Masarykova univerzita • Přírodovědecká fakulta

Jádrové odhady a binární data

Bakalářská práce

BRNO 15. května 2006 Jan Orava

Prohlašuji, že jsem tuto práci vypracoval sám a čerpal jsem pouze z materiálů uvedených v seznamu literatury.

Tímto chci poděkovat paní RNDr. Marii Forbelské, Ph.D. za odborné vedení, cenné rady a připomínky k bakalářské práci, stejně tak za čas, který mi věnovala.

Obsah

Seznam použitého značení

$\mathbb{N}$	množina všech přirozených čísel
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{R}^k$	reálný k-rozměrný euklidovský prostor
$\mathbf{X}$	náhodný výběr
$\mathbf{x}$	realizace náhodného výběru $\mathbf{X}$
$\mathcal{B}$	σ-algebra borelovských množin
$[a, b]$	uzavřený interval
(a, b)	otevřený interval
$(\Omega, \mathcal{A}, \mathsf{P})$	pravděpodobnostní prostor
$\mathsf{P}(A)$	pravděpodobnost jevu A
$\mathsf{E}X$	střední hodnota náhodné veličiny X
$\{x_n\}_{i=1}^n$	posloupnost
$A(\pi)$	alternativní rozdělelní
$I(x_i - x_k)$	indikátorová funkce
$f(x)$	hustota náhodné veličiny x
$\hat{f}(x)$	odhad hustoty náhodné veličiny x
$p(x)$	pravděpodobnostní funkce náhodné veličiny x
$\hat{p}(x)$	odhad pravděpodobnostní funkce náhodné veličiny x
$\mathsf{K}(x)$	jádrová funkce
K_b	binární jádro
K_b^m	m-rozměrné binární jádro
K_{nk}^m	m-rozměrné jádro pro neuspořádané kategoriální data
K_{uk}	jádro pro uspořádané kategoriální data
h	šířka vyhlazovacího okna spojitého jádra
λ	vyhlazovací parametr diskrétního jádra

Úvod

V matematické statistice jsme často postaveni před úlohu, kdy na základě naměřených hodnot musíme určit pravděpodobnostní rozdělení náhodné veličiny.

Jednou z metod odhadu pravděpodobnostní funkce je neparametrický odhad pomocí jader. Jádrový odhad pravděpodobnostní funkce je poměrně mladá disciplína, k podrobnému zkoumání této problematiky došlo teprve v posledních 30 letech. Přestože odhad pomocí jádrových odhadů je poměrně jednoduchý, teorie zabývající se vhodnou volbou paramterů je dosti složitá.

Úvodní kapitola je pouhým připomenutím základních pojmů z pravděpodobnosti a matematické statistiky.

V druhé kapitole jsou definovány *jádrové odhady hustoty*. Zavádíme zde odhady hustoty náhodné veličiny a náhodného vektoru pomocí funkcí zvaných jádra.

Třetí kapitola je věnována hlavní části - *jádrovým odhadům pravděpodobnostní funkce*. V kapitole je odvozen jádrový odhad pravděpodobnostní funkce náhodné veličiny a náhodného vektoru s diskrétním rozdělením pravděpodobnosti. Pozornost také věnujeme vhodné volbě vyhlazovacího parametru.

V poslední kapitole jsou uvedeny ilustrativní příklady jádrových odhadů pravděpodobností funkce binárního náhodného vektoru.

V příloze jsou uvedeny zdrojové texty programů v MATLABU použitých při výpočtu ilustrativních příkladů.

K práci je připojen seznam literatury zabývající se problematikou jádrových odhadů.

Bakalářská práce byla vysázena v systému LaTeX 2_ε.

Kapitola 1

Základní pojmy

1.1 Základní pojmy z pravděpodobnosti

Nejdříve si připomeňme některé pojmy z pravděpodobnosti.

Definice 1.1. Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Reálnou funkci $X : \Omega \to \mathbb{R}$ nazveme *náhodnou veličinou*, jestliže pro každé $x \in \mathbb{R}$ platí

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}. \tag{1.1}$$

Množinu $\mathsf{M} \subset \mathbb{R}$ všech hodnot náhodné veličiny X nazýváme *obor hodnot náhodné veličiny* X.

Definice 1.2. Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$. Reálná funkce F_X definovaná na $\mathbb{R}$ předpisem

$$F_X(x) = \mathsf{P}(X \leq x), \qquad x \in \mathbb{R}, \tag{1.2}$$

se nazývá *distribuční funkce náhodné veličiny* X.

Definice 1.3. Náhodná veličina X s distribuční funkcí $F = F_X(x)$ se nazývá *diskrétní náhodná veličina*, jestliže existuje neprázdná nejvýše spočetná podmnožina $\mathbb{M}$ množiny reálných čísel $\mathbb{R}$ a funkce $p : \mathbb{R} \to \mathbb{R}$ s následujícími vlastnostmi:

$$p(t) > 0 \text{ pro } \forall t \in \mathbb{M}, \tag{1.3}$$

$$p(t) = 0 \text{ pro } \forall t \in \mathbb{R} - \mathbb{M}, \tag{1.4}$$

$$\sum_{t \in \mathbb{M}} p(t) = 1 = \sum_{t \in \mathbb{R}} p(t) = 1 \tag{1.5}$$

a pro $\forall x \in \mathbb{R}$ platí

$$F_X(x) = \sum_{t \leq x} p(t). \tag{1.6}$$

Funkce p se nazývá *pravděpodobnostní funkce* náhodné veličiny X. Říkáme, že náhodná veličina X má *diskrétní rozdělení pravděpodobnosti.*

Definice 1.4. Náhodná veličina X s distribuční funkcí $F = F_X(x)$ je (absolutně) spojitá, jestliže existuje nezáporná funkce s vlastnostmi:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1, \quad F(x) = \int_{-\infty}^{\infty} f(t)\mathrm{d}t. \tag{1.7}$$

Funkce $f(t)$ se nazývá *hustota náhodné veličiny* X.

