Kapitola 5

Jádrové odhady dvourozměrných hustot

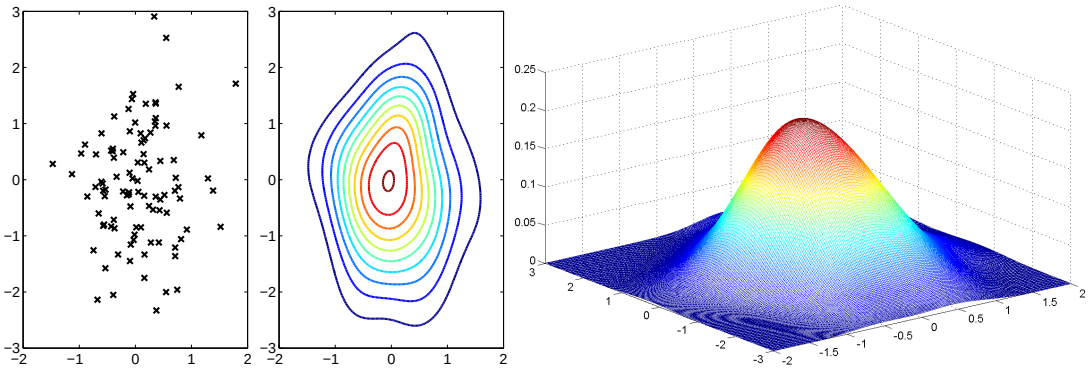
Výstupy z výukové jednotky

Student

- bude znát součinová a sférická dvourozměrná jádra pro odhady dvourozměrných hustot.
- porozumí principu vyhlazování dvourozměrných hustot.
- pochopí nejjednodušší metody pro volbu prvků diagonální vyhlazovací matice.
- zvládne použití příslušného toolboxu v Matlabu pro simulační studii i pro zpracování reálných dat.

1 Motivace

V této kapitole se budeme zabývat rozšířením jádrových odhadů pro jednorozměrné hustoty na odhad vícerozměrných hustot. Ovšem ve vícerozměrném případě nevystačíme s jedním vyhlazovacím parametrem, ale je třeba specifikovat matici vyhlazovacích parametrů. Tato matice řídí jak hladkost, tak i orientaci vícerozměrného vyhlazení. Budeme se zabývat jádrovým odhadem, který je přímým rozšířením jednorozměrného odhadu (3.1) v kapitole 3, a zaměříme se zejména na odhad dvourozměrné hustoty.



Obrázek 5.1: Náhodný výběr a jeho jádrový odhad

Poznámka 1.1. Jádrové odhady dvourozměrných hustot se obvykle znázorňují pomocí vrstevnic, které umožňují snazší náhled na odhadnutou funkci.

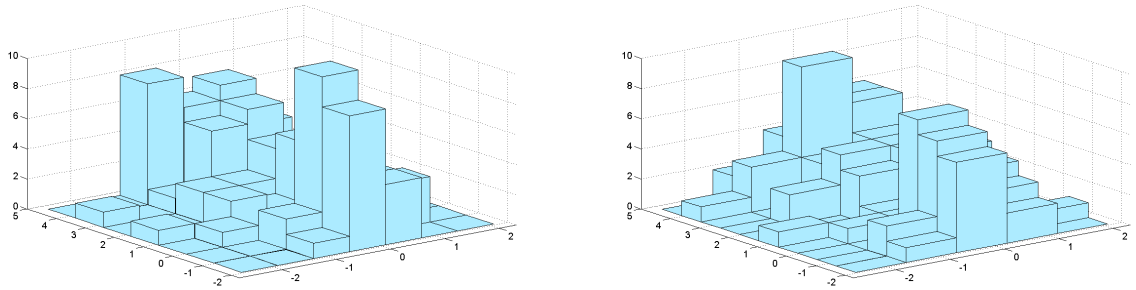
2 Základní typy odhadů

Podobně jako u odhadů hustoty můžeme použít *histogram*, ale ten má zmíněné nevýhody – jde o schodovitou funkci a je citlivý na volbu počtu a šířky třídicích obdélníků – viz obrázek 5.2.

