Diskriminační analýza (DA)

testování hypotéz

(a) interpretace rozdílů - kanonická diskriminační analýza
(aa) zda a do jaké míry je možné odlišit stanovené skupiny objektů na základě znaků, které máme k dispozici,
(ab) které ze znaků k tomuto odlišení přispívají největší mírou.

(b) identifikace objektů - klasifikační diskriminační analýza
odvození jedné nebo více rovnic za účelem identifikace objektů

Požadavky na data:

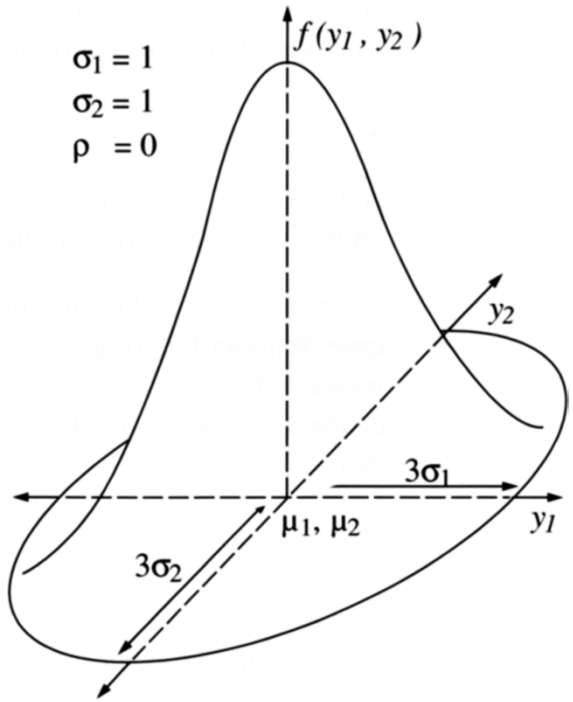
(a) kvantitativní anebo binárními znaky

(b) žádný ze znaků nesmí být lineární kombinací jiného znaku nebo jiných znaků

(c) nelze současně používat dva nebo více velmi silně korelovaných znaků

(d) kovarianční matice pro jednotlivé skupiny musí být přibližně shodné

(e) znaky charakterizující každou skupinu by měly splňovat požadavek mnohorozměrného normálního rozdělení



Pro počty skupin (g), počty znaků (p), počty objektů v skupinách a celkové počty objektů v analýze (n) v diskriminačních analýzách musí platit:

(a) musí být alespoň dvě skupiny objektů: $g \geq 2$;

(b) v každé ze skupin musí být nejméně 2 objekty;

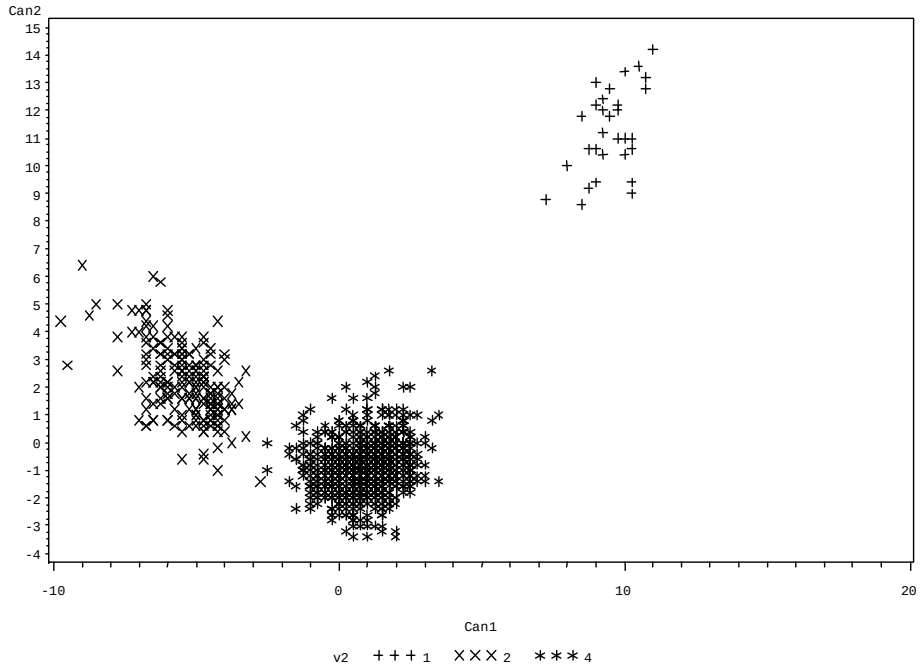
(c) počet znaků použitých v analýze musí být menší než počet objektů zmenšený o počet skupin: $0 < p < (n - g)$;

(d) žádný znak by neměl být v některé skupině konstantní

Kanonická diskriminační analýza (CDA – *canonical discriminant analysis*, *canonical variates analysis*)

umožňuje sledovat vztahy mezi objekty v prostoru definovaném kanonickými osami

ordinační procedura, která maximalizuje rozdíly mezi skupinami



Kanonická diskriminační analýza (CDA – *canonical discriminant analysis*, *canonical variates analysis*)

kanonická diskriminační funkce

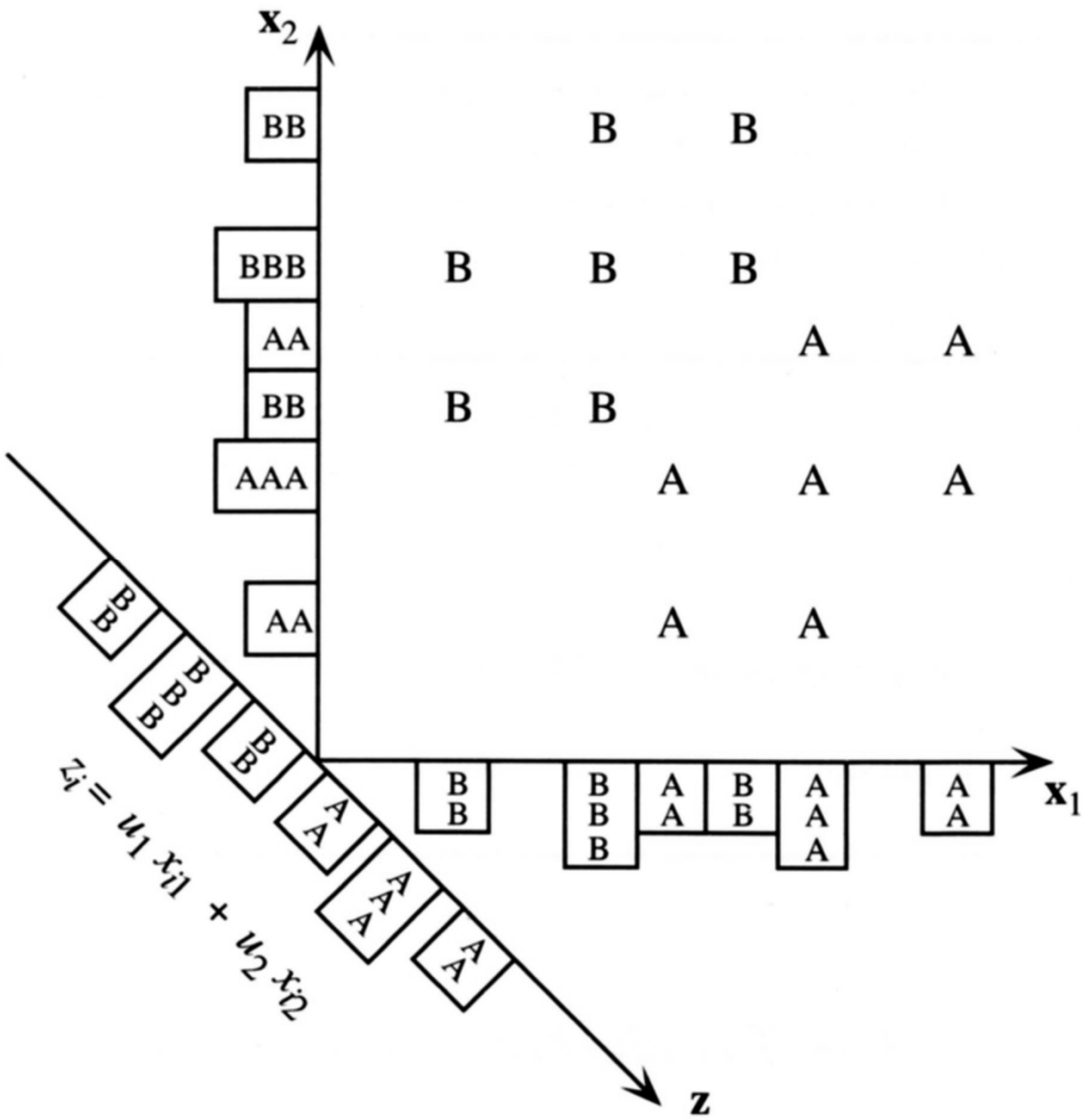
$$f_{km} = a_0 + a_1 x_{1km} + a_2 x_{2km} + \ldots + a_p x_{pkm},$$

