NEPARAMETRICKÁ DISKRIMINAČNÍ ANALÝZA

MARIE FORBELSKÁ

Abstrakt. In the paper the attention is focused to the application of kernel density estimators to statistical discrimination. After a brief description of the discriminant analysis problem the nonparametric approach to discriminant analysis is described. The multivariate product polynomial kernels with data-driven choices of the bandwidth are used for density estimators and this nonparametric approach are compared with classical one by some simulated data.

Резюме: Цель этой статьи касается приложения оценки плотности вероятности при помоши ядер в дискриминантном анализу. В статье сначала рассматриваются элементарные сведения по дискриминантному анализу и потом исследуется непараметрический подход при помоши многомерных полиномиальных ядер, построенных как произведение одномерных ядер, вместе с автоматическим выбором оптимального сглаживающего параметра. Параметрический и непараметрический подходы сровниваются при помоши имитирующих данных.

1. Podstata diskriminační analýzy

Uvažujme danou množinu n objektů, označme ji $\mathcal{S}$ a předpokládejme, že $\mathcal{S}$ je tvořená objekty k různých typů. Budeme říkat, že objekt patří do třídy $\mathcal{S}_j$, je-li typu j $(j = 1, \ldots, k)$. O třídách $\mathcal{S}_j$ budeme předpokládat, že jsou po dvou disjunktní a $\mathcal{S} = \bigcup_{j=1}^{k} \mathcal{S}_j$. Na každém objektu zjišťujeme dva statistické znaky X a $\mathbf{Y}$. X určuje příslušnost objektu do dané třídy, je to diskrétní náhodná veličina a $X = j$ právě když daný objekt patří do třídy $\mathcal{S}_j$, $j = 1, \ldots, k$. $\mathbf{Y} = (Y_1, \cdots, Y_m)'$ je m-rozměrný náhodný vektor, který nějak charakterizuje příslušnou třídu objektů. Označme dále $(X_i, \mathbf{Y}_i')'$ hodnoty znaků X a $\mathbf{Y}$ na i-tém objektu a předpokládejme, že $(X_i, \mathbf{Y}_i')'$ jsou nezávislé náhodné vektory, které tvoří náhodný výběr z rozdělení náhodného vektoru $(X, \mathbf{Y}')'$.

Cílem diskriminační analýzy je stanovit na základě daného náhodného výběru optimální klasifikační pravidlo, které by při pozorování vektoru $\mathbf{Y}$ na nějakém daném objektu, který již nepatří do třídy $\mathcal{S}$, umožnilo jeho zařazení do příslušné třídy s minimální ztrátou.

Při konstrukci takového klasifikačního pravidla vyjdeme z úplného rozkladu $\mathbb{S} = \{\mathbb{S}_1, \ldots, \mathbb{S}_k\}$ prostoru $\mathbb{R}^m$ možných hodnot vektoru $\mathbf{Y}$ do k disjunktních tříd $\mathbb{S}_1, \ldots, \mathbb{S}_k$. Když na uvažovaném objektu zjistíme hodnotu znaku $\mathbf{Y}$, která patří do třídy $\mathbb{S}_j$, rozhodneme, že tento objekt patří do třídy $\mathcal{S}_j$. S užitím tohoto klasifikačního pravidla spojíme ztrátu, která bude způsobena chybnou klasifikací objektu. Je-li daný objekt charakterizován vektorem $(X, \mathbf{Y}')'$ a máme-li klasifikační pravidlo dané rozkladem $\mathbb{S}$, pak příslušnou ztrátu definujeme jako transformovanou diskrétní

2000 *Mathematics Subject Classification.* Primary 62H30; Secondary 30C40.

Klíčová slova. Lineární a kvadratická diskriminační analýza, neparametrická diskriminační analýza, jádrové odhady hustot, součinová jádra.

Příspěvek vznikl s podporou výzkumného záměru MŠMT, CEZ: J07/98:143100001.

náhodnou veličinu $Z_{\mathbb{S}}$ danou předpisem

$$Z_{\mathbb{S}} = Z_{\mathbb{S}}(X, \mathbf{Y}) = z_{jl} \quad \text{pokud } X = j \text{ a } \mathbf{Y} \in \mathbb{S}_l \qquad l, j = 1, \ldots, k,$$

kde z_{jl} jsou daná reálná čísla, charakterizující reálnou ztrátu při zařazení objektu ze třídy $\mathcal{S}_j$ do třídy $\mathcal{S}_l$. V diskriminační analýze se často volí $z_{ll} = 0$ a $z_{jl} = 1$, $l, j = 1, \ldots, k;\ l \neq j$. Klasifikační pravidlo, které minimalizuje střední hodnotu ztráty, pak nazýváme **optimálním**.

Abychom odvodili optimální klasifikační pravidlo, vyjdeme z následujících předpokladů a značení.