Příklad 2.1. Mějme dán datový soubor o velikosti $n = 100$ generovaný ze směsi tří normálních hustot[1]. $N(0, -1; 1/3, 1/3, 0)$, $N(0, 2; 1, 1, 0)$ a $N(0, 4; 1/3, 1/3, 0)$

$$f(x,y) = \frac{1}{3}\frac{3}{2\pi}\,\mathrm{e}^{-\frac{3}{2}\left(x^2+(y+1)^2\right)} + \frac{1}{3}\frac{1}{2\pi}\,\mathrm{e}^{-\frac{1}{2}\left(x^2+(y-2)^2\right)} + \frac{1}{3}\frac{3}{2\pi}\,\mathrm{e}^{-\frac{3}{2}\left(x^2+(y-4)^2\right)}$$

(Data jsou v tabulce 6.4.) Z obrázku 5.2 je patrné, že histogram nepostihuje charakteristické rysy hustoty pravděpodobnosti dat.



(a) 7×7 obdélníků (b) 5×10 obdélníků

Obrázek 5.2: Histogramy s různými počty třídicích obdélníků

Předpokládejme, že máme k dispozici náhodný výběr $([X_1, Y_1], \ldots, [X_n, Y_n])$ z dvourozměrného spojitého rozdělení s hustotou $f(x, y)$. Jádrový odhad hustoty f v bodě $[x, y] \in \mathbb{R}^2$ je definovaný vztahem

$$\hat{f}(x, y; \mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(x - X_i, y - Y_i), = \frac{1}{n|\mathbf{H}|}\sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(x - X_i, y - Y_i)^T\right) \qquad (5.1)$$

přičemž $\mathbf{H}$ je matice vyhlazovacích parametrů a K je dvourozměrné jádro.

Jádro K je dvourozměrná funkce, kterou můžeme získat pomocí jednorozměrného symetrického jádra K_1 ($K_1 \in S_{02}$). Existují dva typy těchto jader:

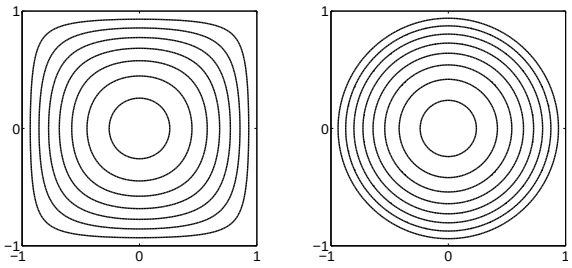
- *součinové jádro* $K^P(x, y) = K_1(x) \cdot K_1(y)$,
- *sféricky symetrické jádro* $K^S(x, y) = c_k^{-1} K_1\left(\sqrt{x^2 + y^2}\right)$, $c_k = \iint K_1\left(\sqrt{x^2 + y^2}\right) \mathrm{d}x\,\mathrm{d}y$.

Příklad 2.2. Epanečnikovo jádro, které je v jednorozměrném případě tvaru $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, má následující dvourozměrné varianty

$$K^P(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2) \quad \text{pro} \quad -1 \le x, y \le 1,$$
$$K^S(x, y) = \frac{2}{\pi}(1 - x^2 - y^2) \quad \text{pro} \quad x^2 + y^2 \le 1.$$

Na obrázku 5.3 jsou zobrazeny vrstevnice těchto jader.

[1] Používáme zde zkrácený zápis pro dvourozměrnou hustotu normálního rozdělení, a to $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \varrho)$



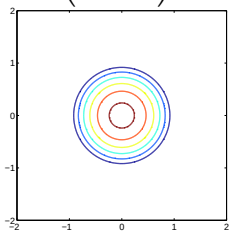
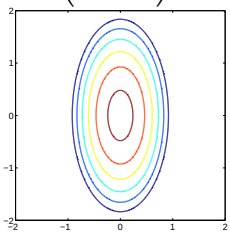
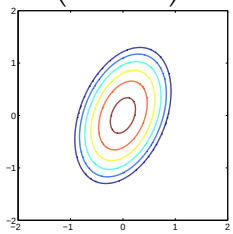
Obrázek 5.3: Součinové (vlevo) a sféricky symetrické (vpravo) dvourozměrné Epanečnikovo jádro

Podívejme se blíže na matici $\mathbf{H}$. Jde o matici vyhlazovacích parametrů, které řídí hladkost výsledného odhadu. Navíc také udávají orientaci odhadnuté hustoty. Rozlišujeme tři základní třídy vyhlazovacích matic:

- třída $\mathcal{S}$, která obsahuje matice s jediným vyhlazovacím parametrem,
- třída $\mathcal{D}$, která zahrnuje diagonální matice,
- třída $\mathcal{F}$, která obsahuje tzv. plné matice.