Definice 1.5. Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$. Uspořádanou m-tici $\mathbf{X} = (X_1, \ldots, X_m)'$ náhodných veličin $X_1, \ldots, X_m$ definovaných na $(\Omega, \mathcal{A}, \mathsf{P})$ nazýváme *(m-rozměrný) náhodný vektor.*

Definice 1.6. Nechť $\mathbf{X} = (X_1, \ldots, X_m)'$ je náhodný vektor definovaný na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$. *Distribuční funkce náhodného vektoru* $\mathbf{X}$ je reálná funkce $F_{\mathbf{X}}$ definovaná na R^m vztahem

$$\begin{aligned} F_{\mathbf{X}}(x_1, \ldots, x_m) &= \mathsf{P}(X_1 \leq x_1, \ldots, X_m \leq x_m) = \mathsf{P}(\mathbf{X} \leq \mathbf{x}), \\ \mathbf{x} &= (x_1, \ldots, x_m)' \in R^m. \end{aligned} \tag{1.8}$$

1.7 Věta (o vlastnostech distribuční funkce). *Distribuční funkce* $F_{\mathbf{X}}(x_1, \ldots, x_m)$ *m-rozměrného náhodného vektoru* $\mathbf{X}$ *má tyto vlastosti:*

1. $\lim\limits_{\forall i\, x_i \to \infty} F_{\mathbf{X}}(x_1, \ldots, x_m) = 1, \qquad \lim\limits_{\exists i\, x_i \to -\infty} F_{\mathbf{X}}(x_1, \ldots, x_m) = 0.$
2. $F_{\mathbf{X}}(x_1, \ldots, x_m)$ *je zprava spojitá v každé proměnné (při pevných hodnotách ostatních* $n - 1$ *proměnných).*
3. $F_{\mathbf{X}}(x_1, \ldots, x_m)$ *je neklesající funkcí každé své proměnné (při pevně daných hodnotách ostatních* $m - 1$ *proměnných).*

Důkaz. viz [9] . □

Definice 1.8. Nechť $\mathbf{X}$ je m-rozměrný náhodný vektor. Náhodný vektor $\mathbf{X}$ se nazývá *diskrétní*, jestliže existuje neprázdná nejvýše spočetná podmnožina $\mathbb{M}$ množiny $\mathbb{R}^m$ a zobrazení $p : \mathbb{M} \to \mathbb{R}$ tak, že platí:

$$p(\mathbf{x}) > 0 \; pro \; \forall \mathbf{x} \in \mathbb{M}, \tag{1.9}$$

$$p(\mathbf{x}) = 0 \; pro \; \forall \mathbf{x} \in \mathbb{R}^m - \mathbb{M}, \tag{1.10}$$

$$\sum_{\mathbf{x} \in \mathbb{M}} p(\mathbf{x}) = 1 \tag{1.11}$$

a pro distribuční funkci $F_{\mathbf{X}}$ náhodné veličiny X máme

$$F_{\mathbf{X}}(x_1, \ldots, x_m) = \sum_{t_1 \leq x_1} \cdots \sum_{t_m \leq x_m} p(t_1, \ldots, t_m), \tag{1.12}$$

kde $x_1, \ldots, x_m, t_1, \ldots, t_m \in \mathbb{R}$.

Funkce p se nazývá *pravděpodobnostní funkce náhodného vektoru* $\mathbf{X}$. Říkáme také, že náhodný vektor $\mathbf{X}$ má *diskrétní rozdělení.*

Definice 1.9. Nechť $\mathbf{X}$ je m-rozměrný náhodný vektor. Náhodný vektor $\mathbf{X}$ se nazývá *(absolutně) spojitý*, jestliže existuje nezáporná funkce $f(x_1, \ldots, x_m) : \mathbb{R}^m \to \mathbb{R}$ taková, že

$$\underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{m} f(t_1, \cdots, t_m) \mathrm{d}t_1 \cdots \mathrm{d}t_m = 1 \tag{1.13}$$

a pro distribuční funkci $F_{\mathbf{X}}$ náhodného vektoru $\mathbf{X}$ máme

$$F_{\mathbf{X}}(x_1, \ldots, x_m) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f(t_1, \cdots, t_m) \mathrm{d}t_1 \cdots \mathrm{d}t_m \tag{1.14}$$

pro $\forall x_1, \ldots, x_m \in \mathbb{R}$. Funkce f se nazývá *hustota.*

1.2 Základní pojmy matematické statistiky

V matematické statistice máme k dispozici výsledky n nezávislých pozorování hodnot sledované náhodné veličiny

$$x_1 = X(\omega_1), \ldots, x_n = X(\omega_n), \quad \omega_i \in \Omega, \; \forall i = 1, \ldots, n$$

a chceme učinit výpověď o rozdělení pravděpodobnosti této zkoumané náhodné veličiny.

Definice 1.10. Náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$, jehož složky jsou nezávislé náhodné veličiny, které mají stejné rozdělení pravděpodobnosti jako zkoumaná náhodná veličina X, se nazývá *náhodný výběr* rozsahu n.

Množinu všech hodnot, kterých může náhodný vektor nabýt, nazýváme výběrový prostor.

Libovolný bod $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ budeme nazývat realizací náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)$.

Kapitola 2

Jádrové odhady hustoty náhodného vektoru

Nejznámější a nejlépe prostudovanou třídou neparametrických odhadů hustot tvoří tzv. *Jádrové odhady*. První odhady tohoto druhu byly původně navrženy počátkem 50. let pro odhad spektrální hustoty. Nezávisle na sobě je odvodili NADARAYA a WATSON. Od té doby byly velmi zdokonaleny. Základní myšlenky si přitom našli řadu analogických použití v mnoha oblastech matematické statistiky.

2.1 Jádrové odhady hustoty náhodné veličiny

Jádrové odhady hustoty náhodné veličiny jsou pojmenovány podle funkcí zvaných jádra. Nejprve si tedy nadefinujeme jádro.

Definice 2.1. Jádrem nazveme libovolnou funkci

$$\mathsf{K} : (\mathbb{R}, \mathcal{B}) \to [0, +\infty), \tag{2.1}$$

která je symetrická, ohraničená a pro niž platí

$$\int_{-\infty}^{\infty} \mathsf{K}(x)\mathrm{d}x = 1 \tag{2.2}$$

a

$$\lim_{x \to \pm\infty} |x| \cdot \mathsf{K}(x) = 0, \tag{2.3}$$

kde $\mathcal{B}$ je σ-algebra borelovských množin na přímce.

Definice 2.2. Nechť $\{h_n\}_{n=1}^{\infty}$ je posloupnost kladných čísel taková, že

$$\lim_{n\to\infty} h_n = 0 \quad \text{a} \quad \lim_{n\to\infty} nh_n = \infty,$$

$\mathsf{K}(x)$ je jádro, $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení s hustotou $f(x)$. *Jádrový odhad hustoty* $f(x)$ je definován vztahem:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \mathsf{K}\left(\frac{x - X_i}{h_n}\right), \qquad x \in \mathbb{R}. \tag{2.4}$$

Seznam nejdůležitějších jader je uveden v tabulce 2.1. Parametr h_n má roli měřítka, které umožňuje pružně měnit tvar jádra, nazývá se šířka okna nebo též vyhlazovací parametr.

