LINEÁRNÍ DISKRIMINAČNÍ FUNKCE

Tomáš Havránek, Jaroslav Vorlíček

Matematické středisko biologických ústavů ČSAV při Fyziologickém ústavu ČSAV, 142 20 Praha 4, Vídeňská 1083

V následujícím výkladu se věnujeme některým aspektům klasické lineární diskriminační analýzy, zejména hodnocení chybné klasifikace, výběru veličin pro klasifikaci a otázkám robustnosti. Výklad je vzhledem k rozsahu nutně zjednodušený a neúplný. Snažíme se alespoň zachytit podstatné otázky. V základním výkladu se držíme knihy [Lachenbruch, 1975] .

I. ÚVOD

Zopakujme si stručně a zhruba o co jde. Zavedeme si následující označení: Uvažujeme populace Π_i , i=1,...,g. Nechť x je vektor pozorování (kX1) k rozměrného náhodného vektoru. Nechť μ_i , Σ_i je vektor středních hodnot a kovarianční matice v i-té populaci. Předpokládejme, že v i-té populaci máme k-rozměrnou hustotu f_i (vůči nějaké společné σ-konečné míře) . Označme si n_i rozsah výběru z populace Π_i a $\bar{x}_i$, S_i odpovídající výběrový průměr a kovarianční matici.

Předpokládejme, že $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$. Pro g=2 můžeme používat lineární (teoretickou) diskriminační funkci

$$D_T(x) = \left(x - \tfrac{1}{2}(\mu_1 + \mu_2)\right)' \Sigma^{-1} (\mu_1 - \mu_2) .$$

Je-li $D_T(x) > 0$, klasifikujeme pozorování do populace Π_1, jinak do Π_2. Této diskriminační funkci odpovídá výběrová lineární diskriminační funkce

$$D_S(x) = \left(x - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\right)' S^{-1} (\bar{x}_1 - \bar{x}_2) .$$

Pro více populací používáme diskriminační funkce (pro i=1,...,g) :

$$D_T^i(x) = \left(x - \tfrac{1}{2}\mu_i\right)' \Sigma^{-1} \mu_i$$

(resp. jejich výběrové protějšky). Klasifikujeme do té populace, pro kterou je hodnota $D_T^i(x)$ největší (v případě rovnosti můžeme znáhodňovat) .

1.1 Vraťme se nyní ke dvěma populacím. Máme k rozměrný výběrový prostor R. Je nutné ho rozdělit na dvě množiny R_1, R_2 takové, že $R_1 \cup R_2 = R$ a $R_1 \cap R_2 = \emptyset$. Je-li $x \in R_1$, klasifikujeme do první populace, je-li $x \in R_2$, klasifikujeme do druhé populace. Zaveďme si pravděpodobnost chybné klasifikace (budiž p_i apriorní pravděpo-

dobnost pro Π_i):

$$T(R,f) = p_1 \int_{R_2} f_1(x)\,dx + p_2 \int_{R_1} f_2(x)\,dx =$$

$$= p_1 + \int_{R_1} (p_2 f_2(x) - p_1 f_1(x))\,dx .$$

Minimalizace: R_1 tak, aby $p_2 f_2(x) - p_1 f_1(x) < 0$ pro $x \in R_1$, t.j. klasifikujeme do Π_1, je-li $f_1(x) / f_2(x) > p_2/p_1$.

Budeme se dále v zásadě zabývat případem, kdy $p_1 = p_2$. Označíme si pak

$$p_1 = \int_{R_2} f_1(x)\,dx \qquad \text{a} \qquad p_2 = \int_{R_1} f_2(x)\,dx$$

1.2 Pro normální rozložení se společnou kovarianční maticí Σ máme:

$$f_i(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(x-\mu_i)' \Sigma^{-1} (x-\mu_i)\right)$$

a tedy

$$\frac{f_1(x)}{f_2(x)} = \exp\left(x' \Sigma^{-1}(\mu_1-\mu_2) - \tfrac{1}{2}(\mu_1+\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)\right)$$

a z toho máme klasifikační pravidlo

$$D_T(x) = \left(x - \tfrac{1}{2}(\mu_1+\mu_2)\right)' \Sigma^{-1}(\mu_1-\mu_2) > \ln(p_2/p_1)$$

(pro $p_1 = p_2$ je $\ln(p_2/p_1) = 0$).

Rozložení $D_T(X)$ (zkomáme ho pro stanovení pravděpodbnosti chybné klasifikace):

$$E(D_T(X) \mid \Pi_1) = \left(\mu_1 - \tfrac{1}{2}(\mu_1+\mu_2)\right)' \Sigma^{-1}(\mu_1-\mu_2) = \tfrac{1}{2}\delta^2$$

Podobně $E(D_T(X) \mid \Pi_2) = -\tfrac{1}{2}\delta^2$ a $\mathrm{VAR}(D_T(X) \mid \Pi_i) = \delta^2$,
kde δ^2 je teoretická Mahalanobisova vzdálenost, t.j.

$$\delta^2 = (\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2).$$

$D_T(X)$ je lineární funkce normálně rozložené náhodné veličiny; máme pak

$$p_1 = P\left(D_T(X) < \ln\frac{1-p_1}{p_1}\right) = \Phi\left(\frac{\ln\frac{1-p_1}{p_1} - \frac{\delta^2}{2}}{\delta}\right),$$

kde Φ je distribuční funkce normovaného normálního rozložení. Je-li $p_1 = p_2$, je $p_1 = p_2 = \Phi(-\frac{\delta}{2})$.