f_{km} = hodnota (skóre) kanonické diskriminační funkce pro případ m v skupině k;

x_{ikm} = hodnota diskriminačního znaku x_i pro případ m v skupině k

a_i = koeficienty diskriminační funkce (i = 0, 1 ..., p);

Koeficienty (a) pro první funkci se odvodí tak, aby skupinové těžiště (centroidy, průměry) byly maximálně vzdálené (ve smyslu Mahalanobisovy vzdálenosti). Koeficienty vypočtené pro druhou funkci musí dále maximalizovat rozdíly mezi skupinovými centroidy a současně hodnoty obou funkcí nesmí být korelovány.


x2
BB
BBB
AA
BB
AAA
AA
B B
B B
B B
B B B
A A
B B
A A A
A A
z
x1
zi = u1 xi1 + u2 xi2

	PCA, PCoA, NMDS	DA
Předem stanovené skupiny	ne	ano
Vysvětlení maximální variability	celkové	meziskupinové
Vážení znaků	ne	ano

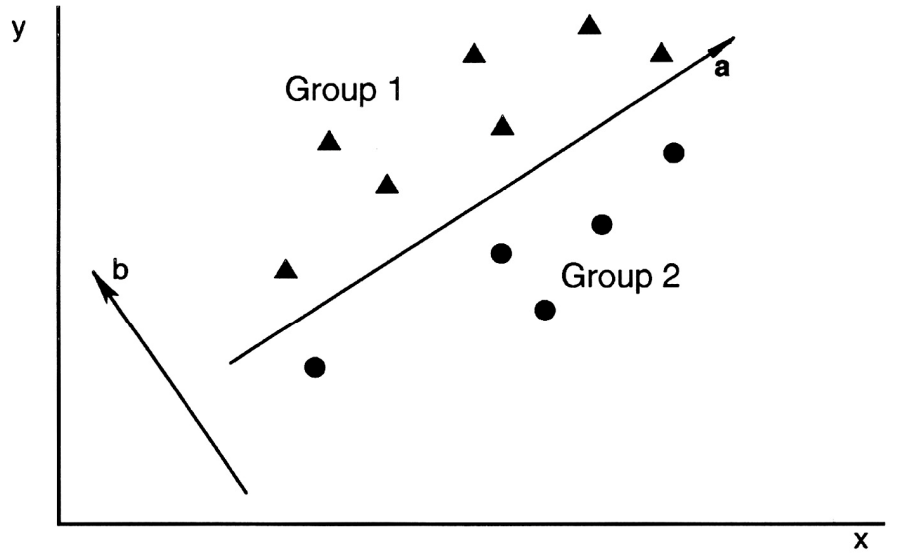


Figure 7.22. Comparison of the underlying ideas in PCA and CVA by an artificial example with two original dimensions. Component 1 (**a**) coincides with the main trend of variation in the entire sample, whereas canonical variate 1 (**b**, there is only one in this case) explains the optimum separation of the two groups.

Koeficienty „*a*“ diskriminační funkce

nestandardizované koeficienty diskriminační funkce (*unstandardized coefficients*)

neadjustované - *raw coefficients*

adjustované

Adjustované koeficienty jsou upravené tak, aby se počátek diskriminační funkce (tj. místo, kde mají všechny kanonické osy nulové hodnoty) nacházel v místě hlavního těžiště (centroidu) (*grand centroid*, tj. v místě průměrné hodnoty všech znaků). POZOR: SAS adjustované koeficienty uvádí jako raw coefficients!

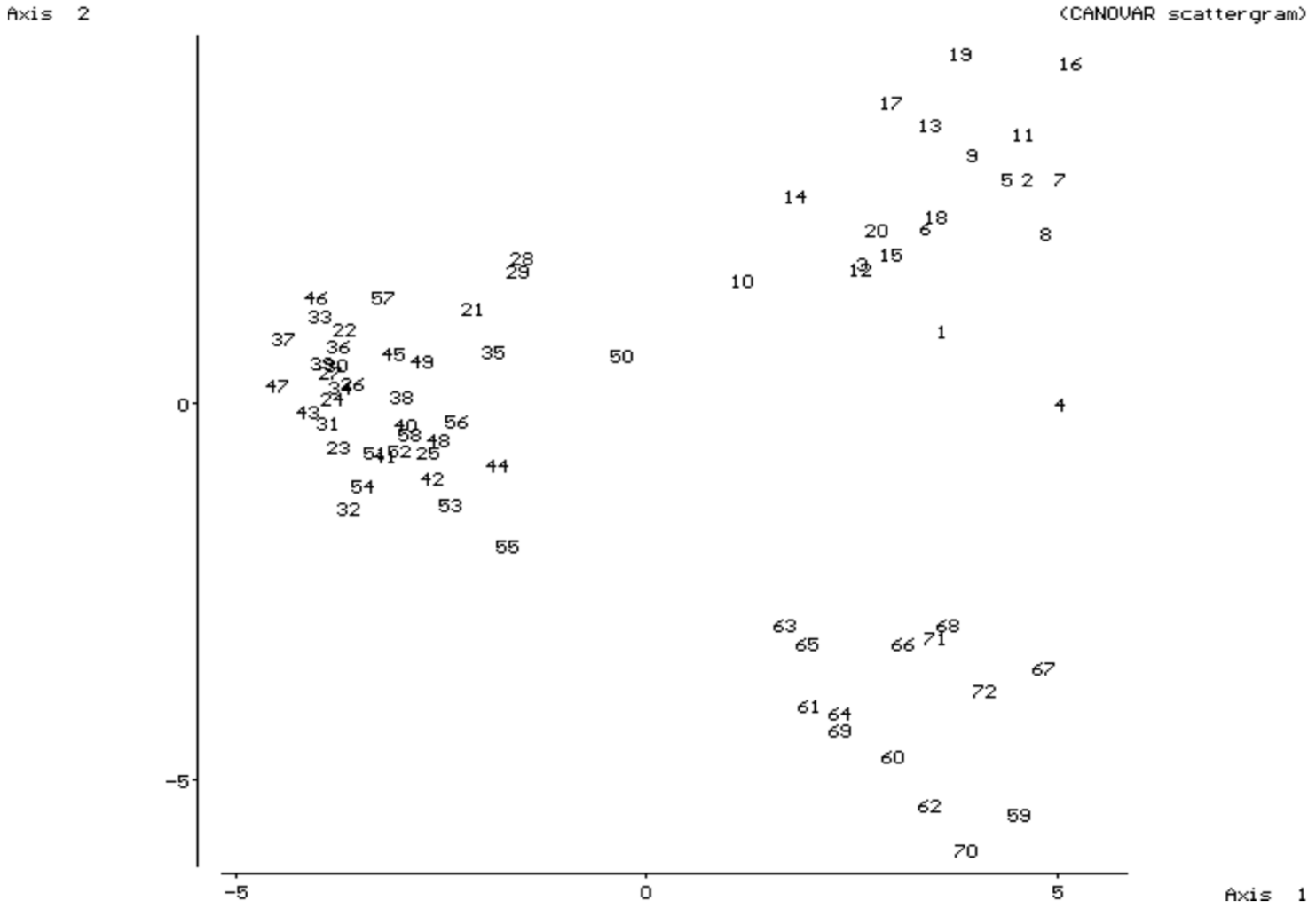
standardizované koeficienty (*standardized coefficients*)