Nechť náhodný vektor $(X, \mathbf{Y}')'$ definovaný na nějakém pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ má hustotu $f_{X\mathbf{Y}}(j, \mathbf{y})$ vzhledem k součinové míře $\mu = \nu_X \times \mu_{\mathbf{Y}}$, kde ν_X je čítací míra a $\mu_{\mathbf{Y}}$ je Lebesquova míra, přičemž tato hustota je tvaru $f_{X\mathbf{Y}}(j, \mathbf{y}) = p_j f_j(\mathbf{y}),\ j = 1, \ldots, k,\ \mathbf{y} \in \mathbb{R}^m,\ p_1 + \cdots + p_k = 1,\ p_j > 0$ a $f_j(\mathbf{y})$ pro každé $j = 1, \ldots, k$ je hustota rozdělení pravděpodobností vzhledem k Lebesquově míře. Zřejmě $f_j(\mathbf{y})$ je podmíněná hustota $\mathbf{Y}$, když $X = j$. Pak náhodná veličina X má marginální pravděpodobnostní funkci $P(X = j) = p_j\ (j = 1, \ldots, k)$ a marginální rozdělení pravděpodobností u náhodného vektoru $\mathbf{Y}$ je dáno hustotou $f_{\mathbf{Y}}(\mathbf{y}) = \sum_{j=1}^{k} p_j f_j(\mathbf{y}),\ \mathbf{y} \in \mathbb{R}^m$. Tedy v uvedeném značení má ztráta $Z_{\mathbb{S}}$ pravděpodobnostní funkci tvaru $p_{Z_{\mathbb{S}}}(j, l) = P(X = j, \mathbf{Y} \in \mathbb{S}_l) = \int_{\mathbb{S}_l} p_j f_j(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y})$, $l, j = 1, \ldots, k$. Snadno nahlédneme, že

$$\begin{aligned} E(Z_{\mathbb{S}}) = L_{\mathbb{S}} &= \sum_{j=1}^{k} \sum_{l=1}^{k} z_{jl} P(Z_{\mathbb{S}} = z_{jl}) = \sum_{j=1}^{k} \sum_{l=1}^{k} z_{jl} P(X = j, Y \in \mathbb{S}_l) = \\ &= \sum_{j=1}^{k} \sum_{l=1}^{k} z_{jl} p_{Z_{\mathbb{S}}}(j, l) = \sum_{j=1}^{k} \sum_{i=l}^{k} z_{jl} \int_{S_l} p_j f_j(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) = \\ &= \sum_{l=1}^{k} \sum_{j=1}^{k} z_{jl} \int_{S_l} p_j f_j(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) = \sum_{l=1}^{k} \int_{S_l} \sum_{j=1}^{k} z_{jl} p_j f_j(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) \end{aligned}$$

Funkci $q_l(\mathbf{y}) = \sum_{j=1}^{k} z_{jl} p_j f_j(\mathbf{y})$ nazveme ***l*-tý skór vektoru Y** a při konstrukci klasifikačního pravidla hraje centrální roli.

Cílem nyní je určit optimální úplný rozklad $\mathbb{S}^* = \{\mathbb{S}_1^*, \ldots, \mathbb{S}_k^*\}$ m-rozměrného euklidovského prostoru $\mathbb{R}^m$ tak, aby střední hodnota ztráty $E(Z_{\mathcal{S}^*})$ byla minimální Důležitou roli hraje následující lemma (viz. [1]).

Lemma 1.1. *Nechť $\mathbb{S}^* = \{\mathbb{S}_1^*, \ldots, \mathbb{S}_k^*\}$ je takový rozklad $\mathbb{R}^m$, že pro $\forall t \in \{1, \ldots, k\}$ platí*

$$\mathbf{y} \in \mathbb{S}_t^* \quad \Rightarrow \quad q_t(\mathbf{y}) \leq q_j(\mathbf{y}), \qquad j = 1, \ldots, k. \tag{1.1}$$

Pak tento rozklad minimalizuje $E(Z_{\mathbb{S}})$, tj.

označíme-li $L^* = E(Z_{\mathbb{S}^*}) = \sum_{i=1}^{k} \int_{\mathbb{S}_i^*} q_i(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y})$,

pak platí $L_{\mathbb{S}} = E(Z_{\mathbb{S}}) \geq L^* = E(Z_{\mathbb{S}^*})$ *pro každý rozklad* $\mathbb{S}$.

Důkaz. $L = E(Z_{\mathbb{S}}) = \sum_{l=1}^{k} \int_{\mathbb{S}_l} q_l(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) = \sum_{l=1}^{k} \sum_{t=1}^{k} \int_{\mathbb{S}_l \cap \mathbb{S}_t^*} q_l(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) \geq$
$\geq \sum_{l=1}^{k} \sum_{t=1}^{k} \int_{\mathbb{S}_l \cap \mathbb{S}_t^*} q_t(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) = \sum_{t=1}^{k} \int_{\mathbb{S}_t^*} q_t(\mathbf{y}) d\mu_{\mathbf{Y}}(\mathbf{y}) = E(Z_{\mathbb{S}^*}) = L^*$ □

Je zřejmé, že hodnota L^* je stejná pro všechny rozklady splňující podmínku předchozího lemmatu.

Z lemmatu 1.1 plyne, že **klasifikační pravidlo dané rozkladem (1.1) je optimální**. Pokud tedy při daném $\mathbf{Y} = \mathbf{y}$ pro všechna $j \neq t$ platí $q_t(\mathbf{y}) < q_j(\mathbf{y})$, pak optimálním rozhodnutím je zařadit daný objekt do t-té třídy. V případě, že v předchozím vzorci platí rovnost i pro další $j (j \neq t)$, je lhostejné, podle kterého pravidla budeme z těchto minimalizujících indexů vybírat.