Název jádra	$\mathsf{K}(\mathbf{x})$	**obor hodnot**
1. Gaussovo	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	$x \in \mathbb{R}^1$
2. Laplaceovo	$\frac{1}{2} e^{-\lvert x\rvert}$	$x \in \mathbb{R}^1$
3. Cauchyovo	$\frac{1}{\pi(1+y^2)}$	$x \in \mathbb{R}^1$
4. Epanechnikovo	$\frac{3}{4\sqrt{5}}(1 - \frac{1}{5}y^2)$ 0	$\lvert x\rvert \leq \sqrt{5}$ $\lvert x\rvert > \sqrt{5}$
5. Trojúhelníkové	$1 - \lvert x\rvert$ 0	$\lvert x\rvert \leq 1$ $\lvert x\rvert > 1$
6. Obdelníkové	$\frac{1}{2}$ 0	$\lvert x\rvert \leq 1$ $\lvert x\rvert > 1$
7. Kosinové	$\frac{1}{2}\cos(x)$ 0	$\lvert x\rvert \leq \pi/2$ $\lvert x\rvert > \pi/2$
8. Kvartické	$\frac{15}{16}(1 - x^2)^2$ 0	$\lvert x\rvert \leq 1$ $\lvert x\rvert > 1$

Tabulka 2.1: Přehled nejdůležitějších jader

2.2 Jádrové odhady hustoty náhodného vektoru

Nechť je dán spojitý náhodný vektor $\mathbf{X} = (X_1, \ldots, X_m)'$ s hustotou $f(\mathbf{x})$, kde $\mathbf{x} = (x_1, \ldots, x_m)'$. Mějme k dispozici náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$, pak

jádrový odhad hustoty náhodného vektoru má tvar

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^m} \sum_{i=1}^{n} \mathsf{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right). \tag{2.5}$$

Parametr h se nazývá šířka okna, K je m-rozměrná jádrová funkce.

Jako parametr je také možné použít m-rozměrný vektor $\mathbf{h} = (h_1, \ldots, h_m)$, použijeme tedy jinou hodnotu parametru šířky okna pro každou proměnnou. Odhad má pak tvar

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 \ldots h_m} \sum_{i=1}^{n} \mathsf{K}\left(\frac{x_1 - X_{i,1}}{h_1}, \ldots, \frac{x_p - X_{i,m}}{h_m}\right). \tag{2.6}$$

Často se používají jádra s kompaktním nosičem (jádro nabývá nenulových hodnot pouze na ohraničeném intervalu). Pro tento případ se používá indikátorová funkce množiny, kterou označíme symbolem I s danou charakterizující množinou.

Mezi nejpoužívanější typy jader patří:

- konstantní jádro

$$\mathsf{K}(\mathbf{x}) = \frac{1}{2^m} I(max\{|x_i| \leq 1, i = 1 \ldots p\}), \tag{2.7}$$

- gaussovské jádro

$$\mathsf{K}(\mathbf{x}) = (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}\mathbf{x}'\mathbf{x}} \tag{2.8}$$

- Epanechnikovo jádro

$$\mathsf{K}(\mathbf{x}) = \frac{m+2}{2c_m}(1 - \mathbf{x}'\mathbf{x}) I(\mathbf{x}'\mathbf{x} \leq 1), \tag{2.9}$$

kde

$$c_m = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2} + 1)}. \tag{2.10}$$

Podrobnější přehled používaných jader je možné najít například v [12].

2.3 Součinové jádro

Při odhadu hustoty náhodného vektoru se také často používá jádro m proměnných ve tvaru součinu m jader jedné proměnné. Nazývá se součinové jádro.

Součinové jádro píšeme ve tvaru

$$\mathsf{K}_m(\mathbf{x}) = \prod_{i=1}^{m} \mathsf{K}_i(x_i), \tag{2.11}$$

kde K_i je jednorozměrné jádro a $\mathbf{x} = (x_1, \ldots, x_m)'$. Odtud dostaneme odhad vícerozměrné hustoty pomocí součinového jádra

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_i} \sum_{j=1}^{n} \prod_{i=1}^{m} \mathsf{K}\left(\frac{x_i - X_{ji}}{h_i}\right). \tag{2.12}$$

2.4 Odhad šířky vyhlazovacího okna

Při jádrových odhadech hustoty musí být věnována velká pozornost šířce vyhlazovacího okna h_n, protože podstatným způsobem ovlivňuje kvalitu odhadu. Odhad šířky okna můžeme řešit metodou *pokus – omyl*, za optimální hodnotu h_n zvolíme tu hodnotu, pro niž je křivka opticky nejhladší.

Přesnější výsledky získáme například při použití *minimalizace odhadu střední kvadratické chyby.*

Střední kvadratickou chybu definujeme

$$MSE(\hat{f}_n(x)) = \mathsf{E}\left(\hat{f}_n(x) - f(x)\right)^2. \tag{2.13}$$

Hledáme tedy takovou velikost šířky vyhlazovacího okna, pro niž je (2.13) minimální.

Jinou metodou odhadu šířky vyhlazovacího okna je *metoda křížového ověřování.* Hledáme konstantu h_n, jež minimalizuje věrohodnostní funkci

$$L(h_n) = \prod_{i=1}^{n} \hat{f}_{ni}(X_i), \tag{2.14}$$

kde

$$\hat{f}_{ni}(x) = \frac{1}{nh_n} \sum_{j=1, j\neq i}^{n} \mathsf{K}\left(\frac{x - X_i}{h_n}\right), \qquad x \in \mathbb{R}. \tag{2.15}$$

Kapitola 3

Jádrové odhady pravděpodobnostní funkce

3.1 Úvod

Nechť X je diskrétní náhodná veličina s pravděpodobnostní funkcí $p(x)$, která nabývá nenulových hodnot na množině bodů $\{x_k\}_{k=1}^{\infty}$. Pravděpodobnostní funkci $p(x)$ diskrétní náhodné veličiny označíme

$$p(x) = \begin{cases} \pi_k & \text{pro } x = x_k, \quad k = 1, 2, \dots \ , \\ 0 & \text{jinak,} \end{cases} \tag{3.1}$$

kdy

$$\sum_{k=1}^{\infty} \pi_k = 1. \tag{3.2}$$

Máme-li k dispozici náhodný výběr $\mathbf{X} = (X_1, \dots, X_n)'$, pak přirozeným odhadem π_k jsou

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x_k), \tag{3.3}$$

kde I je indikátorová funkce definovaná

$$I(X_i = x_k) = \begin{cases} 1 & \text{pokud } X_i = x_k, \\ 0 & \text{jinak.} \end{cases} \tag{3.4}$$

Tentýž odhad dostaneme, použijeme-li jádrový odhad

$$\hat{\pi}_k = \frac{1}{nh} \sum_{i=1}^{n} \mathsf{K}\left(\frac{x - X_i}{h}\right) \tag{3.5}$$

s jádrem

$$\mathsf{K}(x) = \begin{cases} 1 & \text{pro } x = X_i, \\ 0 & \text{pro } x \neq X_i, \end{cases} \tag{3.6}$$

a s vyhlazovacím parametrem $h = 1$. Jádrový odhad je v tomto případě ekvivalentní s odhadem pomocí relativních četností.