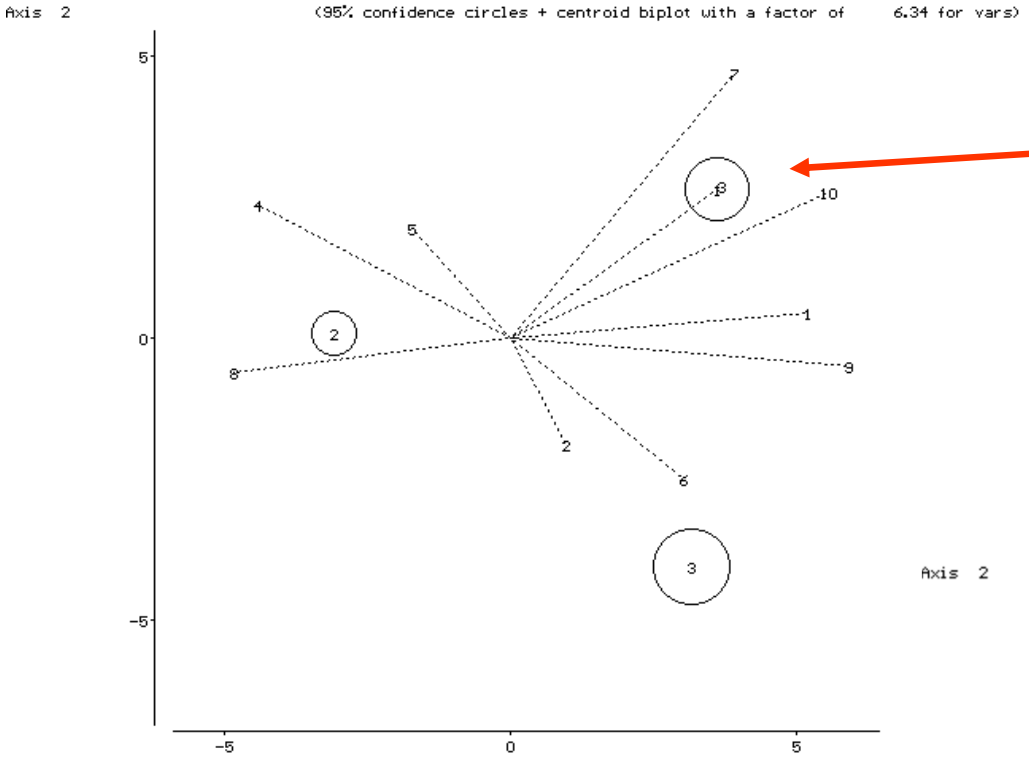
maximální počet diskriminačních funkcí (maximální počet os, maximální počet nenulových vlastních čísel)

$s = \min(p; g - 1)$

Interpretace kanonických os (kanonických diskriminačních funkcí)

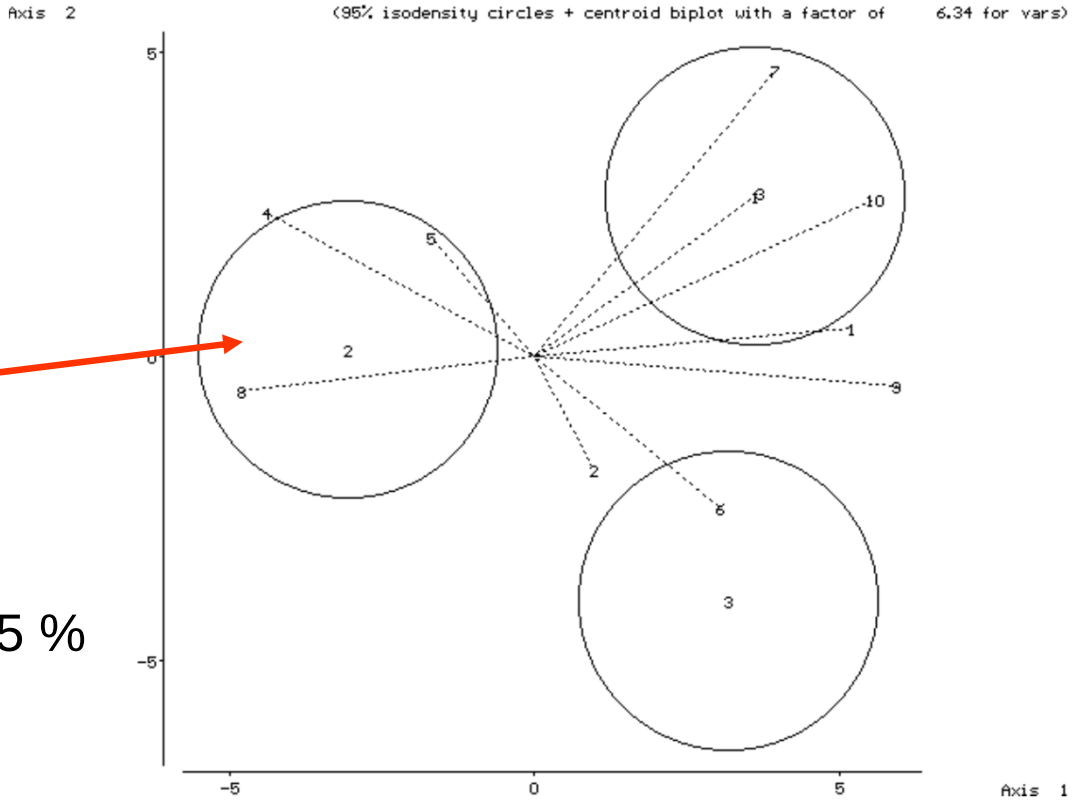
(a) relativní pozice objektů, pozice těžišť (centroidů)