Při **volbě** $\mathbf{z_{ll}=0}$ a $\mathbf{z_{jl}=1}$, $l,j=1,\dots,k,\ l\neq j$, kdy

$$q_l(\mathbf{y})=\sum_{j=1}^{k}p_jf_j(\mathbf{y})-p_lf_l(\mathbf{y})=f_{\mathbf{Y}}(\mathbf{y})-p_lf_l(\mathbf{y}),$$

snadno dostaneme další ekvivalentní **optimální klasifikační pravidlo** založené na nerovnosti

$$p_tf_t(\mathbf{y})\geq p_jf_j(\mathbf{y})\qquad \text{pro } j=1,\dots,k. \tag{1.2}$$

2. Rozhodovací pravidla v případě normálních rozdělení

V tomto odstavci budeme předpokládat, že podmíněné rozdělení náhodného vektoru $\mathbf{Y}$ za podmínky, že $X=j$, je m-rozměrné normální rozdělení $N_m(\boldsymbol{\mu}_j,\mathbf{V}_j)$ se známým vektorem středních hodnot $E(\mathbf{Y}|X=j)=\boldsymbol{\mu}_j$ a známou varianční maticí $var(Y|X=j)=\mathbf{V}_j$ $(j=1,\dots,k)$. Pak hustota tohoto podmíněného rozdělení je dána vzorcem

$$f_j(\mathbf{y})=(2\pi)^{-\frac{m}{2}}|\mathbf{V}_j|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{y}-\boldsymbol{\mu}_j)\right].$$

Klasifikační pravidlo (1.2) lze v tomto případě vyjádřit jako

$$\log p_t+\log f_t(\mathbf{y})>\log p_j+\log f_j(\mathbf{y}),\qquad j=1,\dots,k,\ j\neq t.$$

Označme $\quad D_j=-\frac{1}{2}\log|\mathbf{V}_j|-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{y}-\boldsymbol{\mu}_j)+\log p_j.$

Pak klasifikační pravidlo (1.2) odpovídá

$$D_t>D_j,\qquad j=1,\dots,k,\ j\neq t. \tag{2.1}$$

Diskriminační metoda založená na nerovnosti (2.1) se nazývá **kvadratická diskriminační analýza**.

Pokud jsou si **všechny varianční matice rovny**, tj. $\mathbf{V}_1=\dots=\mathbf{V}_k=\mathbf{V}$, potom

$$\begin{aligned}D_j&=-\tfrac{1}{2}\log|\mathbf{V}|-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_j)'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}_j)+\log p_j=\\&=-\tfrac{1}{2}\log|\mathbf{V}|+\log p_j-\tfrac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}+\mathbf{y}'\mathbf{V}^{-1}\boldsymbol{\mu}_j-\tfrac{1}{2}\boldsymbol{\mu}_j'\mathbf{V}^{-1}\boldsymbol{\mu}_j\end{aligned}$$

Označme $\quad d_j=\mathbf{y}'\mathbf{V}^{-1}\boldsymbol{\mu}_j-\frac{1}{2}\boldsymbol{\mu}_j'\mathbf{V}^{-1}\boldsymbol{\mu}_j+\log p_j.$

Pak $\quad D_j=d_j-\frac{1}{2}\log|\mathbf{V}|-\frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$

a klasifikační pravidlo (2.1) je v tomto speciálním případě ekvivalentní s nerovností

$$d_t>d_j,\qquad j=1,\dots,k,\ j\neq t. \tag{2.2}$$

Diskriminační metoda založená na nerovnosti (2.2) se nazývá **lineární diskriminační analýza**.

3. Diskriminace z experimentálních dat

Při praktickém provádění diskriminační analýzy máme k dispozici k souborů objektů, přičemž víme, který objekt do které třídy patří. Těmto souborům se někdy říká **trénovací**. Počet objektů v j-tém souboru označme n_j a realizace vektoru $\mathbf{Y}$ v j-tém souboru označme $\mathbf{y}_{j1},\dots,\mathbf{y}_{jn_j}$. Mějme dále realizaci $\mathbf{y}\in\mathbb{R}^m$ náhodného vektoru $\mathbf{Y}$, o které nevíme, odkud pochází.

Protože obvykle neznáme rozdělení náhodného vektoru $(X,\mathbf{Y}')'$, pak se při klasifikaci neznámého objektu nabízí dva možné přístupy :

Parametrický přístup: Předpokládáme, že podmíněné rozdělení náhodného vektoru $\mathbf{Y}$ za podmínky, že $X = j$, je normální. V konkrétních situacích obvykle neznáme vektory středních hodnot $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ a varianční matice $V_1, \ldots, V_k$. K dispozici však máme trénovací soubory a pomocí nich určíme $\bar{\mathbf{y}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{y}_{ji}$, $\mathbf{C}_j = \sum_{i=1}^{n_j} (\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)'$, $\hat{p}_j = \frac{n_j}{n_1 + \cdots + n_k}$. Jestliže vektor $\boldsymbol{\mu}_j$ odhadneme vektorem $\bar{\mathbf{y}}_j$, matici $\mathbf{V}_j$ maticí $\hat{\mathbf{V}}_j = \frac{1}{n_j - 1} \mathbf{C}_j$ a apriorní pravděpodobnost p_j relativní četností $\hat{p}_j$, můžeme pro zařazení objektu, jehož příslušnost nepoznáme, použít postupy předchozího odstavce tak, že neznámé parametry nahradíme jejich odhady.

Neparametrický přístup: Nebudeme předpokládat určitý typ podmíněného rozdělení vektoru $\mathbf{Y}$ za podmínky, že $X{=}j$, ale pomocí trénovacích dat odhadneme neznámé podmíněné hustoty $f_j(\mathbf{y})$. Přirozeně se nabízí použít neparametrické metody odhadu hustot, např. jádrové odhady hustoty, odhady hustoty pomocí k nejbližších sousedů a další (viz. [7]).