Rozdíly mezi jednotlivými maticemi jsou patrné z tabulky 5.1, kde jsou zobrazeny vrstevnice součinového Epanečnikova jádra v závislosti na třídě matic.

Tabulka 5.1: Třídy vyhlazovacích matic

$\mathcal{S}$	$\mathcal{D}$	$\mathcal{F}$
$\begin{pmatrix} h^2 & 0 \\ 0 & h^2 \end{pmatrix}$	$\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$	$\begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$
		

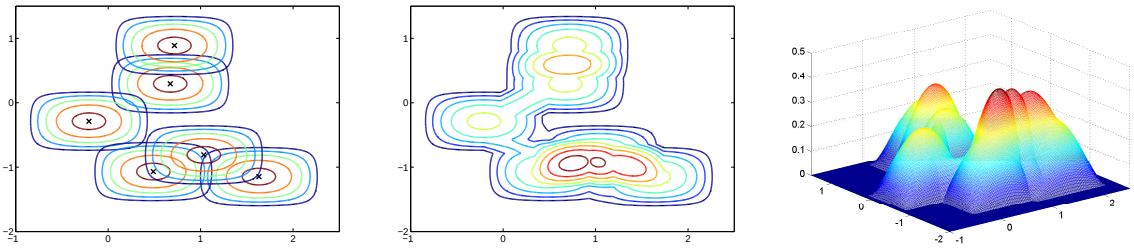
Budeme se zabývat jádrovými odhady s diagonální vyhlazovací maticí. Jádrový odhad s maticí třídy $\mathcal{S}$ dává ve všech směrech stejnou míru vyhlazení, což neponechává příliš mnoho prostoru pro zachycení variabilty dat. Na druhou stranu při použití matice třídy $\mathcal{F}$ je potřeba odhadnout větší počet parametrů, což znamená vyšší výpočetní náročnost.

Konstrukce jádrového odhadu je analogická konstrukci jednorozměrného odhadu. Tedy v každém bodě $[X_i, Y_i]$ sestrojíme jádro $K_{\mathbf{H}}$ a odhad v bodě $[x, y]$ je průměr n hodnot jader v tomto bodě – viz obrázek 5.4.

3 Statistické vlastnosti jádrových odhadů hustoty

Stejně jako u jádrových odhadů jednorozměrných hustot můžeme kvalitu jádrového odhadu hustoty popsat lokálně pomocí střední kvadratické chyby:

$$\operatorname{MSE} \hat{f}(x, y; \mathbf{H}) = \frac{1}{n} \underbrace{\left((K_{\mathbf{H}}^2 * f)(x, y) - (K_{\mathbf{H}} * f)^2(x, y)\right)}_{\text{var}} + \underbrace{\left((K_{\mathbf{H}} * f)(x, y) - f(x, y)\right)}_{\text{bias}}^2,$$



Obrázek 5.4: Konstrukce jádrového odhadu hustoty

nebo globálně pomocí střední integrální kvadratické chyby

$$\mathrm{MISE}\,\hat{f}(x,y;\mathbf{H}) = \iint \mathrm{MSE}\,\hat{f}(x,y;\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y.$$

Poznámka 3.1. Podobně jako v jednorozměrném případě se definuje konvoluce funkcí dvou proměnných. Nechť jsou dány funkce f a g, pro které platí $\iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y < \infty$ a $\iint g^2(x,y)\,\mathrm{d}x\,\mathrm{d}y < \infty$. *Konvoluci* $f * g$ definujeme vztahem

$$(f*g)(x,y) = \iint_{-\infty}^{\infty} f(t,u)g(x-t,y-u)\,\mathrm{d}t\,\mathrm{d}u.$$

Optimální vyhlazovací matice minimalizuje MISE. Je zřejmé, že tyto optimální hodnoty vyhlazovacích parametrů není možné z MISE přímo vyjádřit. Stejně jako u odhadu jednorozměrných hustot se budeme zabývat asymptotickou střední integrální kvadratickou chybou AMISE.