3.2 Jádrové odhady pravděpodobnostní funkce binární náhodné veličiny

Nechť je dána náhodná veličina $X \sim A(\pi)$, která má binární (alternativní) rozdělení s parametrem $\pi \in (0, 1)$.

Pravděpodobnostní funkce $p(x)$ náhodné veličiny X je rovna

$$p(x) = \begin{cases} \pi & \text{pro } x = 1, \\ 1 - \pi & \text{pro } x = 0, \\ 0 & \text{jinak.} \end{cases} \tag{3.7}$$

Pravděpodobnostní funkci $p(x)$ můžeme také psát ve tvaru

$$p(x) = \begin{cases} \pi^x (1 - \pi)^{1-x} & \text{pro } x = 0, 1, \\ 0 & \text{jinak.} \end{cases} \tag{3.8}$$

Nechť je dána náhodná veličina $X \sim A(\pi)$ s binárním rozdělením s parametrem $\pi \in (0, 1)$. Máme-li k dispozici náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$, pak jádrový odhad pravděpodobnostní funkce $p(x)$ je tvaru

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_b(x - X_i; \lambda), \tag{3.9}$$

kde

$$\mathsf{K}_b(x - X_i; \lambda) = \begin{cases} \lambda & \text{pro } x = X_i, \\ 1 - \lambda & \text{pro } x \neq X_i, \end{cases} \tag{3.10}$$

a

$$\frac{1}{2} \leq \lambda \leq 1. \tag{3.11}$$

Funkce $\mathsf{K}_b(x - X_i; \lambda)$ se nazývá *binární jádro*, parametr λ se nazývá *vyhlazovací parametr.*

V případě, kdy vezmeme $\lambda = 1$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_b(x - X_i; 1) = \begin{cases} 1 & \text{pro } x = X_i, \\ 0 & \text{pro } x \neq X_i, \end{cases} \tag{3.12}$$

což je odhad ekvivalentní s odhadem $\hat{\pi}_k$ pomocí relativních četností.

V případě, kdy vezmeme $\lambda = \frac{1}{2}$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_b(x - X_i; \frac{1}{2}) = \begin{cases} \frac{1}{2} & \text{pro } x = X_i, \\ \frac{1}{2} & \text{pro } x \neq X_i, \end{cases} \tag{3.13}$$

což můžeme zjednodušeně zapsat ve tvaru

$$\mathsf{K}_b(x - X_i; \frac{1}{2}) = \frac{1}{2}. \tag{3.14}$$

Pro vyhlazovací parametr $\lambda = \frac{1}{2}$ má pravděpodobnostní funkce náhodné veličiny rovnoměrné rozdělení.

3.3 Jádrové odhady pravděpodobnostních funkcí binárního náhodného vektoru

Je dán náhodný vektor $\mathbf{X} = (X_1, \ldots, X_m)'$, kde $X_i \sim A(\pi)$ je náhodná veličina s alternativním rozdělením pravděpodobnostní funkce pro $\forall i = 1, \ldots, m$, $0 < \pi < 1$.

Máme-li k dispozici náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$, pak jádrový odhad pravděpodobnostní funkce realizace náhodného vektoru $\mathbf{x}$ píšeme ve tvaru

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda), \tag{3.15}$$

kde

$$\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \lambda^{m - d_i^2}(1 - \lambda)^{d_i^2} \tag{3.16}$$

je m-rozměrné jádro a

$$d_i^2 = |\mathbf{x} - \mathbf{X}_i|^2 = (\mathbf{x} - \mathbf{X}_i)'(\mathbf{x} - \mathbf{X}_i). \tag{3.17}$$

Hodnotu vyhlazovací parametru volíme z intervalu

$$\frac{1}{2} \leq \lambda \leq 1. \tag{3.18}$$

Druhá mocnina eukleidovské vzdálenosti d_i^2 mezi vektory $\mathbf{x}$ a $\mathbf{X}_i$ udává počet rozdílných hodnot v jednotlivých složkách vektoru $\mathbf{x}$ a $\mathbf{X}_i$.

Poznámka. *Například pokud jsou vektory* $\mathbf{x}$ *a* $\mathbf{X}_i$ *shodné ve všech složkách, pak je hodnota jádra* $\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \lambda^m$. *Pokud se vektory* $\mathbf{x}$ *a* $\mathbf{X}_i$ *liší právě v jedné složce, pak je* $\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \lambda^{m-1}(1 - \lambda)$. *Pokud se liší ve dvou složkách, pak je* $\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \lambda^{m-2}(1 - \lambda)^2$ *a tak dále.*

Pak m-rozměrnou jádrovou funkci K_b^m můžeme psát také jako součin jednorozměrných binomických jader K_b.

$$\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \prod_{j=1}^{m} \mathsf{K}_b(\mathbf{x}_j - \mathbf{X}_{ji}; \lambda), \tag{3.19}$$

což můžeme psát ve tvaru

$$\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \lambda) = \prod_{j=1}^{m} \lambda^{1-w_{ij}}(1-\lambda)^{w_{ij}}, \tag{3.20}$$

kde

$$w_{ij} = |(\mathbf{x})_j - (\mathbf{X}_i)_j|. \tag{3.21}$$

$(\mathbf{x})_j$ značí j-tou složku vektoru $\mathbf{x}$.

Pokud je vyhlazovací parametr $\lambda = 1$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; 1) = \begin{cases} 1 & \text{pro } \mathbf{x} = \mathbf{X}_i, \\ 0 & \text{pro } \mathbf{x} \neq \mathbf{X}_i. \end{cases} \tag{3.22}$$

Odhad pravděpodobnostní funkce náhodného vektoru pak můžeme psát ve tvaru

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; 1) = \frac{n(\mathbf{x})}{n}, \tag{3.23}$$

kde

$$n(\mathbf{x}) = \sum_{i=1}^{n} I(\mathbf{x} = \mathbf{X}_i) \tag{3.24}$$

je počet členů vzorku, pro které platí rovnost $\mathbf{x} = \mathbf{X}_i$. Jádrový odhad pravděpodobnostní funkce se pro $\lambda = 1$ redukuje na odhad pravděpodobnostní funkce pomocí příslušných relativních četností.

Pokud je vyhlazovací parametr $\lambda = \frac{1}{2}$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_b^m(\mathbf{x} - \mathbf{X}_i; \frac{1}{2}) = \left(\frac{1}{2}\right)^m \qquad \forall i = 1, \ldots, n. \tag{3.25}$$

Odhad pravděpodobnostní funkce náhodného vektoru pak můžeme psát ve tvaru

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2}\right)^m = \left(\frac{1}{2}\right)^m. \tag{3.26}$$

Pravděpodobnostní funkce má pro $\lambda = \frac{1}{2}$ rovnoměrné rozdělení.