Pro zařazení objektu, jehož příslušnost nepoznáme, potom použijeme **rozhodovací pravidlo** (1.2) s tím, že neznámé podmíněné hustoty $f_j(\mathbf{y})$ nahradíme neparametrickým odhadem $\hat{f}_j(\mathbf{y})$ a apriorní pravděpodobnost p_j relativní četností $\hat{p}_j = \frac{n_j}{n_1 + \cdots + n_k}$. Dostaneme tak **rozhodovací pravidlo**: realizaci $\mathbf{y}$ zařadíme do t-té skupiny, pokud pro všechna $j \neq t$ bude platit $\hat{p}_t \hat{f}_t(\mathbf{y}) \geq \hat{p}_j \hat{f}_j(\mathbf{y})$.

Obvykle, před zařazováním nových objektů, ověříme klasifikační proceduru na samotných objektech z trénovacích souborů a registrujeme procento nesprávných zatřídění. Jestliže soubor trénovacích dat neumožňuje vytvořit spolehlivou klasifikační proceduru ani pro trénovací data samotná, nelze samozřejmě klasifikaci realizovat.

4. Jádrové odhady použité v neparametrické diskriminační analýze

V tomto odstavci zavedeme jednorozměrná (resp. vícerozměrná) jádra pro odhad hustoty pravděpodobnosti náhodných veličin (resp. náhodných vektorů) a popíšeme speciální typy jader, která budou použita pro neparametrickou diskriminaci.

Nechť $y_1, \ldots, y_n$ jsou nezávislá pozorování náhodné veličiny Y s hustotou $f(y)$. **Jádrem** rozumíme libovolnou funkci $K : (R, \mathcal{B}) \to (0, +\infty)$, jež je symetrická, ohraničená a pro niž

$$\int_{-\infty}^{\infty} K(y)dy = 1, \quad a \lim_{y \to \pm\infty} |y| K(y) = 0, \tag{4.1}$$

Nechť $\{h_n\}_{n=1}^{\infty}$ je posloupnost kladných čísel taková, že $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n = \infty$ a $K(y)$ je některé jádro.

Jádrový odhad hustoty je definován vztahem (viz. [2] a [5]).

$$\hat{f}_n(y) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h_n}\right) \quad y \in \mathbb{R}. \tag{4.2}$$

Velká pozornost musí být věnována **volbě nejvhodnější konstanty** h_n, tzv. *šířce okna*, neboť podstatným způsobem ovlivňuje kvalitu odhadu. Pro optimální volbu parametru h_n je třeba provést také odhad derivace funkce f (viz. [7] a [8]).

Symbolem $\mathcal{C}^{k_0}$ označme množinu všech k_0-krát spojitě diferencovatelných reálných funkcí, kde $k_0 > 0$ je celé číslo. Jsou-li navíc tyto funkce nulové vně intervalu $[-1, 1]$, označme množinu těchto funkcí symbolem $\mathcal{C}^{k_0}[-1, 1]$.

Nechť $y_1, \ldots, y_n$ jsou nezávislá pozorování náhodné veličiny Y s hustotou $f(y) \in \mathcal{C}^{k_0}$. Jádrový odhad **derivace** $f^{(\nu)}$ pro pevné $0 \leq \nu < k_0$ je definován

vztahem

$$\hat{f}^{(\nu)}_{h,K}(y) = \frac{1}{nh^{\nu+1}} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right). \tag{4.3}$$

Označme $Lip[a,b]$ třídu spojitých funkcí splňujících Lipschitzovu podmínku na $[a,b]$:

$$|f(x) - f(y)| \le L|x - y| \quad \forall x, y \in [a,b], \quad L > 0.$$

Nechť ν, k jsou nezáporná celá čísla, $0 \le \nu < k < k_0$ a jádro $K \in Lip[-1,1]$, přičemž nosič jádra $support(K) \subseteq [-1,1]$. Nechť K splňuje následující momentové podmínky

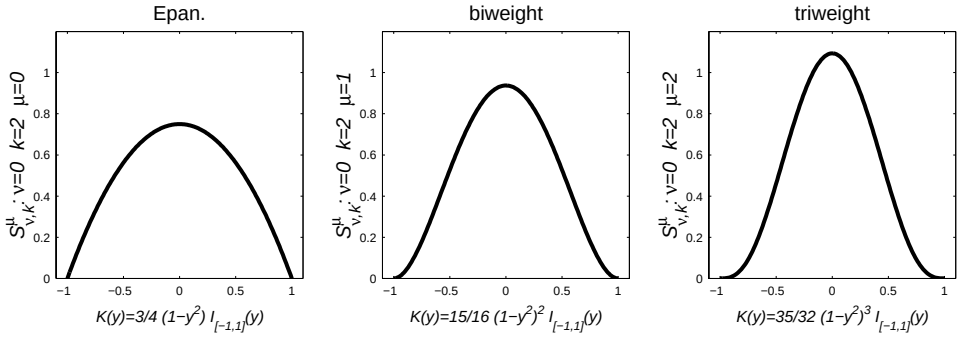
$$\int_{-1}^{1} x^j K(x)dx = \begin{cases} 0 & 0 \le j < k, j \ne \nu \\ (-1)^\nu \nu! & j = \nu \\ \beta_k \ne 0 & j = k \end{cases} \tag{4.4}$$

pak říkáme, že **jádro K je řádu** (ν, k) a píšeme $K \in \mathcal{S}^0_{\nu,k}$.
Pro $\mu \ge 1$ nechť $K \in \mathcal{C}^\mu[-1,1]$, $K \in \mathcal{S}^0_{\nu,k}$. Navíc nechť platí $K^{(j)}(1) = K^{(j)}(-1) = 0$, $j = 0, 1, \ldots, \mu - 1$, $0 \le \nu \le k - 2$ a $\nu + k$ je sudé. Pak takové jádro se nazývá **jádro hladkosti** μ a píšeme $K \in \mathcal{S}^\mu_{\nu,k}$. Příkladem jádra $\mathcal{S}^0_{0,2}$ je Epanečnikovo jádro, $\mathcal{S}^1_{0,2}$ kvartické (biweight) jádro a $\mathcal{S}^2_{0,2}$ triweight jádro (viz. [3]).