Věta 3.1. *Předpokládejme, že funkce f, jádro K a matice vyhlazovacích parametrů*[2] $\mathbf{H} = \mathrm{diag}(h_1^2, h_2^2)$ *splňují následující předpoklady.*

(i) Nechť $\mathbf{H} = \mathbf{H}_n$ je posloupnost matic takových, že $(n|\mathbf{H}|)^{-1}$ a prvky matice $\mathbf{H}$ konvergují k nule pro $n \to \infty$.

(ii) Dále nechť všechny druhé parciální derivace funkce f jsou ohraničené, spojité a integrovatelné se čtvercem.

(iii) Jádro K splňuje

$$\iint xK(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint yK(x,y)\,\mathrm{d}x\,\mathrm{d}y = 0$$
$$\iint x^2K(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint y^2K(x,y)\,\mathrm{d}x\,\mathrm{d}y = \beta_2(K).$$

Pak platí

$$\mathrm{MISE}(\mathbf{H}) = \mathrm{AMISE}(\mathbf{H}) + o(h_1^2+h_2^2) + o\big((h_1h_2n)^{-1}\big),$$

kde

$$\mathrm{AMISE}(\mathbf{H}) \equiv \mathrm{AMISE}\,\hat{f}(\cdot,\mathbf{H}) = \frac{V(K)}{nh_1h_2} + \frac{1}{4}\beta_2^2(K)\big(h_1^4V(f_{xx}) + 2h_1^2h_2^2V(f_xf_y) + h_2^4V(f_{yy})\big), \quad (5.2)$$

přičemž označení je ve shodě s předchozími kapitolami, tj. $V(g) = \iint g^2(x,y)\,\mathrm{d}x\,\mathrm{d}y$.

[2] Užíváme zde zkráceného zápisu: $\mathrm{diag}(h_1,h_2) = \begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}$

Důkaz věty o tvaru AMISE je založen na Taylorově rozvoji funkce $f(x, y)$ a lze jej nalézt např. v knize [14].

Hodnoty parametrů h_1, h_2, pro které $\mathrm{AMISE}(h_1, h_2)$ nabývá minimální hodnoty, jsou dány vztahy:

$$h_{1,opt} = \left(\frac{V^{3/4}(f_{yy})V(K)}{n\beta_2^2(K)V^{3/4}(f_{xx})\left[V(f_x f_y) + \sqrt{V(f_{xx})V(f_{yy})}\right]} \right)^{1/6}, \tag{5.3}$$

$$h_{2,opt} = \left(\frac{V^{3/4}(f_{xx})V(K)}{n\beta_2^2(K)V^{3/4}(f_{yy})\left[V(f_x f_y) + \sqrt{V(f_{xx})V(f_{yy})}\right]} \right)^{1/6}. \tag{5.4}$$

Z těchto vztahů plyne, že množina přípustných vyhlazovacích parametrů je tvaru $a_1 n^{-1/6} \leq h_1 \leq b_1 n^{-1/6}$, $a_2 n^{-1/6} \leq h_2 \leq b_2 n^{-1/6}$ pro vhodné konstanty $0 < a_1 < b_1 < \infty$, $0 < a_2 < b_2 < \infty$.

4 Volba jádra

Podobně jako u odhadu jednorozměrné hustoty není volba jádra podstatná. Je vhodné zvolit součinový tvar optimálního jádra. Tím zajistíme jistou hladkost výsledného odhadu a navíc výpočty s využitím součinových jader jsou jednodušší.

Poznámka 4.1. V literatuře se také využívá Gaussovo jádro $K(x, y) = (2\pi)^{-1} \mathrm{e}^{-(x^2+y^2)/2}$, které se zdá být výhodnějším při studiu asymptotických vlastností jádrového odhadu. Na druhou stranu má nevýhodu, že jeho nosičem je celá reálná osa, což způsobuje „nedokonalosť“ při odhadech hustot s omezeným definičním oborem.