3.4 Jádrové odhady pravděp. funkce pro neuspořádané kategoriální proměnné

Každou náhodnou diskrétní veličinu, která má více než dvě kategorie, můžeme převést na binární data. AITCHISON a AITKEN [1] navrhli způsob, jak rozšířit jejich metodu jádrového odhadu pro neuspořádané a uspořádané kategoriální náhodné veličiny.

Neuspořádaná kategoriální data také nazýváme nominální data.

Nechť je dána náhodná veličina X, která může nabývat a různých hodnot. Máme-li k dispozici náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$, pak metodu odhadu pravděpodobnostní funkce binární náhodné veličiny pomocí jádra můžeme rozšířit následujícím způsobem. Odhad pravěpodobnostní funkce $\hat{p}(x)$ je

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{nk}(x - X_i; \lambda), \tag{3.27}$$

kde

$$\mathsf{K}_{nk}(x - X_i; \lambda) = \begin{cases} \lambda & \text{pro } x = X_i, \\ \frac{1-\lambda}{a-1} & \text{pro } x \neq X_i, \end{cases} \tag{3.28}$$

kde $x, X_i\{1, \ldots, a\}; i = 1, \ldots, n$.

Abychom zajistili, že největší váha padne na aktuální pozorování, musí vyhlazovacího parametru λ splňovat podmínku

$$\frac{1-\lambda}{a-1} \leq \lambda \leq 1, \tag{3.29}$$

vyhlazovací parametr λ tedy musí ležet v intervalu

$$\frac{1}{a} \leq \lambda \leq 1. \tag{3.30}$$

Pokud zvolíme vyhlazovací parametr $\lambda = 1$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_{nk}(x - X_i; 1) = \begin{cases} 1 & \text{pro } x = X_i, \\ 0 & \text{pro } x \neq X_i. \end{cases} \tag{3.31}$$

Jádrový odhad pravděpodobnostní funkce náhodného vektoru je pak ve tvaru

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{nk}(x - X_i; 1) = \frac{n(x)}{n}. \tag{3.32}$$

Jádrový odhad pravděpodobnostní funkce náhodného vektoru se pro $\lambda = 1$ redukuje na odhad pravděpodobnostní funkce pomocí příslušných relativních četností.

Pokud zvolíme vyhlazovací parametr $\lambda = \frac{1}{a}$, dostaneme hodnotu jádra ve tvaru

$$\mathsf{K}_{nk}(x - X_i; \frac{1}{a}) = \begin{cases} \frac{1}{a} & \text{pro } x = X_i, \\ \frac{1-\frac{1}{a}}{a-1} = \frac{1}{a} & \text{pro } x \neq X_i. \end{cases} \tag{3.33}$$

Odhad pravděpodobnostní funkce náhodného vektoru pak můžeme psát ve tvaru

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{a}\right)^m = \left(\frac{1}{a}\right)^m. \tag{3.34}$$

Pravděpodobnostní funkce má pro $\lambda = \frac{1}{a}$ rovnoměrné rozdělení pravděpodobnosti.

Nechť je dán náhodný vektor $\mathbf{X} = (X_1, \ldots, X_m)'$, kde X_i je kategoriální náhodná veličina a a_i značí počet hodnot, které může nabývat X_i. Označme $\mathbb{A}_i = \{1, \ldots, a_i\}$ pro $i = 1, \ldots, m$. Nechť je dán náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Označme $a = \prod_{i=1}^{m} a_i$ a $\mathbb{A} = \{1, \ldots, a\}$. Pak a značí počet možných hodnot nové kategoriální náhodné veličiny $Y \in \mathbb{A}$. Odhad pravděpodobnostní funkce nové kategoriální náhodné veličiny píšeme ve tvaru

$$\hat{p}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{nk}(y - Y_i; \lambda), \tag{3.35}$$

kde

$$\mathsf{K}_{nk}(y - Y_j; \lambda) = \begin{cases} \lambda & \text{pro } y = Y_j, \\ \frac{1-\lambda}{a-1} & \text{pro } y \neq Y_j, \end{cases} \tag{3.36}$$

$y, Y_j \in \mathbb{A}$ a vyhlazovací parametr volíme z intervalu

$$\frac{1}{a} \leq \lambda \leq 1. \tag{3.37}$$

Při odhadu pravděpodobnostní funkce náhodného vektoru pomocí součinového jádra použijeme následující postup.

Nechť je dán náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$, kde X_i je náhodná veličina, která může nabývat a_i různých hodnot.

Máme-li k dispozici náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$, pak jádrový odhad pravděpodobnostní funkce pomocí součinového jádra je ve tvaru

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}_{nk}^m(\mathbf{x} - \mathbf{X}_i; \lambda) \tag{3.38}$$

kde

$$\mathsf{K}_{nk}^{m}(\mathbf{x}-\mathbf{X}_i;\lambda)=\prod_{j=1}^{m}\lambda^{1-w_{ij}}\left(\frac{1-\lambda}{a_j-1}\right)^{w_{ij}}, \tag{3.39}$$

kde

$$w_{ij}=|(\mathbf{x})_j-(\mathbf{X}_i)_j| \tag{3.40}$$

a vyhlazovací parametr λ musí splňovat podmínku

$$\max_{j=1,\ldots,m}\frac{1}{a_j}\leq\lambda\leq 1. \tag{3.41}$$

V případě, kdy vyhlazovací parametr λ nabývá různých hodnot pro jednotlivé marginální náhodné veličiny použijeme jádro ve tvaru

$$\mathsf{K}_{nk}^{m}(\mathbf{x}-\mathbf{X}_i;\boldsymbol{\lambda})=\prod_{j=1}^{m}\lambda_j^{1-w_{ij}}\left(\frac{1-\lambda_j}{a_j-1}\right)^{w_{ij}}, \tag{3.42}$$

kde $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_m)'$ je vektor, který obsahuje m vyhlazovacích parametrů příslušných k jednotlivým náhodným veličinám. Každé λ_j musí splňovat podmínku

$$\frac{1}{a_j}\leq\lambda_j\leq 1,\qquad\forall j=1,\ldots,m. \tag{3.43}$$

3.5 Jádrové odhady pravděp. funkce pro uspořádané kategoriální proměnné

Diskrétní jádro (3.42) můžeme upravit pro použití při odhadu pravděpodobnostní funkce uspořádané kategorické náhodné veličiny. Aitchison a Aitken [1] ukázali způsob, jakým lze jádro upravit pro trojčlennou uspořádanou kategoriální náhodnou veličinu.