1.3 Jak vypadá rozložení $D_S(X)$? Podmíněno $\bar{x}_1, \bar{x}_2$ a S je

$$E(D_S(X) \mid \Pi_1) = \left(\mu_1 - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)\right)' S^{-1}(\bar{x}_1-\bar{x}_2)$$

$$\text{VAR}\,(D_S(x)\,|\,\Pi_i) = (\bar{x}_1 - \bar{x}_2)' S^{-1} \Sigma S^{-1} (\bar{x}_1 - \bar{x}_2)$$

a rozložení je opět normální. Nepodmíněně je

$$E(D_S(x)\,|\,\Pi_1) = \tfrac{1}{2} c_1 \left(\delta^2 - \frac{k(n_2-n_1)}{n_1 n_2}\right), \quad E(D_S(x)\,|\,\Pi_2) = \tfrac{1}{2} c_1 \left(-\delta^2 - \frac{k(n_2-n_1)}{n_1 n_2}\right)$$

kde $c_1 = (n_1+n_2-2)/(n_1+n_2-k-3)$ a $\text{VAR}\, D_S(x) = c_2 \left(\delta^2 + k(n_1+n_2)/(n_1 n_2)\right)$, kde

$$c_2 = \frac{(n_1+n_2-3)(n_1+n_2-2)^2}{(n_1+n_2-k-2)(n_1+n_2-k-3)(n_1+n_2-k-5)} .$$

Rozložení D_S není exaktně známo. Víme, že D_S je konsistentním odhadem D_T a je známa aproximace rozložení D_T [Okamoto, 1963].

II. HODNOCENÍ DISKRIMINAČNÍ FUNKCE

2.1 Testování rozdílu mezi skupinami je proveditelné za výše uvedených předpokladů pomocí výběrové Mahalanobisovy vzdálenosti

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

resp. statistiky $F = \frac{n_1 n_2}{n_1+n_2} \frac{n_1+n_2-k-1}{(n_1+n_2-2)k} D^2$. Je známo její aproximativní zobecnění pro více skupin [2] (Wilksovo Λ). Problémem zůstává odhad míry chyb klasifikace a otázka výběru veličin pro klasifikaci. Věnujme se nyní první z těchto otázek.

2.2 Je možné používat následujících "měr" chybné klasifikace:

1: Optimální (známe např. μ_1, μ_2, Σ): $T(R,f)$

2: Aktuální míru chyb:

$$T(\hat{R},f) = p_1 \int_{\hat{R}_2} f_1(x)\,dx + p_2 \int_{\hat{R}_1} f_2(x)\,dx, \text{ kde}$$

$\hat{R}_1, \hat{R}_2$ je rozklad výběrového prostoru vytvořený na základě výběru.

3: Očekávaná míra chyb: $E\,T(\hat{R},f)$

4: Odhadnutá míra chyb "dosazením":

$$T(\hat{R},\hat{f}) = p_1 \int_{\hat{R}_2} \hat{f}_1(x)\,dx + p_2 \int_{\hat{R}_1} \hat{f}_2(x)\,dx$$

(t.j. například použitím $\bar{x}_1, \bar{x}_2, S$ v normálních hustotách).

5: Nalezená míra chyb: pro daný výběr (učící, se známou klasifikací) je diskriminační funkce (rozklad R_1, R_2) odhadnuta a každý vektor

pozorování je klasifikován pomocí $\hat{R}_1, \hat{R}_2$; je stanoven podíl chybných zařazení.

6: Odhad pomocí testovacího souboru: pomocí učícího souboru je nalezeno $\hat{R}_1$, $\hat{R}_2$ a v dalším výběru se známou klasifikací (testovacím) je stanoven podíl chybných klasifikací.

7: Míra chyb nalezená na učícím souboru s vylučováním: pomocí učícího souboru je stanoveno $\hat{R}_1, \hat{R}_2$ .Učící soubor je pak klasifikován,ale pro každý vektor je vždy upraveno $\hat{R}_1$ a $\hat{R}_2$ tak,jako by bylo vypočteno ze souboru,který tento vektor neobsahoval.

8: Odhad použítím Okamotovy formule pro rozložení výběrové diskriminační funkce (resp. pomocí jiné aproximace rozložení D_S).

Je patrně zřejmé,že 1,2,3,4 a 8 výše jsou silně závislé na rozložení; za robustní lze považovat 5,6 a 7.

2.3 V práci [Lachenbruch,Mickey,1968] jsou tyto metody hodnoceny z hlediska přiblížení k $T(R,f)$ (resp P_1) ,vypočtené při úplné informaci o rozložení.Jde vždy o normální rozložení. V práci se nehodnotí metoda 6 , protože vyžaduje nový testovací soubor.Tento požadavek je v praxi často nereálný nebo nepopulární.Je-li splnitelný, je 6 nejrozumnější metoda při dostatečné velikosti testovacího souboru (jde pak o odhad binomické pravděpodobnosti, jehož přesnost je zřejmá) .

Hodnoceny jsou tedy metody,kterými lze "nalézt" míru chybné klasifikace na základě jediného učícího souboru. Pro označení metod je použito jednoduchých symbolů. Jde o metody:

4 D : dosazování,

5 R : nalezená míra chyb,

7 U : vylučování,

DS: jiné odhady Mahalanobisovy vzdálenosti v $\hat{P}_1 = \Phi(-D^2/2)$; konkrétně jde o korekci na nestrannost

$$D_s^2 = \frac{n_1+n_2-k-3}{n_1+n_2-2} D^2$$

8 O : D^2 pro δ^2 v Okamotově rozvoji při výpočtu

$$P_1 = P(D_S(X)<0/\pi_1) .$$

OS: D_s^2 pro δ^2 v Okamotově rozvoji.

Ū : kombinace empirického přístupu metody 7 a využití normálního rozložení: $P_1 = \Phi(-\bar{D}_1/S_{D_1})$ kde $\bar{D}_1$ a S_{D_1} je výběrový průměr a rozptyl hodnot diskriminační funkce nalezených (při vylučování) pro vektory pozorování z první skupiny populace .