5 Volba vyhlazovacího parametru

5.1 Metoda referenční hustoty

Předpokládejme, že náhodný výběr $([X_1, Y_1], \ldots, [X_n, Y_n])$ pochází z normálního rozdělení s hustotou

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \mathrm{e}^{-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}}$$

a jádro K je dvourozměrnou standardizovanou normální hustotou, tj.

$$K(x, y) = \frac{1}{2\pi} \mathrm{e}^{-\frac{x^2}{2} - \frac{y^2}{2}}.$$

Pak podle metody referenční hustoty lze získat tyto odhady vyhlazovacích parametrů

$$h_{i,\mathrm{REF}} = \widehat{\sigma}_i n^{-1/6}, \quad i = 1, 2.$$

Tento vztah je také znám jako Scottovo pravidlo ([11]).

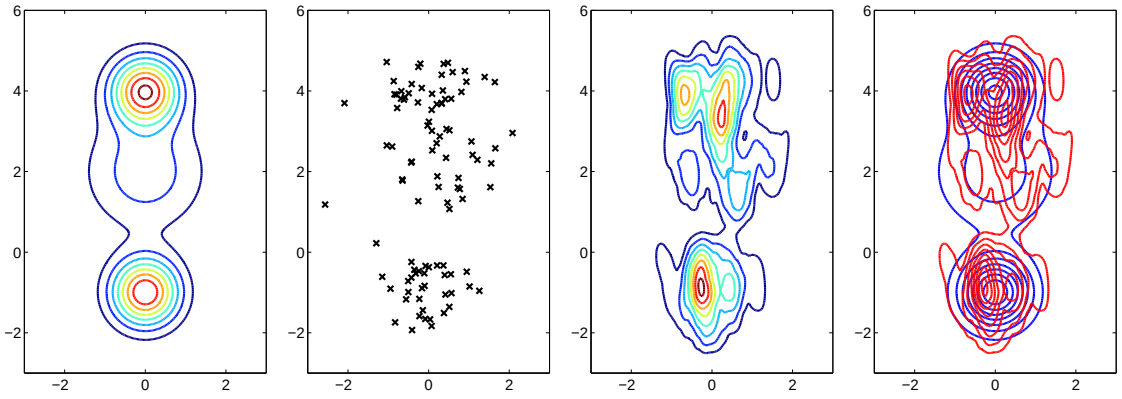
Příklad 5.1. Pro simulovaná data z příkladu 2.1 vychází matice vyhlazovacích parametrů podle metody referenční hustoty s využitím součinového Epanečnikova jádra takto:

$$\mathbf{H}_{\mathrm{REF}} = \begin{pmatrix} 0{,}3595 & 0 \\ 0 & 0{,}9972 \end{pmatrix}.$$

Na obrázku 5.5 je vykreslen odhad hustoty s touto maticí a porovnání odhadu se skutečnou hustotou.

Poznámka 5.1. V toolboxu, který doplňuje tato skripta, se uvádí druhá mocnina matice vyhlazovacích parametrů, tedy $\mathbf{H}^2$.

Z obrázku 5.5 je patrné, že jádrový odhad je podhlazený, zejména ve směru osy x. Je to způsobeno odhadem směrodatné odchylky, která je základem metody referenční hustoty.



(a) Skutečná hustota (b) Data (c) Odhad s $\mathbf{H}_{\mathrm{REF}}$ (d) Porovnání odhadu (—) se skutečnou hustotou (—)

Obrázek 5.5: Jádrový odhad dvourozměrné hustoty – referenční hustota

5.2 Metoda křížového ověřování

Metoda křížového ověřování je založena na odhadu hustoty v bodě $[X_i, Y_i]$ s vynecháním tohoto pozorování. Funkci metody křížového ověřování CV můžeme zapsat ve tvaru

$$\mathrm{CV}(\mathbf{H}) = \iint \left(\hat{f}(x, y, \mathbf{H})\right)^2 \mathrm{d}x\, \mathrm{d}y - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i, Y_i, \mathbf{H}).$$

kde

$$\hat{f}_{-i}(X_i, Y_i, \mathbf{H}) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_{\mathbf{H}}(X_i - X_j, Y_i - Y_j)$$

Někdy se metoda CV nazývá nevychýlená metoda křížového ověřování (*unbiased cross-validation*), důvodem je jednoduchý vztah mezi CV a MISE, který uvádí následující věta.