Předpokládejme, že X_p je tříčlenná uspořádaná náhodná veličina nabývající hodnoty 0, 1 a 2. Pro odhad pravděpodobnostní funkce použijeme jádro

$$\mathsf{K}_{uk}(x-X_i;\lambda)=\prod_{p=1}^{m}\mathsf{K}_p(x_p,X_{ip};\lambda), \tag{3.44}$$

kde hodnotu $\mathsf{K}_p(x_p,X_{ip};\lambda)$ vezmeme z tabulky 3.1. Vyhlazovací parametr musí splňovat podmínku

$$\lambda\geq\frac{2}{3}. \tag{3.45}$$

X_{ip}	$x_p = 0$	$x_p = 1$	$x_p = 2$
0	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$
1	$\frac{1}{2}(1-\lambda^2)$	λ^2	$\frac{1}{2}(1-\lambda^2)$
2	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2

Tabulka 3.1: Hodnoty jader $\mathsf{K}_{uk}(x_p, X_{ip}; \lambda)$ pro tříčlennou uspořádanou kategoriální náhodnou veličinu

Poznámka. *Z tabulky 3.1 vidíme, že pokud rozdíl $|x_p - X_{ip}| = 0$, tak hodnota přírůstku jádra je λ^2. S rostoucí hodnotou rozdílu $|x_p - X_{ip}|$ hodnota přírůstku jádra klesá.*

3.6 Volba vyhlazovacího parametru

Odhad pravděpodobnostní funkce závisí na volbě vyhlazovacího parametru λ. Vhodnou volbou vyhlazovacího paramteru λ docílíme přesnějšího odhadu pravděpodobnostní funkce.

Metody pro výběr vyhlazovacího parametru jsou založeny na diskrétní ztrátové funkci. Nechť

$$\mathbb{A} = \prod_{j=1}^{m} \mathbb{A}_j = \prod_{j=1}^{m} \{1, \ldots, a_j\} = \{1, \ldots, a\} \tag{3.46}$$

značí množinu možných hodnot (variant) diskrétního náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_m)'$ s pravděpodobnostní funkcí $p(\mathbf{x})$. $\mathbb{A}_j = \{1, \ldots, a_j\}$ značí možné hodnoty kategoriální náhodné veličiny $X_j, \forall j = 1, \ldots, m$. Dále mějme k dispozici náhodný výběr $\mathbb{D} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$.

3.6.1 Metoda navržená Aitchisonem a Aitkenem (1976)

Aitchison a Aitken v [1] navrhli postup odhadu λ pomocí metody *křížového ověřování* („cross-validation“ nebo také „leaving-one-out“) vycházející z metody maximální věrohodnosti.

Při této metodě volíme vyhlazovací paramter λ z intervalu $\frac{1}{a} \leq \lambda \leq 1$ tak, aby maximalizoval pseudověrohodnostní funkci

$$CV_{ML}(\lambda, \mathbb{D}) = \prod_{i=1}^{n} \hat{p}(\mathbf{X}_i | \mathbb{D} - \mathbf{X}_i; \lambda), \tag{3.47}$$

kde $\mathbb{D} - \mathbf{X}_i = \{\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n\}$ značí množinu $\mathbb{D}$ s vyloučeným vektorem $\mathbf{X}_i$. Upravením (3.47) dostaneme

$$CV_{ML}(\lambda, \mathbb{D}) = \prod_{i=1}^{n} \frac{1}{n-1} \sum_{j=1,\, j\neq i}^{n} \mathsf{K}(\mathbf{X}_i - \mathbf{X}_j; \lambda). \tag{3.48}$$

Označme $\hat{\lambda}$ odhad vyhlazovacího paramteru λ získaný maximalizací pseudověrohodnostní funkce $CV_{ML}(\lambda, \mathbb{D})$. Aitchison a Aitken ukázali, že pro $n \to \infty$ odhad pravděpodobnosti $\hat{p}(\mathbf{x}|\mathbb{D}, \hat{\lambda})$ konverguje k $p(\mathbf{x})$. Tedy

$$\lim_{n\to\infty} \hat{p}(\mathbf{x}|\mathbb{D}, \hat{\lambda}) = p(\mathbf{x}). \tag{3.49}$$

3.6.2 Metoda navržená Hallem (1981)

Hall v [8] ukázal nedostatky metody navrhnuté Aitchisonem a Aitkenem.

Pokud je několik variant v náhodném výběru prázdných a všechny ostatní varianty obsahují alespoň dvě pozorování, pak je hodnota λ pravděpodobně odhadnuta jako 1. V tomto případě bude odhad pravděpodobnostní funkce náhodného vektoru pomocí jader ekvivalentní s odhadem pomocí relativních četností. Tento odhad přiřadí nulovou přavděpodobnost prázdným buňkám.

Přítomnost varianty právě s jedním pozorováním má přesně opačný efekt. Když hodnota vyhlazovacího paramtru λ roste k 1, odhad pravděpodobnostní funkce náhodného vektoru bude klesat k 0.

Hall navrhl metodu *křížového ověřování* vycházející z metody *nejmenších čtverců* („least-squares cross-validation“). Tato metoda je založena na minimalizaci *střední integrální kvadratické chyby.*

$$\begin{aligned} MISE_{\hat{f}}(\lambda) &= \mathsf{E} \int \left(\hat{f}(x;\lambda) - f(x)\right)^2 \mathrm{d}x \qquad (3.50) \\ &= \mathsf{E} \int \left(\hat{f}(x;\lambda)\right)^2 \mathrm{d}x - 2 \underbrace{\mathsf{E} \int \hat{f}(x;\lambda) f(x)\, \mathrm{d}x}_{=\mathsf{E}\hat{f}(\mathbf{X})} + \\ &\quad + \int (f(x))^2 \, \mathrm{d}x. \qquad (3.51) \end{aligned}$$

Skutečná hodnota f je neznámá, tedy výraz (3.51) musíme odhadnout. Roznásobením mocnin a ignorováním výrazů neobsahujících λ dostaneme

odhad $MISE_{\hat{f}}$

$$CV_{LS}(\lambda) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\mathbf{x}\in\mathbb{A}}\mathsf{K}(\mathbf{x}-\mathbf{X}_i;\lambda)\mathsf{K}(\mathbf{x}-\mathbf{X}_j;\lambda) - \\ - 2\underbrace{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,\,j\neq i}^{n}\mathsf{K}(\mathbf{X}_i-\mathbf{X}_j;\lambda)}_{=\widehat{\mathsf{E}\hat{f}(\mathbf{X})}}. \tag{3.52}$$

3.7 Smíšená data

V praxi se často setkáme s případem, kdy máme nalézt hustotu náhodného vektoru, jehož složky jsou jak spojité náhodné veličiny, tak i diskrétní náhodné veličiny.