V práci použijeme jádra :

Epanečnikovo	$K(y) = \frac{3}{4}\left(1 - y^2\right) I_{[-1,1]}(y)$
kvartické(biweight)	$K(y) = \frac{15}{16}(1 - y^2)^2 I_{[-1,1]}(y)$
triweight	$K(y) = \frac{35}{32}(1 - y^2)^3 I_{[-1,1]}(y)$

kde $I_{[a,b]}(y) = \begin{cases} 1 & y \in [a,b] \\ 0 & jinak \end{cases}$ (viz. obrázek 1.)



Obrázek 1: Ukázka jader typu $K \in \mathcal{S}^\mu_{\nu,k}$

Pro optimální volbu šířky okna tohoto typu jader použijeme algoritmus, který je popsán v práci [4].

Pro odhad hustoty pravděpodobnosti náhodných vektorů jsou definovány **vícerozměrné jádrové odhady** vztahem

$$\hat{f}_n(\mathbf{y}) = \frac{1}{nh_1 \ldots h_m} \sum_{i=1}^{n} K\left(\frac{y_1 - y_{i1}}{h_1}, \cdots, \frac{y_m - y_{im}}{h_m}\right), \tag{4.5}$$

kde $\mathbf{y}_1 = (y_{11}, \ldots, y_{1m}), \ldots, \mathbf{y}_n = (y_{n1}, \ldots, y_{nm})$ je náhodný výběr z m-rozměrného spojitého rozdělení o hustotě $f(\mathbf{y})$, $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$.

V dalším budeme používat jako jádro m-proměnných tzv. *součinové jádro,* které je součinem m jader jedné proměnné, tj.

$$\hat{f}_n(\mathbf{y}) = \frac{1}{nh_1 \dots h_m} \sum_{i=1}^{n} \prod_{j=1}^{m} K\left(\frac{y_j - y_{ij}}{h_j}\right), \tag{4.6}$$

kde $K \in \mathcal{S}^{\mu}_{\nu,k}$, přičemž opět využijeme algoritmus automatického vyhledávání optimální šířky oken pro tento typ jader (viz. [4]).

5. Srovnání parametrické a neparametrické diskriminace

Srovnání parametrické a neparametrické diskriminace je provedeno na simulovaných datech ze směsi **normálních rozdělení**:

$$f_a(y_1, y_2) = \frac{1}{k} \sum_{j=1}^{k} f_{aj}(y_1, y_2) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{2\pi\sigma_{j1}\sigma_{j2}\sqrt{1-\rho_j^2}} \exp\left(-\frac{1}{2} q_j(y_1, y_2)\right) \tag{5.1}$$

kde

$$q_j(y_1, y_2) = \frac{1}{1-\rho_j^2} \left[\left(\frac{y_1 - \mu_{j1}}{\sigma_{j1}}\right)^2 - 2\rho_j \left(\frac{y_1 - \mu_{j1}}{\sigma_{j1}}\right)\left(\frac{y_2 - \mu_{j2}}{\sigma_{j2}}\right) + \left(\frac{y_2 - \mu_{j2}}{\sigma_{j2}}\right)^2 \right]$$

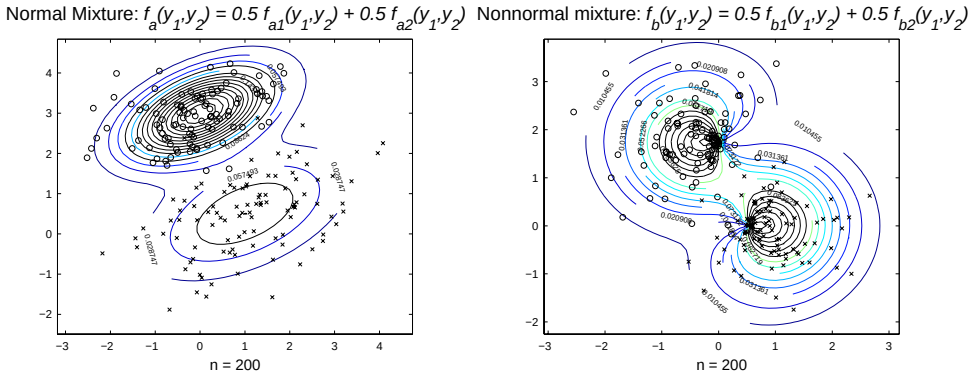
a ze směsi **hustot**:

$$f_b(y_1, y_2) = \frac{1}{k} \sum_{j=1}^{k} f_{bj}(y_1, y_2) \tag{5.2}$$

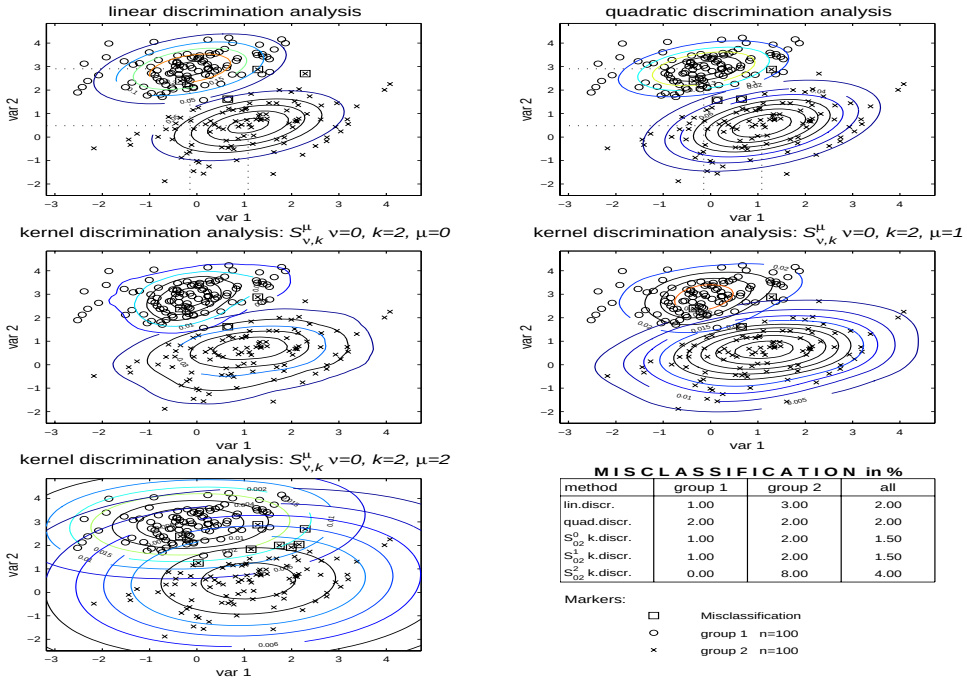
kde

$$f_{bj}(y_1, y_2) = \frac{1}{2\pi} \left(1 + \frac{2\rho_j (y_1 - \mu_{j1})}{\sqrt{(y_1 - \mu_{j1})^2 + (y_2 - \mu_{j2})^2}}\right) \exp\left(-\frac{(y_1 - \mu_{j1})^2 + (y_2 - \mu_{j2})^2}{2}\right).$$

Příklady směsí typu (5.1) a (5.2) pro $k = 2$ jsou uvedeny na obrázku 2 a výsledky diskriminace těchto směsí jsou demonstrovány na obrázcích 3 a 4. Pro generování pseudonáhodných čísel z rozdělení typu (5.2) byl použit algoritmus doporučený v [6].



Obrázek 2: **Simulovaná data spolu s vrstevnicovými grafy funkcí $\mathbf{f_a}(\mathbf{x}, \mathbf{y})$ a $\mathbf{f_b}(\mathbf{x}, \mathbf{y})$** s parametry: μ^a_{11}=0; μ^a_{12}=3; σ^2_{11}=1; σ^2_{12}=0.5; ρ^a_1=0.5; μ^a_{21}=1; μ^a_{21}=0.5; σ^2_{21}=2; σ^2_{22}=1; ρ^a_2=0.5 a μ^b_{11}=0; μ^b_{12}=1.75; ρ^b_1=−0.5; μ^b_{21}=0.5; μ^b_{22}=0; ρ^b_2=0.5.

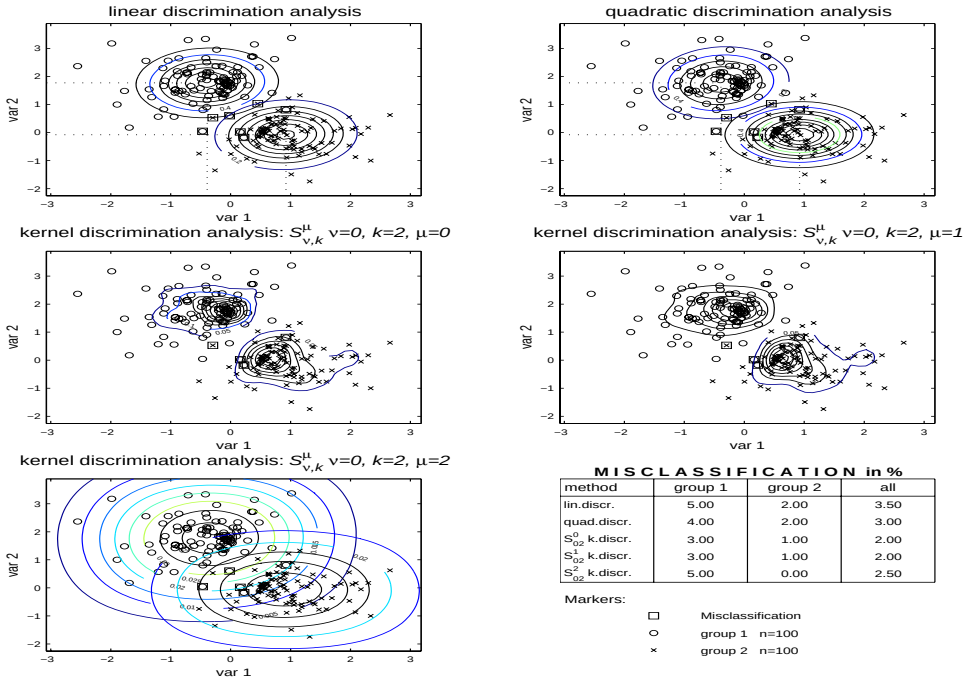


MISCLASSIFICATION in %

method	group 1	group 2	all
lin.discr.	1.00	3.00	2.00
quad.discr.	2.00	2.00	2.00
S^0_{02} k.discr.	1.00	2.00	1.50
S^1_{02} k.discr.	1.00	2.00	1.50
S^2_{02} k.discr.	0.00	8.00	4.00

Markers:
Misclassification
group 1 n=100
group 2 n=100

Obrázek 3: **Srovnání parametrické a neparametrické diskriminace s užitím jader typu** $K \in \mathcal{S}^{\mu}_{\nu,k}(\nu{=}0, k{=}2, \mu{=}0, 1, 2)$ pro simulovaná data ze směsi $f_a(x, y)$ normálních rozdělení .



