotlivých objektů do skupin - > interpretace

Ordinační analýzy: shrnutí

- Analýza hlavních komponent, faktorová analýza, korespondenční analýza, multidimensional scaling i diskriminační analýza se snaží zjednodušit vícerozměrnou strukturu dat výpočtem souhrnných os
- Metody se liší v logice tvorby těchto os
  - Maximální variabilita (analýza hlavních komponent, korespondenční analýza)
  - Maximální interpretovatelnost os (faktorová analýza)
  - Převod asociační matice do Euklidovského prostoru (multidimensional scaling)
  - Odlišení existujících skupin (diskriminační analýza)