Označíme-li si P_1 "optimální" pravděpodobnost chyby , $P_1 = \int_{R_2} f(x)\,dx$ při plné znalosti rozložení a $\hat{P}_{1*}$ odhad, je použito pro hodnocení odhadu vzdálenosti $e = |P_1 - \hat{P}_{1*}|$. Hodnocení je prováděno na základě simulací učícího sou

boru, t.j. generování rozložení se známými parametry a následným stanovením D_S pro každý generovaný soubor. Bylo generováno vždy 48 výběrů s následujícími parametry : $\Sigma = I$, $\mu_1 = 0$, $\mu_2 = (\delta, 0, \dots, 0)$, kde

δ^2	1.098	1.817	2.836	4.293	6.574	11.482
$P_1 = \Phi(\delta/2)$	.30	.25	.20	.15	.10	.05

Pro každou z následujících kombinací velikostí výběru a dimenze bylo generováno 12 výběrů:

	dimenze k.			
	2	4	8	20
n_1, n_2	4,4	8,8	8,8	15,15
	4,8	8,16	8,16	15,30
	4,12	8,24	8,24	15,45
	16,16	20,20	20,20	25,25
	16,32	20,40	20,40	25,50
	16,48	20,60	20,60	25,75

Celkové výsledky jsou vidět z následující tabulky:

Počet výběrů s chybou e v daném rozmezí:

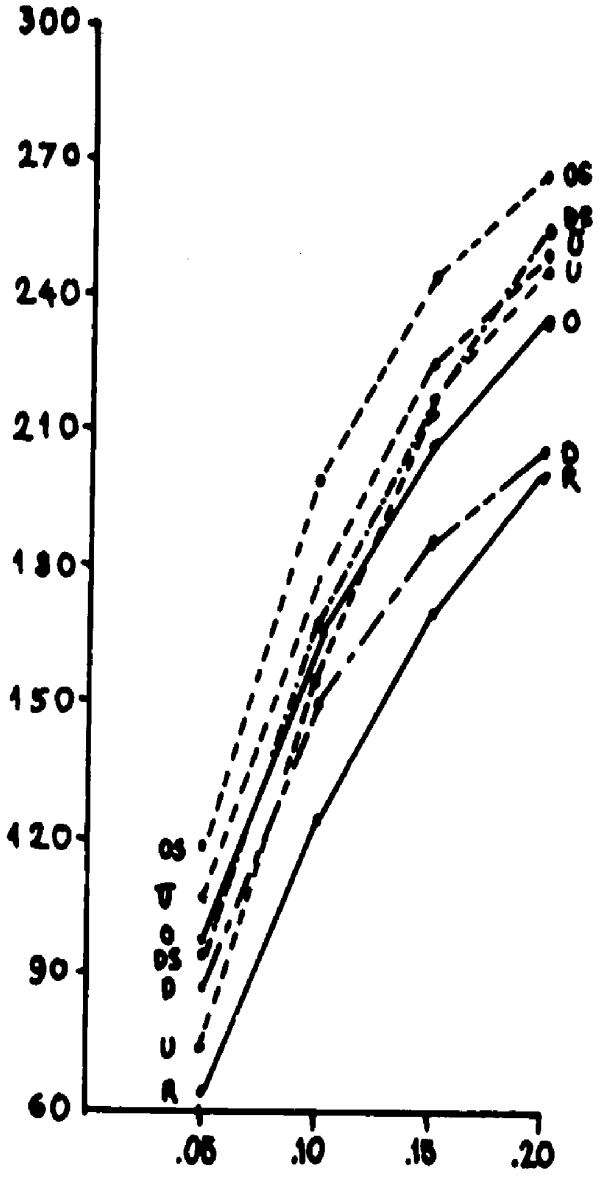
	e<0.05	.05<e<.10	.10<e<.15	.15<e<.20	e>.20
O	97	68	42	28	53
OS	118	81	45	23	21
D	87	52	37	40	72
DS	94	73	50	36	35
$\bar{U}$	107	70	48	25	38
U	73	82	62	29	42
R	64	60	46	30	88

Speciálně například pro $k = 2$ a $n_1 + n_2 - k - 1 \leq 25$ dostáváme při 24 výběrech následující rozložení "velmi špatných" případů odhadu:

	O	OS	D	DS	$\bar{U}$	U	R
e>.20	10	2	13	8	3	3	16

Poznamenejme ještě, že tabulky nejsou stratifikovány vzhledem k δ^2 a tedy "optimální" chybě a zároveň, že je sledována pouze absolutní hodnota odchylky od P_1 ač v praxi záleží i na směru této odchylky. Celá situace vyžaduje další podrobnější zkoumání .

Rozložení vzhledem k e nám ukazuje následující obrázek:



Na základě těchto simulací se zdá, že nejlepší metody jsou U, $\bar{U}$ a OS, kde ovšem U a OS jsou založeny na předpokladu normality. U je metoda v podstatě robustní (navíc je používána ve standardních systémech statistických programů, např. BMDP).

K téže problematice ještě upozorňujeme na práce [Lachenbruch, 1967] a [McLachlan, 1974]. Prrvní z nich se týká metody $\bar{U}$ a druhá jedné další metody založené na předpokladu normality a redukci vychýlenosti odhadu P_1 pomocí metody D (dosazování). Zároveň je nutné upozornit na Cochranovu poznámku [1968] k [Lachenbruch, 1968]. Cochran se domnívá, že $\bar{U}$ metoda vyžaduje příliš mnoho počítání k tomu, aby ji bylo možno považovat za prakticky užitečnou. Dnešní široké používání této metody Cochranovu námitku vyvrací. Práce [McLachlan, 1974] potvrzuje jinými metodami závěry Lachenbruchovy, ovšem pouze pro odhadování založené

na normalitě (D,DS,O,OS).