95 % konfidenční interval

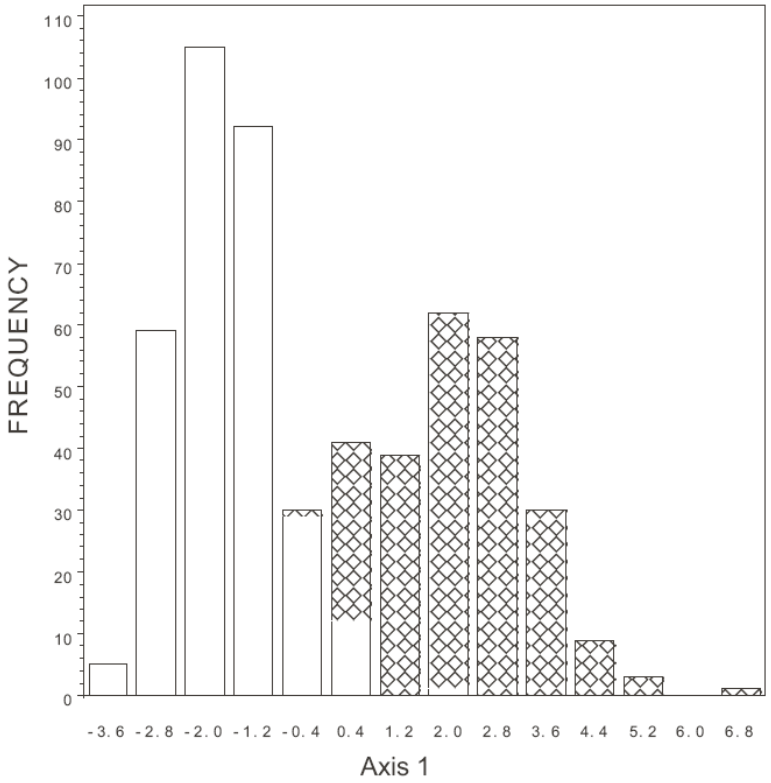
těžiště (centroidu)

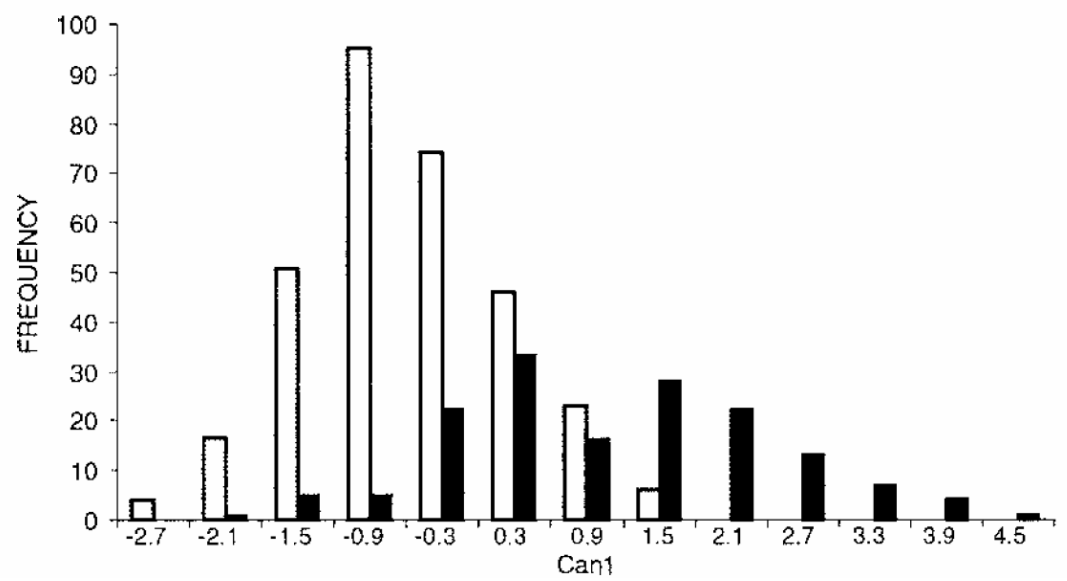


plochy, ve kterých by se mělo nacházet (za předpokladu normálního rozdělení znaků) 95 % objektů dané skupiny

Interpretace kanonických os (kanonických diskriminačních funkcí)

(a) relativní pozice objektů





(b) celkové strukturní koeficienty nebo celková kanonická struktura (*total structure coefficients*, *total canonical structure*)

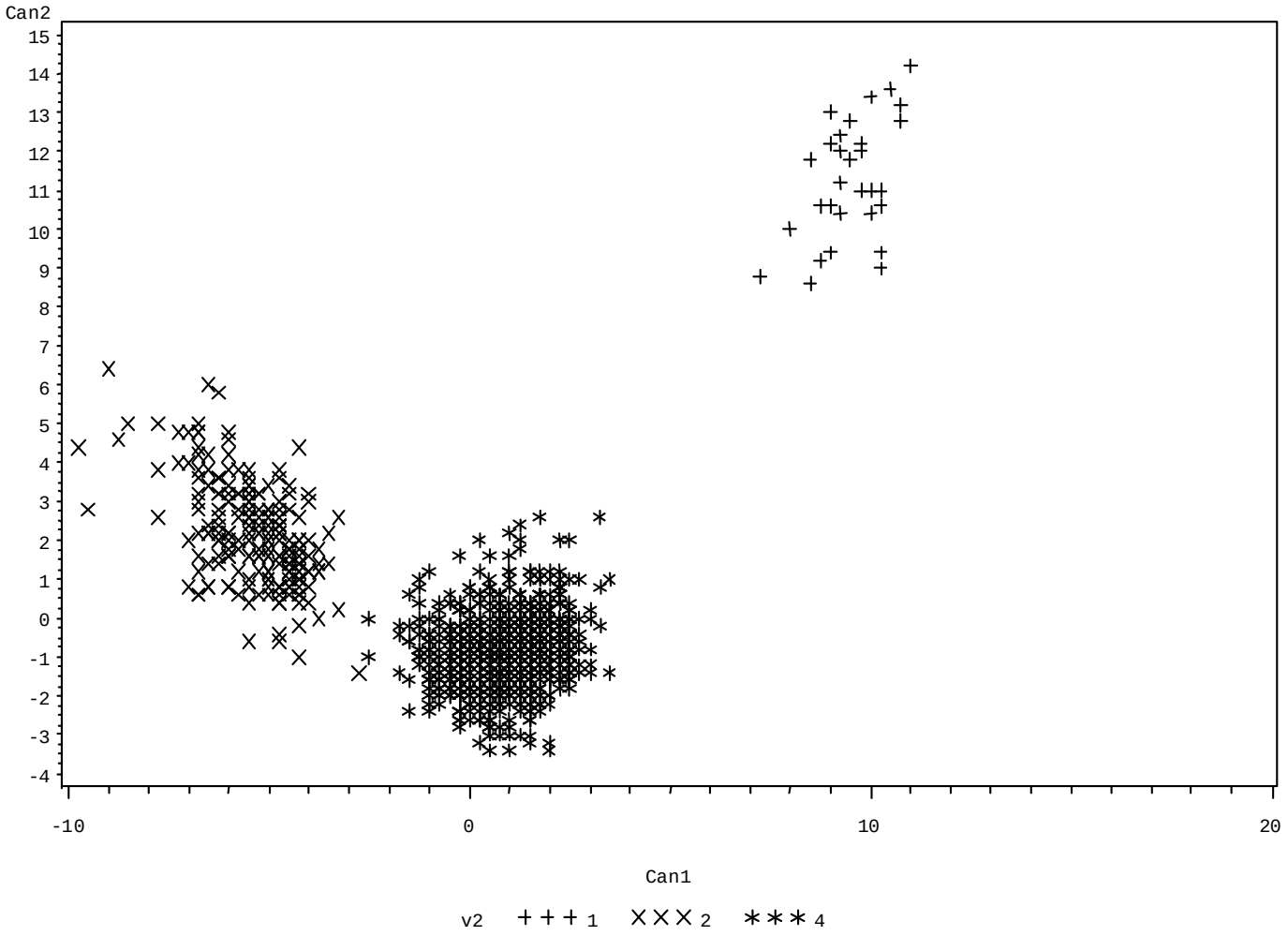
(c) vlastními čísly (*eigenvalues*)