MISCLASSIFICATION in %

method	group 1	group 2	all
lin.discr.	5.00	2.00	3.50
quad.discr.	4.00	2.00	3.00
S^0_{02} k.discr.	3.00	1.00	2.00
S^1_{02} k.discr.	3.00	1.00	2.00
S^2_{02} k.discr.	5.00	0.00	2.50

Markers:
Misclassification
group 1 n=100
group 2 n=100

Obrázek 4: **Srovnání parametrické a neparametrické diskriminace s užitím jader typu** $K \in \mathcal{S}^{\mu}_{\nu,k}(\nu{=}0, k{=}2, \mu{=}0, 1, 2)$ pro simulovaná data ze směsi $f_b(x, y)$.

Bylo provedeno 24 simulací normálních směsí typu (5.1) (viz. řádky 1 až 24 na obrázku 5) a 18 simulací směsí hustot typu (5.2) (viz. řádky 25 až 42 na obrázku 5), kdy se měnily parametry polohy a měřítka hustot, velikost a počet skupin.

Normal and Nonnormal Mixtures

	Group 1						Group 2						Group 3						Misclassification in %				
	μ_1		V_1				μ_2		V_2				μ_3		V_3				class.methods		kernel methods		
	μ_{11}	μ_{12}	σ^2_{11}	σ^2_{12}	ρ_1	n_1	μ_{21}	μ_{22}	σ^2_{21}	σ^2_{22}	ρ_2	n_2	μ_{31}	μ_{32}	σ^2_{31}	σ^2_{32}	ρ_3	n_3	lin.	quad.	S^0_{02}	S^1_{02}	S^2_{02}
1	0	3	1	0.5	0.5	30	1	0	2	1	0.5	30							3.33	3.33	1.67	1.67	1.67
2	0	3	1	0.5	0.5	50	1	0	2	1	0.5	50							1.00	1.00	1.00	1.00	1.00
3	0	3	1	0.5	0.5	70	1	0	2	1	0.5	70							1.43	0.00	0.00	0.71	0.71
4	0	3	1	0.5	0.5	100	1	0	2	1	0.5	100							1.50	0.50	0.50	0.50	0.50
5	0	3	1	0.5	0.5	150	1	0	2	1	0.5	150							1.67	2.00	1.33	2.00	1.33
6	0	3	1	0.5	0.5	200	1	0	2	1	0.5	200							0.75	0.75	0.75	0.75	2.00
7	0	3	1	0.5	0.5	30	1	0.5	2	1	0.5	30							3.33	0.00	0.00	0.00	1.67
8	0	3	1	0.5	0.5	50	1	0.5	2	1	0.5	50							4.00	1.00	1.00	2.00	2.00
9	0	3	1	0.5	0.5	70	1	0.5	2	1	0.5	70							1.43	1.43	1.43	1.43	2.14
10	0	3	1	0.5	0.5	100	1	0.5	2	1	0.5	100							2.50	2.00	2.50	2.50	2.00
11	0	3	1	0.5	0.5	150	1	0.5	2	1	0.5	150							2.67	2.00	2.67	2.33	2.33
12	0	3	1	0.5	0.5	200	1	0.5	2	1	0.5	200							3.25	2.75	2.75	2.50	10.25
13	0	3	1	0.5	0.5	30	1	0	2	1	0.5	30	3	3	0.7	0.3	−0.5	30	11.11	6.67	5.56	6.67	6.67
14	0	3	1	0.5	0.5	50	1	0	2	1	0.5	50	3	3	0.7	0.3	−0.5	50	8.00	6.67	6.67	6.67	8.00
15	0	3	1	0.5	0.5	70	1	0	2	1	0.5	70	3	3	0.7	0.3	−0.5	70	8.10	6.67	5.71	6.19	7.62
16	0	3	1	0.5	0.5	100	1	0	2	1	0.5	100	3	3	0.7	0.3	−0.5	100	8.00	7.67	8.00	8.33	8.67
17	0	3	1	0.5	0.5	150	1	0	2	1	0.5	150	3	3	0.7	0.3	−0.5	150	6.67	5.56	5.56	5.78	6.44
18	0	3	1	0.5	0.5	200	1	0	2	1	0.5	200	3	3	0.7	0.3	−0.5	200	9.17	7.83	7.67	7.67	15.50
19	0	3	1	0.5	0.25	30	1	0.5	2	1	0.75	30	3	3	0.7	0.3	−0.5	30	8.89	7.78	3.33	6.67	8.89
20	0	3	1	0.5	0.25	50	1	0.5	2	1	0.75	50	3	3	0.7	0.3	−0.5	50	13.33	6.67	8.67	9.33	9.33
21	0	3	1	0.5	0.25	70	1	0.5	2	1	0.75	70	3	3	0.7	0.3	−0.5	70	10.00	6.19	5.24	7.62	8.10
22	0	3	1	0.5	0.25	100	1	0.5	2	1	0.75	100	3	3	0.7	0.3	−0.5	100	7.33	6.67	5.67	6.00	6.67
23	0	3	1	0.5	0.25	150	1	0.5	2	1	0.75	150	3	3	0.7	0.3	−0.5	150	8.22	6.89	6.67	6.67	6.67
24	0	3	1	0.5	0.25	200	1	0.5	2	1	0.75	200	3	3	0.7	0.3	−0.5	200	8.00	7.00	7.17	7.17	10.17
25	0	3			−0.5	30	1	0			0.5	30							0.00	0.00	0.00	0.00	0.00
26	0	3			−0.5	50	1	0			0.5	50							0.00	0.00	0.00	1.00	1.00
27	0	3			−0.5	70	1	0			0.5	70							2.14	2.14	0.00	2.14	1.43
28	0	3			−0.5	100	1	0			0.5	100							1.50	1.00	0.50	1.00	1.00
29	0	3			−0.5	150	1	0			0.5	150							0.00	0.33	0.00	0.00	0.00
30	0	3			−0.5	200	1	0			0.5	200							1.00	1.00	0.75	0.75	1.00
31	0	2.5			−0.5	30	0.5	0			0.5	30							1.67	3.33	1.67	1.67	1.67
32	0	2.5			−0.5	50	0.5	0			0.5	50							1.00	1.00	1.00	1.00	3.00
33	0	2.5			−0.5	70	0.5	0			0.5	70							3.57	3.57	2.14	3.57	4.29
34	0	2.5			−0.5	100	0.5	0			0.5	100							3.00	2.50	1.50	2.00	2.00
35	0	2.5			−0.5	150	0.5	0			0.5	150							2.67	2.00	2.00	2.67	2.67
36	0	2.5			−0.5	200	0.5	0			0.5	200							3.00	3.00	2.50	2.50	2.75
37	0	1.75			−0.5	30	0.5	0			0.5	30							6.67	6.67	3.33	6.67	6.67
38	0	1.75			−0.5	50	0.5	0			0.5	50							6.00	6.00	5.00	7.00	7.00
39	0	1.75			−0.5	70	0.5	0			0.5	70							6.43	7.14	5.00	6.43	7.14
40	0	1.75			−0.5	100	0.5	0			0.5	100							5.00	5.00	3.50	5.50	5.00
41	0	1.75			−0.5	150	0.5	0			0.5	150							4.33	4.33	2.67	5.00	5.67
42	0	1.75			−0.5	200	0.5	0			0.5	200							4.75	4.50	3.00	3.00	3.75