Nechť je dán náhodný vektor $\mathbf{Z} = (X, Y)$, kde X je binární náhodná veličina a Y je spojitá náhodná veličina. Mějme k dispozici náhodný výběr

$$\mathbb{D} = \{(x_i, y_i), i = 1, \ldots, n\}. \tag{3.53}$$

Jádrový odhad hustoty pak můžeme pomocí součinového jádra spočítat jako

$$p(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{K}_1(x - X_i;\lambda)\frac{1}{2}\mathsf{K}_2\left(\frac{y-Y_i}{h}\right), \tag{3.54}$$

kde K_1 je diskrétní jádro a K_2 je vhodné spojité jádro.

Kapitola 4

Ilustrativní příklady

Všechny ilustrativní příklady byly spočítány pomocí výpočetního systému MATLAB. Zdrojové texty použitých programů jsou uvedeny v příloze A.

Simuluji náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_{50}$, kde $\mathbf{X}_i, i = 1, \ldots, 50$ jsou realizace náhodného vektoru $\mathbf{X} = (X_1, X_2, X_3)$. $X_1 \sim A(\pi_1)$, $X_2 \sim A(\pi_2)$, $X_3 \sim A(\pi_3)$ jsou náhodné veličiny s alternativním rozdělením s parametry $PI = (\pi_1, \pi_2, \pi_3) = (0.4, 0.2, 0.7)$. Pro simulaci náhodného výběru použiji příkaz

```
> D=nahvyber(50,3,PI)
```
(viz sekce A.1).

4.1 Odhad pravděpodobnostní funkce s optimální volbou vyhlazovacího parametru

Pro odhad vyhlazovacího parametru λ s přesností na tři desetiné místa použiji příkaz

```
> h=lambda(D,0.001)
```
(viz sekce A.2)

Odhad vyhlazovacího parametru je

```
> h = 0.910.
```

Nyní již můžu provézt samotný odhad pravděpodobnostní funkce pomocí jádra.

```
> Xp=jadra(h,D,PI)
```
(viz sekce A.3)

Jako výstup programu dostaneme matici

$$X_p = \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 0.0888 & 0.0800 & 0.0560 \\ 0 & 1 & 1 & 0.0798 & 0.0600 & 0.0840 \\ 1 & 0 & 1 & 0.2350 & 0.2600 & 0.2240 \\ 0 & 0 & 1 & 0.2932 & 0.3400 & 0.3360 \\ 1 & 1 & 0 & 0.0308 & 0.0200 & 0.0240 \\ 0 & 1 & 0 & 0.0218 & 0 & 0.0360 \\ 1 & 0 & 0 & 0.0962 & 0.0800 & 0.0960 \\ 0 & 0 & 0 & 0.1544 & 0.1600 & 0.1440 \end{array}\right). \tag{4.1}$$

První tři sloupce matice X_p udávají souřadnice vektorů, jejichž pravděpodobnost odhadujeme. Čtvrtý sloupec udává pravděpodobnosti odhadnuté pomocí jader. Pátý sloupec udává pravděpodobnosti spočítané pomocí relativních četností. Poslední sloupec udává skutečné pravděpodobnosti jednotlivých realizací náhodného vektoru.

Vektor $(0, 1, 0)$ se v náhodném výběru nevyskytl. Odhad pomocí relativních četností mu tedy přiřadil nulovou pravděpodobnost, přestože jeho skutečná pravděpodobnost je nenulová. Odhad pomocí jader tento problém eliminoval a přiřadil vektoru (0,1,0) pravděpodobnost nenulovou.

4.2 Odhad pravděpodobnostní funkce pro krajní hodnoty vyhlazovacího parametru

V případě, kdy zvolíme pro stejný náhodný výběr hodnotu vyhlazovacího parametru $\lambda = 1$, dostaneme hodnoty jádrových odhadů stejné, jako hodnoty odhadů pomocí relativních četností, jak je vidět v (4.2).

```
> Xp=jadra(1,D,PI)
```
(viz sekce A.3)

$$X_p = \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 0.0800 & 0.0800 & 0.0560 \\ 0 & 1 & 1 & 0.0600 & 0.0600 & 0.0840 \\ 1 & 0 & 1 & 0.2600 & 0.2600 & 0.2240 \\ 0 & 0 & 1 & 0.3400 & 0.3400 & 0.3360 \\ 1 & 1 & 0 & 0.0200 & 0.0200 & 0.0240 \\ 0 & 1 & 0 & 0 & 0 & 0.0360 \\ 1 & 0 & 0 & 0.0800 & 0.0800 & 0.0960 \\ 0 & 0 & 0 & 0.1600 & 0.1600 & 0.1440 \end{array}\right) \tag{4.2}$$

V případě, kdy zvolíme pro stejný náhodný výběr hodnotu vyhlazovacího parametru $\lambda = \frac{1}{2}$, bude mít náhodný vektor rovnoměrné rozdělení pravděpodobnosti, jak je vidět v (4.3).

```
> Xp=jadra(0.5,D,PI)
```
(viz sekce A.3)

$$X_p = \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 0.1250 & 0.0800 & 0.0560 \\ 0 & 1 & 1 & 0.1250 & 0.0600 & 0.0840 \\ 1 & 0 & 1 & 0.1250 & 0.2600 & 0.2240 \\ 0 & 0 & 1 & 0.1250 & 0.3400 & 0.3360 \\ 1 & 1 & 0 & 0.1250 & 0.0200 & 0.0240 \\ 0 & 1 & 0 & 0.1250 & 0 & 0.0360 \\ 1 & 0 & 0 & 0.1250 & 0.0800 & 0.0960 \\ 0 & 0 & 0 & 0.1250 & 0.1600 & 0.1440 \end{array}\right) \tag{4.3}$$

Příloha A

Programy v MATLABU

Nechť je dán náhodný vektor $\mathbf{X} = (X_1, \ldots, X_m)$, kde $X_i \sim A(\pi_i)$ má alternativní rozdělení $\forall i = 1, \ldots, n$ a $\pi_i \in [0, 1]$.

A.1 Generování náhodného výběru

Program `[D]=nahvyber(n,m,PI)` generuje náhodný výběr příslušející náhodnému vektoru $\mathbf{X}$. `n` udává počet generovaných prvků náhodného výběru, `m` udává počet členů náhodného vektoru $\mathbf{X}$ a `PI` $= (\pi_1, \ldots, \pi_m)$ je vektor udávající parametry alternativního rozdělelní příslušející jednotlivým náhodným veličinám X_i.

Výstup programu je matice `D`, jejíž řádky jsou jednotlivé prvky náhodného výběru.