(d) kanonické korelační koeficienty (*canonical correlation coefficients*)

Druhou mocninu kanonických korelačních koeficientů je možné interpretovat jako podíl variability diskriminační funkce, která je vysvětlena skupinami, resp. rozdíly mezi skupinami. Tato charakteristika může být někdy užitečnější než procentuální vyjádření vlastních čísel. Pokud se totiž skupiny liší v analyzovaných znacích jen málo, hodnoty kanonických korelačních koeficientů budou nízké

(e) statistické významnosti diskriminačních funkcí (os)

je kritérium Wilks' lambda, chí kvadrát, případně poměr věrohodnosti (*likelihood ratio*)



diploidní *O. macrocarpon* +

diploidní *O. microcarpus* x

polyploidní *O. palustris* *

Oxycoccus - kanonicka diskriminacni analyza

The CANDISC Procedure

	Canonical correlation	Adjusted canonical correlation	Approximate standard error	Squared canonical correlation
1	0.942491	0.940682	0.003217	0.888290
2	0.905916	0.903497	0.005164	0.820683

Test of H0: The canonical correlations in the current row and all that follow are zero

Values of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	7.9518	3.3750	0.6347	0.6347	0.02003149	202.76	70	2340	<.0001
2	4.5767		0.3653	1.0000	.17931699	157.63	34	1171	<.0001

The CANDISC Procedure

Total Canonical Structure

Variable	Can1	Can2
v4	0.672599	0.282658
v5	0.712117	0.099797
v6	0.683018	0.232456
v7	0.693296	0.143058
v8	**0.814472**	0.054974
v9	0.542881	-0.217714
v10	0.266363	-0.256140
v11	0.361661	-0.180027
v12	**0.830531**	-0.094723
v13	-0.126190	-0.090211
v14	0.663007	0.090076
v15	0.760576	-0.095871
v16	0.593574	0.269126
v17	0.451203	-0.111711
v18	0.725532	**0.417065**
v19	0.361339	-0.197338
v20	0.116087	-0.283614
v21	0.512043	**0.345655**
v22	-0.045555	0.264878

(3) Total-Sample Standardized Canonical Coefficients

Variable	Can1	Can2
v4	2.20664064	1.14717014
v5	-1.29441256	**-7.28758830**
v6	-1.88656987	1.41224150
v7	1.19813550	**6.97436468**
v8	0.88095578	0.97932782
v9	-0.53521212	-1.25167526
v10	0.16348543	-0.20978626
v11	-0.48940387	-0.02370969
v12	0.34296483	-0.19078962
v13	0.21085257	-0.63991444
v14	0.20793931	0.40710833
v15	0.17302770	-0.24243622
v16	0.13295992	0.41005661
v17	-0.02748186	-0.00185147
v18	0.42290424	0.71065636
v19	**11.85793927**	**-5.40534569**
v20	**14.40526450**	**-6.42814071**
v21	0.19589232	0.29079279
v22	-0.27396644	-1.17566832

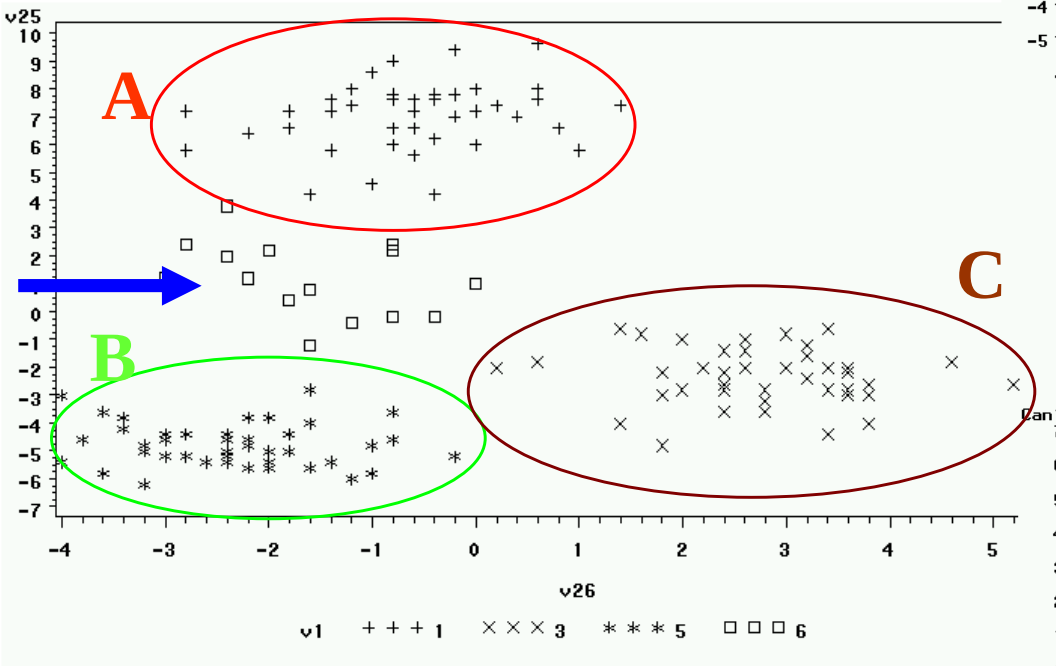
Oxycoccus - kanonicka diskriminacni analyza