Obrázek 5: **Tabulka parametrů simulovaných dat** spolu s celkovým procentem nesprávně klasifikovaných objektů pro klasické i neparametrické metody diskriminace.

Výsledky srovnání neparametrické diskriminace s lineární a kvadratickou diskriminací jsou uvedeny v tabulkách na obrázku 6, kde znaménka po řadě "+", "=" a "-" značí lepší, stejné a horší výsledky neparametrické diskriminace vůči klasickým metodám na základě celkového procenta špatně klasifikovaných objektů.

Pomocí simulací se ukázalo, že neparametrická diskriminace založená na jádrech $\mathcal{S}^0_{0,2}$ dává nejlepší výsledky, o něco horší neparametrická diskriminace založená na jádrech $\mathcal{S}^1_{0,2}$ a výrazně horší výsledky dosahuje neparametrická diskriminace založená na jádrech $\mathcal{S}^2_{0,2}$ (v důsledku příliš širokých vyhlazovacích oken poskytnutých algoritmem popsáným v práci [4]).

Pro tyto prvotní simulace se tedy ukazuje, že neparametrická diskriminace, tak jak je popsaná v předchozím odstavci, může být srovnatelnou náhradou klasické diskriminace dokonce i v případě normálních směsí a může být užitečná v situacích, kdy není dostatečná informace o typu rozdělení ve směsi.

Normal Mixtures

method	lin.discr.			quad.discr.		
	+	=	−	+	=	−
S^0_{02} k.discr.	75.00	12.50	12.50	37.50	41.67	20.83
S^1_{02} k.discr.	75.00	16.67	8.33	29.17	33.33	37.50
S^2_{02} k.discr.	62.50	4.17	33.33	12.50	20.83	66.67

Nonnormal Mixtures

method	lin.discr.			quad.discr.		
	+	=	−	+	=	−
S^0_{02} k.discr.	72.22	16.67	11.11	77.78	22.22	0.00
S^1_{02} k.discr.	27.78	27.78	44.44	38.89	33.33	27.78
S^2_{02} k.discr.	27.78	27.78	44.44	33.33	33.33	33.33

Obrázek 6: **Tabulky srovnání parametrické a neparametrické diskriminace** (hodnoty jsou uvedeny v %).

Literatura

[1] Anděl, J.: *Matematická statistika*. SNTL/ALFA. Praha 1978

[2] Antoch, J.,Vorlíčková, D.: *Vybrané metody statistické analýzy dat*. Academia, Praha 1992

[3] Horová, I.: *Optimatization Problems Connected with Kernel Smoothing, Signal Processing, Communications and Computer Science World*. Scientific and Engineering Press 2000, str. 339-445.

[4] Horová, I., Vieu, P. Zelinka, J.: *Optimal Choice of Nonparametric Estimates of a Density and of its Derivates*, zasláno k tisku

[5] Michálek, J.:*Kernel estimators - basic properties and optimal choice of parameters for estimation*. Proceedings ROBUST 94. Prague, 1994.

[6] Nachtsheim, Ch.,J, Johnson, M.,E.:*A New Family of Multivariate Distributions With Applications to Monte Carlo Studies*. Journal of the American Statistical Association, Volume 83, Issue 404 (Dec.,1988), 984-989

[7] Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1993.

[8] Wand, I.P. and Jones, I.C.: *Kernel Smoothing*. Chapman & Hall, London 1995

MU PřF, KAM, Janáčkovo nám. 2a, 662 95 Brno

E-mail: forbel@math.muni.cz