III. VÝBĚR VELIČIN PRO DISKRIMINACI

3.1 Věnujme se nyní problému výběru veličin nutných pro klasifikaci. Je možné používat různé procedury. Nejjednodušší z nich je použít ty veličiny, pro které $(j=1,\ldots,k)$ je

$$|t_j| \geq C_{1-\alpha/2k,\, n_1+n_2-2}$$

($1-\alpha/2k$ kvantil studentova rozložení s n_1+n_2-2 stupni volnosti), kde

$$t_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s_j\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{a} \quad s_j^2 = \frac{1}{n_1+n_2-2}\left(\sum_{k=1}^{n_1}(x_{1jk}-\bar{x}_{1j})^2 + \sum_{k=1}^{n_2}(x_{2jk}-\bar{x}_{2j})^2\right).$$

Tato metoda má tu nevýhodu, že můžeme snadno zahrnout do veličin vybraných pro konstrukci klasifikačního pravidla např dvě na sobě silně závislé veličiny. Použijeme-li obě tyto veličiny místo jedné z nich, nezískáváme téměř žádný "informační přínos".

3.2 Proto se používají různé metody postupného výběru (s celou řadou variant). V zásadě v každém kroku takové metody je veličina zařazena, jestliže (1) maximalizuje hodnotu nějaké statistiky a (2) tato hodnota přesahuje jistou mez. Pro každý krok jsou hodnoty statistik upravovány v závislosti na tom, které veličiny byly v předchozích krocích zařazeny.

Nejobvyklejší postup tohoto typu je založen na využití F-testů pro podmíněné střední hodnoty. Označme si

$$w_{jj'} = \sum_{i=1}^{g}\sum_{r=1}^{m_i}(x_{ijr} - \bar{x}_{ij})(x_{ij'r} - \bar{x}_{ij'}) \quad \text{a}$$

$$t_{jj'} = \sum_{i=1}^{g}\sum_{r=1}^{m_i}(x_{ijr} - \bar{x}_{j})(x_{ij'r} - \bar{x}_{j'})^2.$$

Příslušné matice $W(A)$ a $T(A)$ pro $A = \{j_1,\ldots,j_p\} \subseteq \{1,\ldots,k\}$ používáme pak pro další výpočty. Wilksovo Λ je

$$\Lambda(A) = \frac{\det W(A)}{\det T(A)}$$

Označme si pro $j \notin A$ $\quad \Lambda(A.j) = \dfrac{\Lambda(A,j)}{\Lambda(A)}$ (Pedanticky vzato bychom místo A,j měli psát $A \cup \{j\}$).
Odpovídající statistika F je

$$F = \frac{n-g-p}{p-1} \quad \frac{1-\Lambda(A.j)}{\Lambda(A.j)}$$

pro testování rozdílu v podmíněné střední hodnotě veličiny j podmíněno veličinami z A . Můžeme rozeznávat F-hodnotu "pro vstup" (přidání veličiny k A) a F-hodnotu"pro vyřazení" (vyřazení veličiny z A). Pravidla pro rozhodování o veličině v daném kroku pak bývají:

(a) nevyřazuj veličinu,je-li F pro vyřazení $\geq mez_1$,

(b) nezařazuj veličinu,je-li F pro zařazení $< mez_2$,

(c) nezařazuj veličinu,je-li tolerance (t.j. $1-R^2(j/A)$) $< mez_3$.

Výběr konkrétní vveličiny se pak řídí maximalizací F pro vstup.

3.3 Numericky je ovšem výpočet F hodnot prováděn optimálněji (viz např. [Jennrich ,1977]).Výpočty se provádějí postupně,tak jak jsou veličiny zahrnovány do diskriminační funkce (tj. do A).Nechť B je regulární matice.Definujeme "setřásání" (sweeping) matice podle k-tého diagonálního prvku b_{kk} jako

$$\tilde{b}_{kk} = -\frac{1}{b_{kk}} \;,\; \tilde{b}_{ik} = \frac{b_{ik}}{b_{kk}} \;,\; \tilde{b}_{kj} = \frac{b_{kj}}{b_{kk}} \;,\; \tilde{b}_{ij} = b_{ij} - \frac{b_{ik}\,b_{kj}}{b_{kk}} \quad (i \neq k, j \neq k).$$

Podobně lze postupovat i pro bloky. Připomeňme si

$$\det B = \det B_{11} \det (B_{22} - B_{21} B_{11}^{-1} B_{12}) .$$

Je-li $A = \{j_1,\ldots,j_p\} \subseteq \{1,\ldots,k\}$ množina indexů veličin zařazených do diskriminace,bylo postupně provedeno "setřásání" matic W a T podle $w_{j_1 j_1}$, $w_{j_2 j_2}$,.... (na pořadí nezáleží). Označme si takto vzniklé matice $\tilde{W}$ a $\tilde{T}$. Pak F pro vstup je ($j \notin \{j_1,\ldots,j_p\}$)

$$F_j = \frac{n-p-g}{p-1} \cdot \frac{1 - V_j}{V_j} \quad ,\text{kde} \quad V_j = \frac{\tilde{w}_{jj'}}{\tilde{t}_{jj'}}$$

(je totiž $\det (W(\{ j_1,\ldots,j_p,j\})) = \det W (\{j_1,\ldots,j_p\})\, \tilde{w}_{jj}$ a $\det (T(\{ j_1,\ldots,j_p,j\})) = \det T(\{j_1,\ldots,j_p\})\, \tilde{t}_{jj}$). Pro F pro vyřazení $j \in \{j_1,\ldots j_p\}$ máme

$$F_j = \frac{n-p-g+1}{p-1} \cdot (V_j - 1)$$

a konečně tolerance je $t_j = \tilde{w}_{jj}/w_{jj}$.