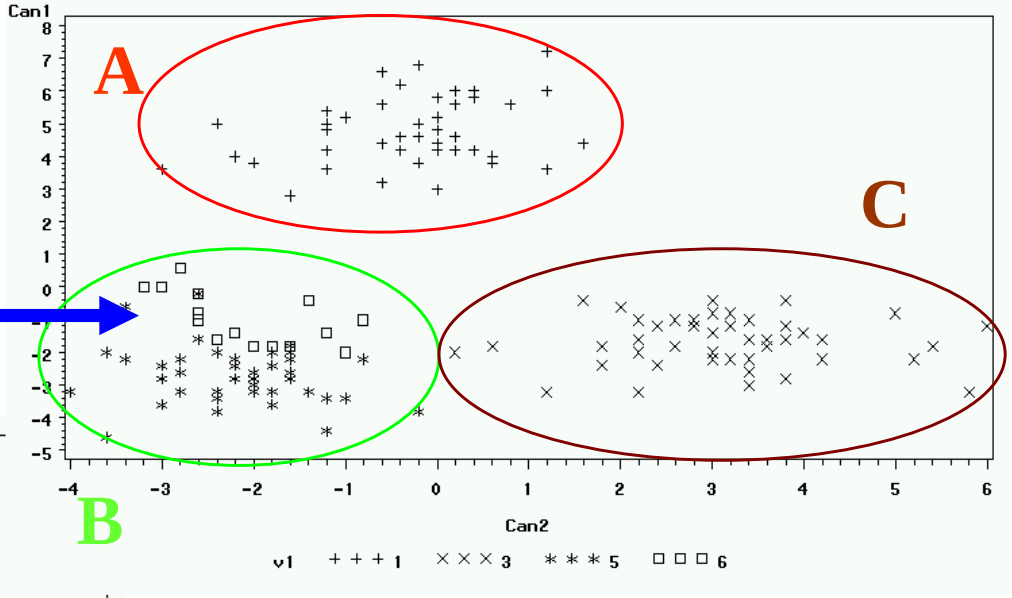
The CANDISC Procedure

Raw Canonical Coefficients

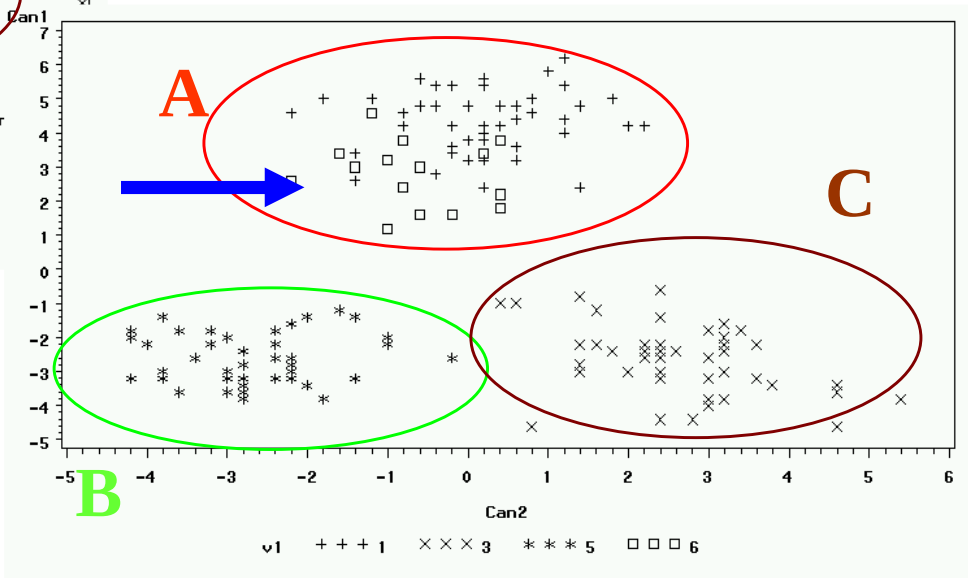
Variable	Can1	Can2
v4	0.838448541	0.435885713
v5	-1.304051753	-7.341857294
v6	-1.452234796	1.087108555
v7	1.134513637	6.604020879
v8	0.784902149	0.872548345
v9	-0.992564918	-2.321264617
V10	0.689107945	-0.884270705
v11	-2.269512407	-0.109948923
v12	0.454311114	-0.252730994
v13	0.793906003	-2.409417741
v14	0.525571082	1.028975078
v15	0.176677291	-0.247549821
v16	0.242091319	0.746624615
v17	-0.102516110	-0.006906569
v18	0.660204493	1.109420248
v19	1.877413657	-0.855803828
v20	1.939521377	-0.865483332
v21	0.199233118	0.295752053
v22	-0.581907220	-2.497130258

nezařazené





zařazené do B



zařazené do A

Pozor: zařazení přechodných objektů do různých skupin může přinést různé diskr. funkce a různé výsledky

Klasifikační diskriminační analýza

(a) hledání identifikačního (klasifikačního) kritéria

skupiny objektů známého zařazení

skupinu objektů neurčitého postavení

(b) zjištění účinnosti klasifikačního kritéria

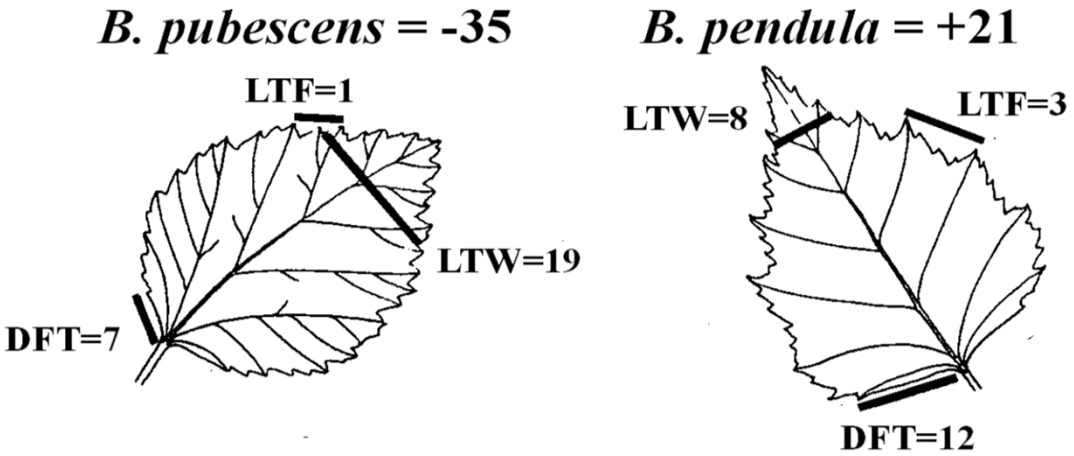
resubstituce (*resubstitution*)

křížové ověření (*cross-validation*)

Účinnost klasifikačního kriteria testujeme na stejném souboru dat, z něhož se toto klasifikační pravidlo odvozuje (tento způsob testu se nazývá **resubstituce**, *resubstitution*). Pokud máme menší počet objektů, je vhodné použít tzv. **křížové ověření** (*cross-validation*): Ze souboru *n* objektů vybereme *n* – 1 objektů, které použijeme jako tréninkový soubor. Na základě tohoto tréninkového souboru odvodíme klasifikační kritérium, které potom aplikujeme na jeden vypuštěný případ. Celý postup opakujeme *n*-krát.