Zdrojový text programu:

```
function [D]=nahvyber(n,m,PI)

%Generuje náhodný výber s alternativnim rozdelenim rozsahu n*m
%PI vektor parametru alternativniho rozdeleni prislusejicich
%jednotlivym slozkam nahodneho vektoru

for i=1:n
  for j=1:m
    D(i,j)=[(rand(1)<PI(j))];
  end;
end;
```

A.2 Odhad vyhlazovacího parametru

Program `[h]=lambda(D,eps)` provádí odhad vyhlazovacího parametru λ pomocí vzorce (3.48) metody navržené AITCHISONEM a AITKENEM. `D` je náhodný výběr na základě kterého provádíme odhad. `eps` je přesnost s jakou provádíme odhad vyhlazovacího paramteru λ. Je důležité si uvědomit, že s vetší přesností roste doba výpočtu.

Výstup programu je odhad vyhlazovacího paramteru `h`.

Zdrojový text programu:

```
function [hn]=lambda(D,eps)

%[hn]=lambda(D,eps)
%D je nahodny vyber rozsahu n*m
%eps je presnost s jakou bude spocitana hodnota
%vyhlazovaciho parametru
%pro vetsi presnost bude vypocet trvat radove 10* vice

h=0.5:eps:1;
k=size(h,2);
[n,m]=size(D);
CV=ones(1,k);

for g=1:k
  for i=1:n
    k=0;
    for j=1:n
      if i==j
      else
        d=(D(i,:)-D(j,:))*(D(i,:)-D(j,:))';
        km=h(g)^(m-d)*(1-h(g))^d;
        k=k+km;
      end;
    end;
    k=k/(n-1);
    CV(g)=CV(g)*k;
  end;
end;
[CVmax,CVindex]=max(CV);
hn=h(CVindex);
```

A.3 Odhad pravděpodobnostní funkce náhodného vektoru

Program `[Xp]=jadra(h,D,PI)` provádí odhad pravděpodobnostní funkce náhodného vektoru $\mathbf{X}$ na základě vzorce (3.15). `h` je hodnota vyhlazovacího paramteru λ. `D` je příslušný náhodný výběr. `PI` je vektor udávající parametry skutečného rozdělení, uvádí se pouze pro porovnání odhadnuté pravděpodobnosti se skutečnými hodnotami.

Výstup programu je matice `Xp`, kde na každém řádku prvních `m` složek udává souřadnice vektoru, jehož pravděpodobnost hledáme. Další složka udává hodnotu pravděpodobnostní funkce příslušného vektoru odhadnutou pomocí jádra. Následující složka udává odhadnutou hodnotu pravděpodobnostní funkce pomocí relativních četností a poslední hodnota na každém řádku udává skutečnou hodnotu pravděpodobnostní funkce příslušného vektoru. Můžeme tedy porovnat úspěšnost jádrového odhadu se skutečnou hodnotou a odhadem pomocí relativních četností.

Zdrojový text programu:

```
function [Xp]=jadra(h,D,PI)

% [Xp]=jadra(h,D,PI)
%
% h vyhlazovaci parametr lambda
%
% D nahodny vyber-cvicne data, radky matice davaji
% jednotlive cleny nahodneho vyberu
%
% PI vektor hodnot parametru alternativniho rozdeleni
% jednotlivych nahodnych velicin
%
% Xp je matice, kde na radku je vzdy uveden vektor
%    s odhadem pravdepodobnosti pomoci jadra,
%    odhadem pomoci relativnich cetnosti
%    a nakonec je uvedena skutecna pravdepodobnost

[n,m]=size(D);

% Generovani matice všech moznych vektoru x
a=1;
c=2;

for j=1:m
```

```
i=1;
  for nn=1:(2^m/c)
    for k=1:a
      X(i,j)=1;
      i=i+1;
    end;
    for l=1:a
      X(i,j)=0;
      i=i+1;
    end;
  end;
a=2*a;
c=2*c;
end;

% odhad pomoci relativnich cetnosti

for i=1:(2^m)
  zm=0;
  for j=1:n
    z=(m==sum((X(i,:)==D(j,:))));
    zm=z+zm;
  end;
  w(i,1)=zm/n;
end;

% odhad pravdepodobnosti nahodneho vektoru

for s=1:2^m
  k=0;
  for fn=1:n
    d=sum(abs(D(fn,:)-X(s,:)));
    km=h^(m-d)*(1-h)^d;
    k=k+km;
  end;
  p(s)=k/n;
end;
p=p';

%skutecna pravdepodobnost
for i=1:2^m
  pr=1;
  for j=1:m
```

```
    if X(i,j)==1
      pr=pr*PI(j);
    else
      pr=pr*(1-PI(j));
    end;
  end;
  PR(i,1)=pr;
end;

%vysledna matice
Xp=X;
Xp(:,m+1)=p;
Xp(:,m+2)=w;
Xp(:,m+3)=PR;
```

Literatura

[1] Aitchison, J., Aitken, C. G. G.: *Multivariete Binary Discimination by the Kernel Method*, Biometrika, vol. 63, No. 3, str. 413-420, 1976

[2] Aitken, C., MacDonald, D. G.: *An Application of Discrete Kernel Methods to Forensic Odontology*, Applied Statistics , Volume 28, Issue 1, str. 55-61, 1979

[3] Anděl, J.: *Matematická statistika*, SNTL/ALFA, Praha 1985

[4] Butler, J.: *Multivariete Discrimination of Binary Data using SAS Software*, Research & Statistics of Income, Internal Revenue Services, Washington D.C.

[5] Čermáková, A., Forbelská, M.: *Diskriminační analýza pro vícerozměrná binární data*, Dep.of Math. Faculty of Mechanical Eng. Slovac University of Techn. in Bratislava, APLIMAT, Bratislava 2004

[6] Čermáková, A., Forbelská, M.: *Paramteric and Nonparametric Disrcimination for Categorical Variables*, Masaryk University, Faculty of Science, Mathematica, DATASTAT, Brno 2003

[7] Forbelská, M.: *Neparametrická diskriminační analýza*, sborník, ROBUST'2000, Nečtiny, str. 50-58, 2001

[8] Hall, P.: *On Nonparametric Multivariate Binary Discrimininination*, Biometrika, vol.68, Issue 1, str. 287-294, 1981

[9] Kunderová, P.: *Základy pravděpodobnosti a matematické statistiky*, Univerzita Palackého v Olomouci, Přírodovědecká fakulta, Olomouc 2004

[10] McLachlan, G. J.: *Discrimination Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York 1992

[11] Murray, G. D., Titterington, D. M.: *Estimation Problems with Data from a Mixture*, Applied Statistics, Volume 27, Issue 3, str. 325-337, 1978

[12] Nadaraya, E. A.: *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academica Publishers, Dordrecht, Boston, London 1989

[13] Zádrapa, T.: *Jádrové odhady*, Diplomová práce, přírodovědecká fakulta MU, Brno 1996

[14] Zelinka, J.: *Jádrové odhady hustoty náhodného vektoru*, KAM, přírodovědecká fakulta MU, Brno, ROBUST 1994