V každém kroku jsou pak matice $\tilde{W}$ a $\tilde{T}$ upravovány setřesením příslušných diagonálních prvků odpovídajících zařazené veličině (resp. zpětně upraveny při vyřazení veličiny.

3.4 Otázkou je volba vhodných mezí. Obvykle se používá příslušných kritických hodnot F rozložení pro konvenční hladinu významnosti $\alpha = 0.05$ .To je chybné jednak z důvodů,že testujeme vždy vlastně nejlepší z několika hodnot F statistiky, jednak z důvodů simultánní statistické inference. Problém prvního typu (výběr nejlepší z několika hodnot) je částečně řešen (pro dvě testové statisti-

ky) pro obdobnou situaci v postupné mnohorozměrné lineární regresi [Draper, Guttman a Lapczak,1979] . Výsledky ukazují,že skutečné hladiny významnosti mohou být o mnoho horší než je stanovené $\alpha = 0.05$ (ve zkoumaných případech od 0.108 do 0.541). Otázkami simultánní inference se zabývá práce [McKay, 1977] ;naneštěstí je tato práce poměrně těžko čitelná. Pro praxi lze snad doporučit používat sice kritických hodnot F statistiky,ale s vysokou hladinou významnosti a být si přitom vědom výše uvedených problémů. Praxe ukazuje,že při volbě např. $\alpha = 0.05$ nebo $\alpha = 0.01$ se při postupném výběru veličin spolehlivost klasifikace (odhadnutá např.metodou $\bar{U}$) nezlepšuje,ale podle F kriteria jsou stále vybírány další a další veličiny. Odhad spolehlivosti klasifikace může tedy při praktických výpočtech korigovat výběr veličin.

3.5 V práci [Murray,1977] je ukazováno,že počet chybných klasifikací neklesá obecně s počtem vybraných veličin. Jde opět o simulační výsledky: jsou použity vždy dva výběry z k dimenzionálního normálního rozložení s kovarianční maticí I_k . Pro p veličin je pak optimální $P_1 = \Phi((-1/4)\sqrt{p})$ při použití středních hodnot 1/4 a -1/4 pro každou veličinu v prvním resp. druhém výběru. Pravděpodobnost chyb tedy roste s p. Byla dále použita lineární diskriminační funkce se známými parametry. Pravděpodobnost chybné klasifikace byla odhadována z výběrů a pomocí ní hodnoceny jednotlivé vybrané podmnožiny veličin.

Byly uvažovány následující procedury pro výběr veličin (při celkové dimensi k):

(a) k = 10 - vyzkoušej každou podmožinu,

(b) k = 10 - postupný výběr veličin,

(c) k = 50 - postupný výběr veličin do lo veličin.

Tyto procedury byly kombinovány s následujícími pravidly pro zastavování:

(1) vyber nejlepší podmnožinu dané velikosti,

(2) vyber nejlepší podmnožinu ,

(3) vyber nejlepší podmnožinu dané velikosti, není-li nejlepší podmnožina o jednu větší "ostře" lepší.

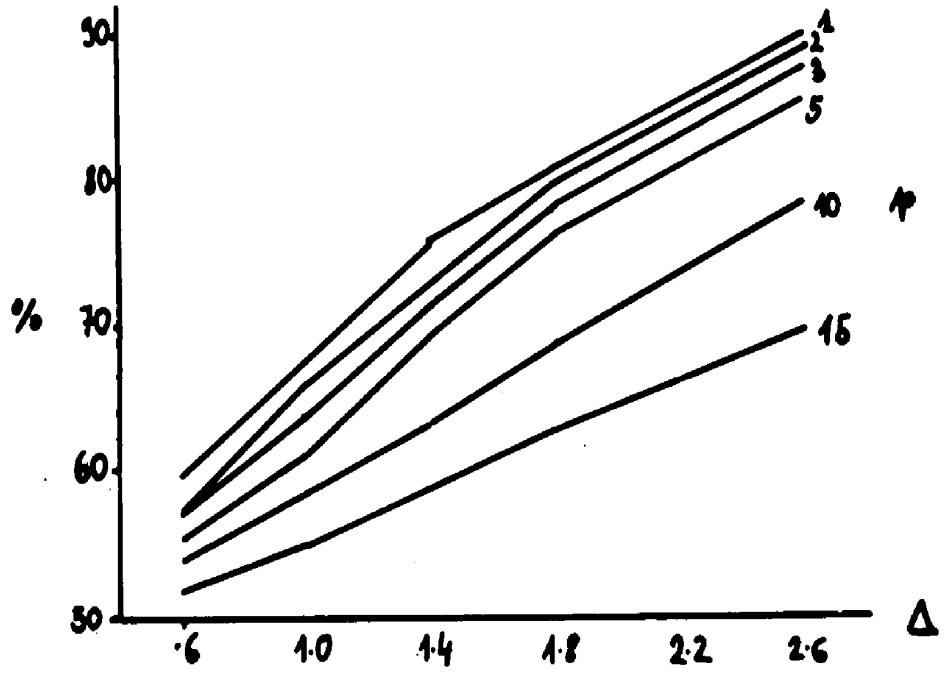
Velikost dat byla n = 25,50,100.Pro každou kombinaci metody,zastavování a velikosti dat bylo generováno 160 výběrů. Částečné výsledky zde uvádíme:

(pro pravidlo (1)):

		p									
		1	2	3	4	5	6	7	8	9	10
n= 25	a	25.2	18.0	13.9	11.5	10.6	10.0	10.8	12.2	15.3	21.4
	b	25.6	19.3	16.1	14.4	13.9	13.4	13.8	15.4	17.3	21.6
	c	18.2	11.0	6.8	4.2	2.2	1.5	1.1	0.8	0.5	0.4
n=50	a	29.5	22.8	19.0	17.0	15.7	15.1	15.3	16.0	17.9	21.8
	b	29.7	24.0	20.8	19.2	18.1	17.4	17.4	17.9	19.1	21.1
	c	25.0	17.9	13.5	10.3	8.2	6.4	5.2	4.0	3.4	3.0
n=100	a	32.8	26.7	23.2	20.6	19.3	18.3	18.0	18.2	19.2	21.6
	b	32.3	27.0	24.2	22.1	20.7	19.9	19.4	19.5	19.8	21.0
	c	29.4	22.9	18.7	15.7	13.1	11.3	9.8	8.4	7.5	6.5
opt.	P_1	40.1	36.2	33.3	30.8	28.8	27.0	25.4	24.0	22.7	21.5

podle mínění Murraye výsledky ukazují mimo jiné,že odhadnutá míra chyb,je-li použita k výběru nejlepší podmnožiny veličin,není dobrým ukazatelem pravděpodobnosti chybné klasifikace pro použití klasifikačního pravidla na další výběry (srovnej podobné problémy s R^2 v mnohorozměrné lineární regresi, viz např. [Rencher a Pun,1980] a [Hill, 1979]).Nejde ovšem o rozpor se závěrem 3.4 (je si nutné ale uvědomit,že i výběr podle nejlepšího F poznamenává odhad míry chyb).

Probléme dimense použitého vektoru veličin se zabývá i práce [Van Ness a Simpson,1979] .Zde jde o otázku,kolik je nutné použít veličin pro různé diskriminační algoritmy k dosažení srovnatelných výsledků.Srovnává se (1) lineární diskriminační funkce s neznámými středními hodnotami a známými kovariancemi, (2) lineární diskriminační funkce s neznámými středními hodnotami i kovariancemi,(3) kvadratická diskriminace s neznámými parametry,(4)metoda jader s normálními jádry, (5) metoda jader s cauchyovskými jádry.Jsou vždy použity učící soubory generované z k-dimensionálního normálního rozložení s nulovou, resp $(\Delta,0,\ldots,0)$ střední hodnotou a kovarianční maticí I_k .Pak jsou generovány testovací data. Výsledky jsou presentovány v grafech,např :



pro lineární diskriminační funkci (% je procento (průměrné) správné klasifikace a p je počet zahrnutých veličin).

Závěrem autorů je,že " vzrůst v Δ je nutný k ospravedlnění růstu dimense použitého vektoru veličin" . Snad by bylo možné shrnout,že nemá smysl zvyšovat příliš dimensi,je-li vzdálenost populací malá; výsledky klasifikace nelze takto příliš vylepšit (pozor: jde o případ,kdy vzdálenost zůstává s rostoucím počtem veličin stejná - je-li např. vzdálenost tvořena stejnými přírůstky od každé veličiny,je situace jiná!).

3.6 Příkladem srovnávání metod výběru v různých konkrétních programech je práce [Habbema,Hermans,1977] . Autoři srovnávají programy obsažené v systémech

BMDP a SPSS a programy ALLOC (vlastní program autorů) a DISCRIM. Základní srovnání obsahuje následující tabulka:

	BMDP/SPSS	DISCRIM	ALLOC
předpoklady	normalita shodné kov.matice	normalita shodné kov.matice	metody jádra (robustní)
metody výběru	F test	U statistika pro vybranou množinu	maximalizace správné klasifikace
	postupný výběr	všechny podmnožiny	postupný výběr
odhad správné klasifikace	U vylučování	R nalezená	U vylučování
zastavování	mez pro F	redukce v U	mez pro růst správné klasifikace

(hodnocení BMDP nezachycuje poslední stav z r.1979, SPSS nemá vylučování).

Programy byly aplikovány na konkrétní datový soubor (12 veličin). Výsledky byly velice pestré, jak ukazuje následující tabulka:

krok:	indexy vybraných veličin:		
	ALLOC	BMDP	DISCR
1.	8	9	9
2.	4 8	4 9	4 9
3.	4 7 8	2 4 9	2 4 9
4.	1 4 7 8	1 2 4 9	1 4 6 8
atd.			

Podobný obraz dává počet správných klasifikací (ze 48 objektů)

krok:	1	2	3	4	5	6	7	8	9
ALLOC	20	25	28	30	28	28	29	31	27
BMDP	12	17	18	17	17	20	23	22	17
DISCR	12	17	18	20	24	23	20	20	17

IV. ROBUSTNOST LINEÁRNÍ DISKRIMINAČNÍ FUNKCE

4.1 Předpoklady "optimality" lineární diskriminační funkce (pro dva výběry) jsou, jak víme :

(1) f_1 a f_2 jsou mnohorozměrné normální hustoty, (2) $\Sigma_1 = \Sigma_2 = \Sigma$, (3) známe apriorní pravděpodobnost p_1 a p_2, (4) Známe parametry μ_1, μ_2, Σ. Je možno zkoumat vliv porušení těchto předpokladů.

4.2. Studie o porušení předpokladů normality. V [Gilbert, 1968] je studován

případ alternativních veličin pro dvě populace. Nechť náhodný vektor $\mathbf{x} = (x_1, \ldots, x_k)$, kde každé x_i je alternativní náhodná veličiny nabývající hodnot 0, 1. $\mathbf{x}$ má multinomické rozložení s 2^k třídami. Označme dále Y náhodnou veličinu, nabývající hodnoty 1, je-li $\mathbf{x}$ z populace Π_1, a hodnotu 0, je-li $\mathbf{x}$ z populace Π_2. Označme $p_{\mathbf{x},Y} = P(\mathbf{x} = \mathbf{x}, Y = y)$. Pravidlo minimalizující pravděpodobnost chybné klasifikace je následující: $\mathbf{x} \in \Pi_1$ je-li $p_{\mathbf{x}1}/p_{\mathbf{x}2} > 1$ a $\mathbf{x} \in \Pi_2$, je-li $p_{\mathbf{x}1}/p_{\mathbf{x}2} < 1$. Pro případ neznámých pravděpodobností $p_{\mathbf{x},Y}$ bylo srovnáváno pět metod klasifikace založených v zásadě na různých odhadech poměru $p_{\mathbf{x}1} / p_{\mathbf{x}2}$:

(1) $p_{\mathbf{x}1}/p_{\mathbf{x}2}$ je odhadnuto podílem četností $n_{\mathbf{x}1}/n_{\mathbf{x}2}$.