Způsoby odvození klasifikačního pravidla:

(1) Kanonická diskriminační funkce - objekty se klasifikují na základě jejich skóre na kanonické diskriminační funkci anebo na základě jejich projekce do kanonického prostoru

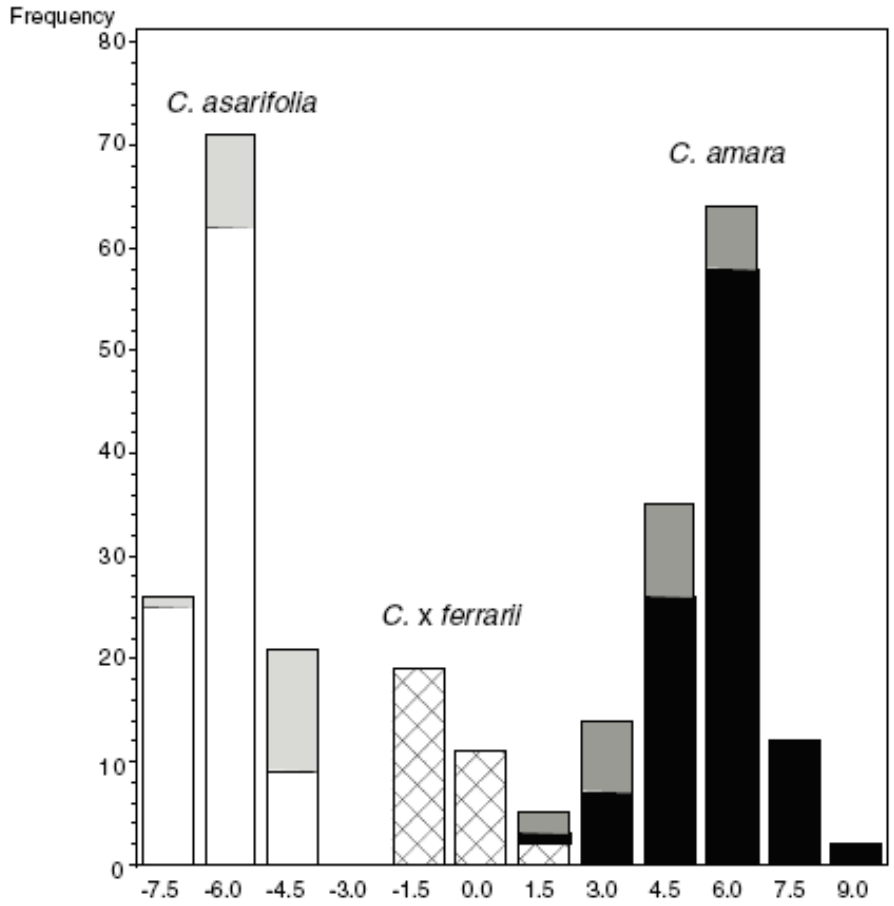


diskriminační funkce na určení druhů *Betula pubescens* a *B. pendula*
12LTF + 2DFT – 2LTW – 23
kladné hodnoty *B. pendula*, záporné hodnoty *B. pubescens*
pravděpodobnost správného určení 93%
(Stace, C. A., 1991, New Flora of the British Isles)

Způsoby odvození klasifikačního pravidla:

(1) Kanonická diskriminační funkce - objekty se klasifikují na základě jejich skóre na kanonické diskriminační funkci anebo na základě jejich projekce do kanonického prostoru

Klasifikovaný objekt se zobrazí v kanonickém prostoru spolu se souborem známých objektů (jejichž příslušnost ke skupinám je známá). Podle vzájemné pozice klasifikovaného objektu a souboru známých objektů se usuzuje na příslušnost tohoto prvku k některé skupině.



(2) výpočet lineární klasifikační funkce pro každou skupinu

Pro každou skupinu objektů se vypočítá samostatná lineární klasifikační funkce. Dále se vypočítá klasifikační skóre neznámého (klasifikovaného) objektu pro každou z těchto funkcí. Objekt bude zařazen do skupiny, pro kterou klasifikační skóre dosáhne nejvyšší hodnoty.

(3) klasifikační pravidla založená na pravděpodobnostních modelech

(i) lineární diskriminační funkce

(ii) kvadratické diskriminační funkce

(iii) neparametrické metody, např. *k*-nejbližších sousedů (*k-nearest neighbors*)

(3) klasifikační pravidla založená na pravděpodobnostních modelech

- $\boldsymbol{v}$ je obecně p-komponentní vektor znaků
- $\boldsymbol{v}_0$ představuje vektor znaků jednoho konkrétního objektu
- $\boldsymbol{v}$ má rozdílné pravděpodobnosti, že patří do π_1 a π_2
- hustoty pravděpodobností (*probability densities*) - $f_1(\boldsymbol{v})$ pro π_1 a $f_2(\boldsymbol{v})$ pro π_2.
- prostor R obsahující všechny objekty má podprostory R_1 a R_2,
  platí $R_1 \cap R_2 = 0$ a $R = R_1 \cup R_2$)
- klasifikační pravidlo bude definovat rozdělení prostoru R na dva vzájemně se vylučující podprostory R_1 a R_2, a současně bude přiřazovat objekty ze skupiny π_1 do R_1 a objekty ze skupiny π_2 do R_2.

podprostor R_1 je definován jako soubor vektorů $\boldsymbol{v}$,

pro které platí: $f_1(\boldsymbol{v}) > f_2(\boldsymbol{v})$

podprostor R_2 je definován jako soubor vektorů $\boldsymbol{v}$,

pro které platí: $f_1(\boldsymbol{v}) \leq f_2(\boldsymbol{v})$.

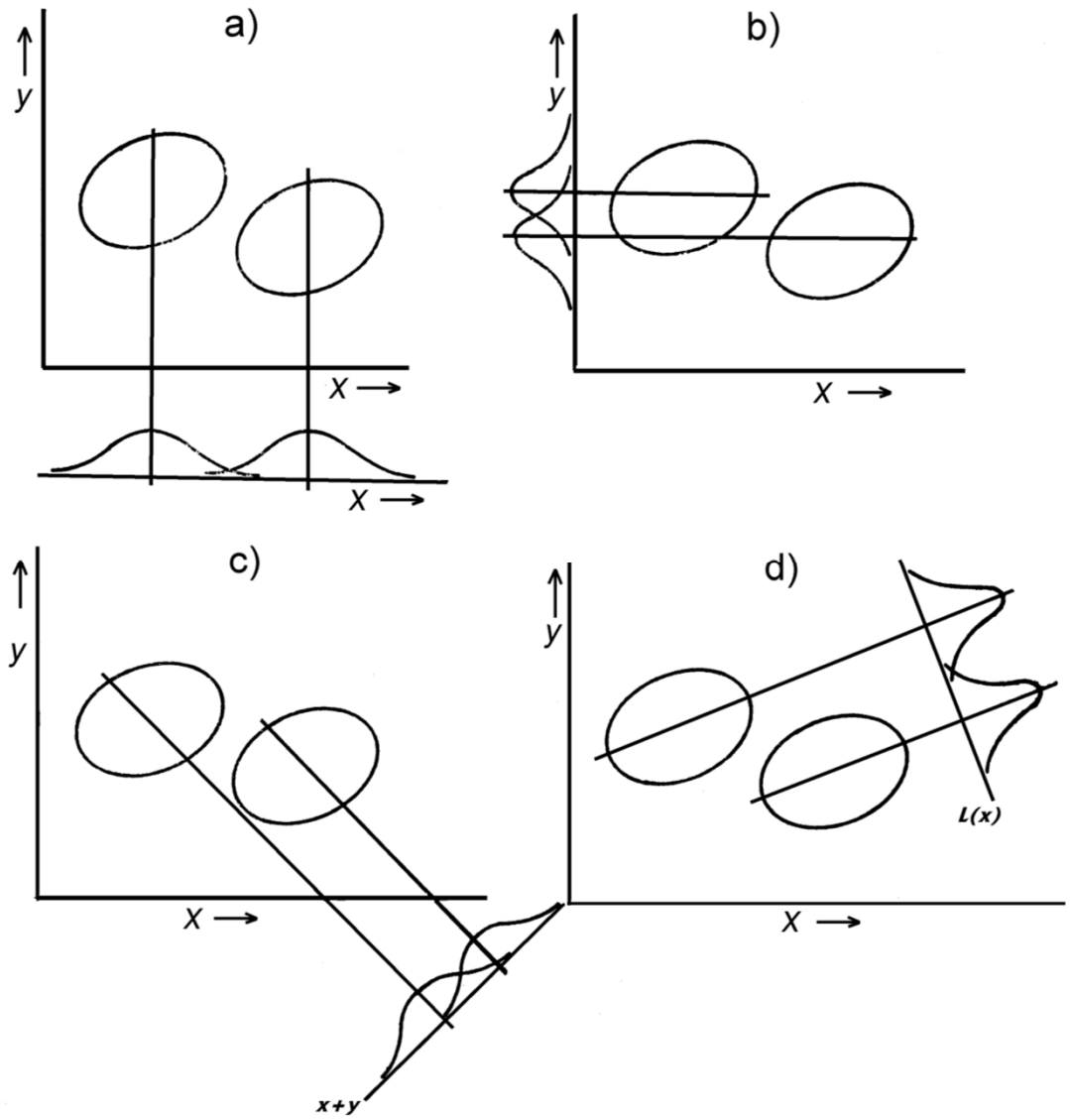
hodnoty $f_1(\boldsymbol{v})$ a $f_2(\boldsymbol{v})$ je možné odhadnout na základě výsledků