(2) V modelu $\log(p_{\mathbf{x}1} / p_{\mathbf{x}2}) = 2\beta'\tilde{\mathbf{x}}$, $\beta' = (\beta_0, \ldots, \beta_k)$, $\tilde{\mathbf{x}} = (1, \mathbf{x})$ je β odhadnuto metodou maximální věrohodnosti,

(3) při stejném modelu je β odhadnuto metodou minimálního logitového χ^2:

$$\sum_{\mathbf{x}} \frac{n_{\mathbf{x}1}\, n_{\mathbf{x}2}}{n_{\mathbf{x}.}} \left[\log \frac{n_{\mathbf{x}1}}{n_{\mathbf{x}2}} - 2\beta'\tilde{\mathbf{x}} \right]^2$$

(4) $p_{\mathbf{x}1} / p_{\mathbf{x}2}$ je odhadnuto metodou maximální věrohodnoti za předpokladu nezávislosti veličin ve vektoru $\mathbf{x}$.

(5) Lineární diskriminační funkce.

Metoda (1) je jednoduchá a asymptoticky optimální, ale pro větší počet veličin je třeba pro odhad pravděpodobností $p_{\mathbf{x},Y}$ výběrů velkého rozsahu. Metoda (4) obsahuje příliš omezující předpoklad nezávislosti. Metody (2) a (3) dovolují závislost, ale neuvažují úplný multinomický model. Pro srovnání bylo třeba počet parametrů ($2^k - 1$) omezit. Byl uvažován pouze model s interakcemi prvního řádu, t.j.

$$(x) \quad \log p_{\mathbf{x}y} = \alpha + \sum_{r=1}^{k+1} (-1)^{x_r} \alpha_r + \sum_{r<s} (-1)^{x_r + x_s} \alpha_{rs}$$

kde $x_{k+1} = y$. Za platnosti tohoto modelu jsou metody (2) a (3) asymptoticky optimální. Jako míra srovnání diskriminačních metod nebyla uvažována pouze pravděpodobnost chybné klasifikace, která uvažuje pouze případ stejných apriorních pravděpodobností, ale též míra shody, jež srovnává přímo hodnoty diskriminačních funkcí. Při platnosti modelu (x) je optimální diskriminační funkce dána příslušným logaritmem poměru věrohodností (při známých paramatrech). Takto obdržená diskriminační funkce je lineární funkcí vektoru pozorovaných hodnot. Metody (2) - (5) rovněž vedou na lineární funkci vektoru pozorování "odhadující" logaritmus poměru $p_{\mathbf{x}1}/p_{\mathbf{x}2}$. Za míru shody byl zvolen korelační koeficient mezi optimální a hodnocenou lineární funkcí. Shoda byla hodnocena při dimensi k= 2,3,6 pro 6 112 populací se známými parametry v modelu (x). Distribuce korelačního koeficientu je uvedena v následující tabulce pro obvyklou lineární diskriminační funkci založenou na známých parametrech:

	k = 2	3	6	
0.99-1.00	92.6	79.5	58.6	(v procentech)
0.98-0.99	5.9	10.3	16.8	
0.97-0.98	1.5	3.5	7.4	
0.96-0.97	-	2.0	5.4	
0.95-0.96	-	1.4	2.5	

Pro další hodnocení pomocí simulací bylo pro k=6 použito 15 různých sad parmetrů v modelu (x). Bylo generováno vždy 100 výběrů rozsahu 100 a 100 výběrů rozsahu 500. Byl vypočten průměrný korelační koeficient (výběrový) pro optimální diskriminační funkci a funkce z metod (1) - (5). Rovněž byla počítána průměrná pravděpodobnost chybné klasifikace při všech pěti výběrových diskriminačních funkcích. Pořadí metod podle korelačního koeficientu bylo (4),(3),(5),(2),(1), přičemž (1) byla výrazně nejhorší. Pro průměrnou pravděpodobnost chybné klasifikace byly obdrženy podobné výsledky s tím, že optimální lineární klasifikační metody se lišila od (2)-(5) jen málo, zpravidla ne o více než 5%. Závěr uvedené studie je, že použití lineární diskriminační funkce (5) dává vzhledem k ostatním metodám jen nepodstanou ztrátu. Metoda založená na předpokladu nezávislosti veličin by patrně v modelech s větší mírou závislosti nedávala tak dobré výsledky jako zde (to je možno ovšem říci i pro lineární diskriminační funkci obecně). Moore [1973] dospěl za poněkud obecnějších předpokladů ke stejným výsledkům.

4.3 Při hodnocení rubustnosti lineární diskriminační funkce vzhledem k <u>porušení předpokladu rovnosti kovariančních matic</u> Marks a Dunnová [1974] srovnávali tři diskriminační funkce

(1) Obvyklou lineární diskriminační funkci, (2) nejlepší lineární diskriminační funkci minimalizující chybu klasifikace za přepokladu různých kovariančních matic [Anderson a Bahadur, 1962]. Podle ní klasifikujeme do Π_1, je-li

$x'\beta < \beta'\mu_1 + t_1\beta'\Sigma_1\beta$

kde $\beta = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}$ a t_1, t_2 jsou konstanty volené tak, aby byla minimalizována chyba klasifikace. (3) Kvadratickou diskriminační funkci. Uvažovali dvě skupiny se středními hodnotami $\mu_1 = 0, \mu_2 = v$ a kovariančními maticemi $\Sigma_1 = I$ a $\Sigma_2 = \Lambda$, kde Λ je diagonální matice s diagonálou $(\underbrace{\lambda, \dots, \lambda}_{k/2}, 1, \dots, 1)$, kde k je dimense rozložení. Bylo generováno 100 dvojic výběrů a byla srovnávána vypočtená (pro (1) a (2)) resp. odhadnutá na testovacím výběru (pro (3)) míra chyb. Výsledky lze shrnout takto: Pro výběry velkého rozsahu je (3) výrazně lepší než (1), jsou-li rozdíly mezi kovariančními maticemi velké (λ velké). Pro výběry malého rozsahu je (3) výrazně horší než (1) při mírně odlišných kovariančních maticích a tento rozdíl se zvětšuje při rostoucí dimensi. Pro malý rozdíl kovariančních matic je (2) o málo lepší než (1), pro velká λ je výrazně lepší, ale pak je (3) ještě

lepší.

Porušením předpokladů ve stejném směru se zabývá práce [Gilbert,1969]. V práci je studován případ,kdy $\Sigma_1 = I$ a $\Sigma_2 = d\,I$. Jsou srovnávány pravděpodobnosti chyb při známých paramatrech, Výsledky odpovídají předchozím závěrům.

4.4 Případem jiného porušení normality se zabývá práce [Lachenbruch,Sneering a Revo,1973].Jde zde o spojitá rozložení vzniklá transformací normálního rozložení (např. $y = \log x$, $y = \log(x/(1-x))$, kde x je normální, tedy lognormální a logitnormální rozložení). Předpoklad shody kovariančních matic nebyl porušen, Toto porušení předpokladů má vliv na míry chyb (jsou rozdíly mezi měrou chyb dosaženou optimálním pravidlem a měrou chyb dosaženou při použití lineární diskriminační funkcí). Metody pro odhad chyb se ch valy různě - metody založené na normalitě špatně, metoda $\bar{U}$ (vylučování)dobře . Kvadratická diskriminační funkce dávala obecně podstatně horší výsledky.

Speciálně vlivu porušení předpokladu normality na kvadratickou diskriminační funkci (pro případ s různými kovariančními maticemi) je věnována práce [Clark,Lachenbruch a Broffitt,1979]. Kromě extrémně šikmých rolození se zdá, že chování kvadratické dikriminační funkce není špatné.

LITERATURA:

1) A.A.Afifi,S.P.Azen: Statistical analysis - a computer oriented approach, Academic Press,New York,1972.

2) T.W.Anderson,R.R.Bahadur: Classification into two multivariate normal distributions with unequal covariance matrices,AMS 33(1962),420-431.

3) W.R.Clarke,P.A.Lachenbruch: How non-normality affects the quadratic discriminant function, Communications in Statistics,Theor.Meth. A8(13)(1979), 1285-1301.

4) W.C.Cochran: Commentary on "Estimation of error rates in discriminant analysis",Technometrics 10(1968),204-206

5) N.R.Draper,I.Guttman,L.Lapczak: Actual rejection levels in a certain stepwise test,Commun.Statist.Theor.Meth. A8(2) (1979),99-105.

6) E.S.Gilbert: On discrimination using qualitative variables, JASA 63(1968), 1399-1407.

7) E.S.Gilbert: The effect of unequal covariance matrices on Fisher's linear discriminat funkction, Biometrics 25(1969),505-516.

8) J.D.F.Habbema,J.Hermans: Selection of variables in discriminant analysis, Technometrics 19(1977),487-493.

9) M.A.Hill: Anotated computer output for regression analysis,Technical rep.48, Health Computer Facility,University of California,Los Angeles,1979.

10) R.I.Jennrich: Stepwise discriminant analysis,in: K.Enslein,A.Ralston, H.S.Wilf(eds.): Statistical methods for digital computers,J.Wiley, New York,1977.

11) P.A.Lachenbruch: Discriminant analysis, Hafner Press, New York,1975,

12) P.A.Lachenbruch:,M.R.Mickey: Estimation of error rates in discriminant analysis, Technometrics 10(1968),1-11.

13) P.A.Lachenbruch: An almost unbiased methods for obtaining confidence intervals for the probability of misclassification in discriminant analysis, Biometrics 23(1967),639-645.

14) P.A.Lachenbruch,C.Sneering,L.T.Revo: Robustness of the linear and quadratic discriminant function to certain types of non-normality,Commun.Statist. 1(1973),39-57.

15) R.J.McKay: Simultaneous test procedures for variable selection in multiple discriminant analysis, Biometrika 64(1977),283-290.

16) S.Marks,O.J.Dunn: Discriminant functions when covariance matrices are unequal, JASA 69(1974),555-559.

17) G.J.McLachlan: Estimation of the errors of misclassification on the criterion of asymptotic mean square error, Technometrics 16(1974),255-260.

18) D.H.Moore: Evaluation of five discrimination procedures for binary variables, JASA 68(1973), 399-404.

19) G.D.Murray: A cautionary note on selection of variables in discriminant analysis, Applied Statistics 26(1977),246-250.

20) M.Okamoto: An asymptotic expression for the distribution of the linear discriminant function, AMS 34(1963),1286-1292.

21) A.C.Rensher, F.C.Pun: Inflation of R^2 in best subset regression, Technometrics 22(1980),49-53.

22) J.W.Van Ness,C.Simpson: On the effects of dimension in discriminant analysis, Technometrics 18(1976),175-187.