měření tréninkového souboru

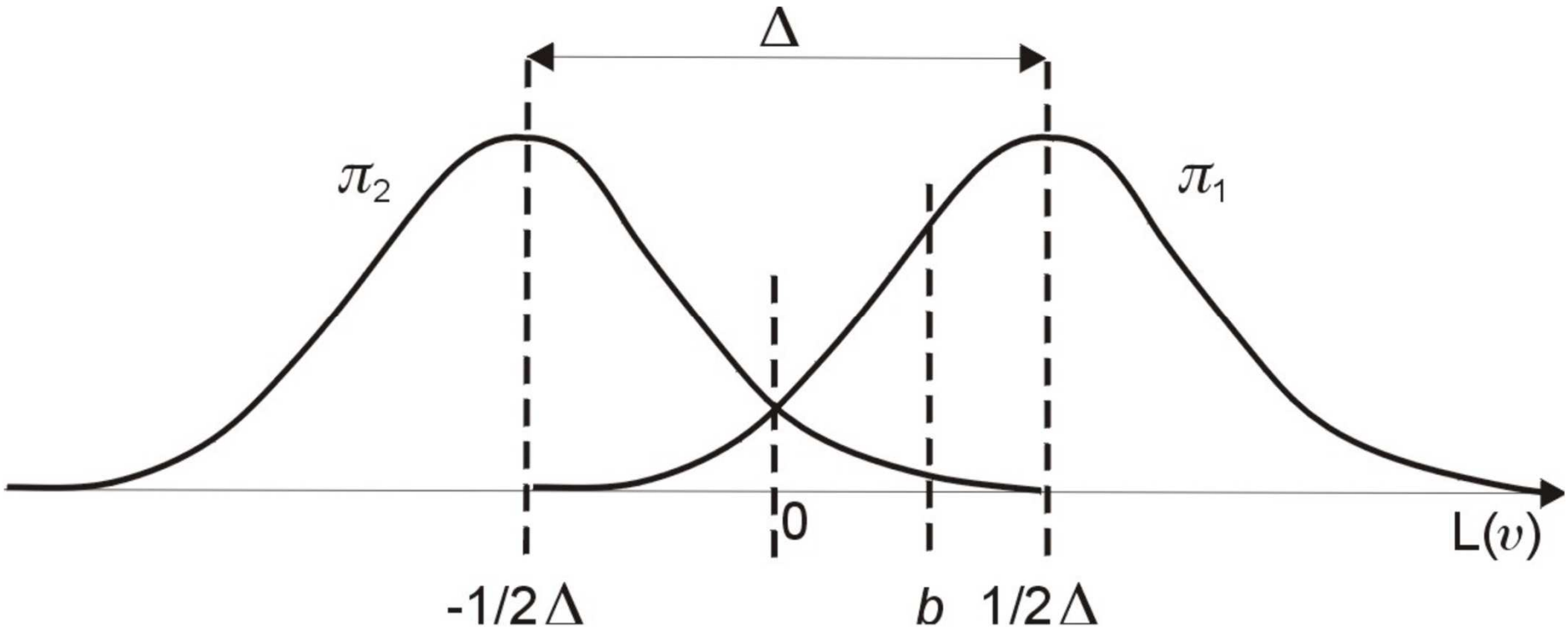
klasifikační pravidlo má potom tvar:

$\boldsymbol{v}$ náleží do π_1, pokud $f_1(\boldsymbol{v})/f_2(\boldsymbol{v}) > 1$

$\boldsymbol{v}$ náleží do π_2, pokud $f_1(\boldsymbol{v})/f_2(\boldsymbol{v}) \leq 1$



$f_i(\boldsymbol{v})$ je v těchto případech založen na předpokladu,
že většina objektů je nahloučena kolem těžiště (centroidu)
a s rostoucí vzdáleností od těžiště jejich hustota klesá



L(v) – lineární diskriminační funkce

Δ – Mahalanobisova vzdálenost vyjadřující odlišení skupin π_1 a π_2

Number of Observations and Percent Classified into v2

From v2	1	2	4	Total
1	30	0	0	30
	100.00	0.00	0.00	100.00
2	0	209	1	210
	0.00	**99.52**	0.48	100.00
4	0	2	965	967
	0.00	0.21	**99.79**	100.00
Total	30	211	966	1207
	2.49	17.48	80.03	100.00
Priors	0.33333	0.33333	0.33333	

[1] Posterior Probability of Membership in v2

Obs	v2	From v2	Classified into v2	1	2	4
200	2	4	*	0.0000	0.4354	0.5646
437	4	2	*	0.0000	0.8598	0.1402
452	4	2	*	0.0000	0.5198	0.4802

* Misclassified observation

Kroková diskriminační analýza (*stepwise discriminant analysis*)

Kroková diskriminanční analýza vyhledává takovou kombinaci znaků, které společně umožňují co nejlepší oddělení stanovených skupin

Soubor nejvhodnějších znaků je vybírán postupně, v jednotlivých krocích

Metoda začíná selekcí znaku, který je nejlepší pro oddělení předem stanovených skupin, v dalším kroku posuzuje všechny zbývající znaky a hledá takový, který skupiny nejlépe odděluje v kombinaci s již vybraným znakem

V každém kroku se počítá statistická významnost vybraných znaků (hodnota „*F-to-remove*“, *statistics for removal*) a statistická významnost znaků ostatních (hodnota „*F-to-enter*“, *statistics for entry*)