Věta 5.1. *Funkce* CV *je nevychýleným odhadem* MISE, *tj. platí*

$$E\,\mathrm{CV}(\mathbf{H}) = \mathrm{MISE}(\mathbf{H}) - \iint f^2(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

Důkaz. Vypočtěme střední hodnotu CV:

$$\begin{aligned} E\,\mathrm{CV}(\mathbf{H}) &= E \iint \hat{f}^2(x, y, \mathbf{H})\, \mathrm{d}x\, \mathrm{d}y - \frac{2}{n} \sum_{i=1}^{n} E\hat{f}_{-i}(X_i, Y_i, \mathbf{H}) \\ &= E \iint \hat{f}^2(x, y, \mathbf{H})\, \mathrm{d}x\, \mathrm{d}y - 2EK_{\mathbf{H}}(X_1 - X_2, Y_1 - Y_2) \\ &= E \iint \hat{f}^2(x, y, \mathbf{H})\, \mathrm{d}x\, \mathrm{d}y - 2 \iiiint K_{\mathbf{H}}(x - u, y - v) f(x, y) f(u, v)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}u\, \mathrm{d}v \end{aligned}$$

a úpravou MISE

$$\begin{aligned}\mathrm{MISE}(\mathbf{H}) &= E\iint \big(\hat{f}(x,y,\mathbf{H}) - f(x,y)\big)^2 \,\mathrm{d}x\,\mathrm{d}y \\ &= E\iint \hat{f}^2(x,y,\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y - 2E\int \hat{f}(x,y,\mathbf{H}) f(x,y)\,\mathrm{d}x\,\mathrm{d}y + \iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= E\iint \hat{f}^2(x,y,\mathbf{H})\,\mathrm{d}x\,\mathrm{d}y - 2\iiiint K_{\mathbf{H}}(x-u,y-v) f(x,y) f(u,v)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}u\,\mathrm{d}v \\ &\quad + \iint f^2(x,y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Porovnáním upravených výrazů dostaneme tvrzení. □

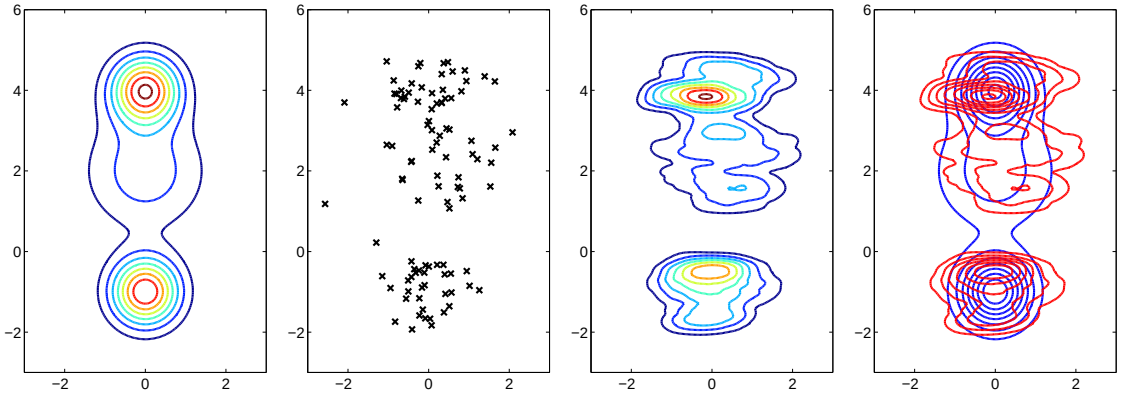
Optimální matici vyhlazovacích parametrů vzhledem k metodě CV označíme $\mathbf{H}_{\mathrm{CV}}$, tj.

$$\mathbf{H}_{\mathrm{CV}} = \arg\min_{\mathbf{H}\in\mathcal{D}} \mathrm{CV}(\mathbf{H}).$$

Příklad 5.2. Použijeme-li součinové Epanečnikovo jádro, pak pro simulovaná data z příkladu 2.1 dostanem matici vyhlazovacích parametrů určenou podle metody křížového ověřování v následujícím tvaru:

$$\mathbf{H}_{\mathrm{CV}} = \begin{pmatrix} 1{,}2055 & 0 \\ 0 & 0{,}3783 \end{pmatrix}.$$

Na obrázku 5.6 je vykreslen odhad hustoty s touto maticí.



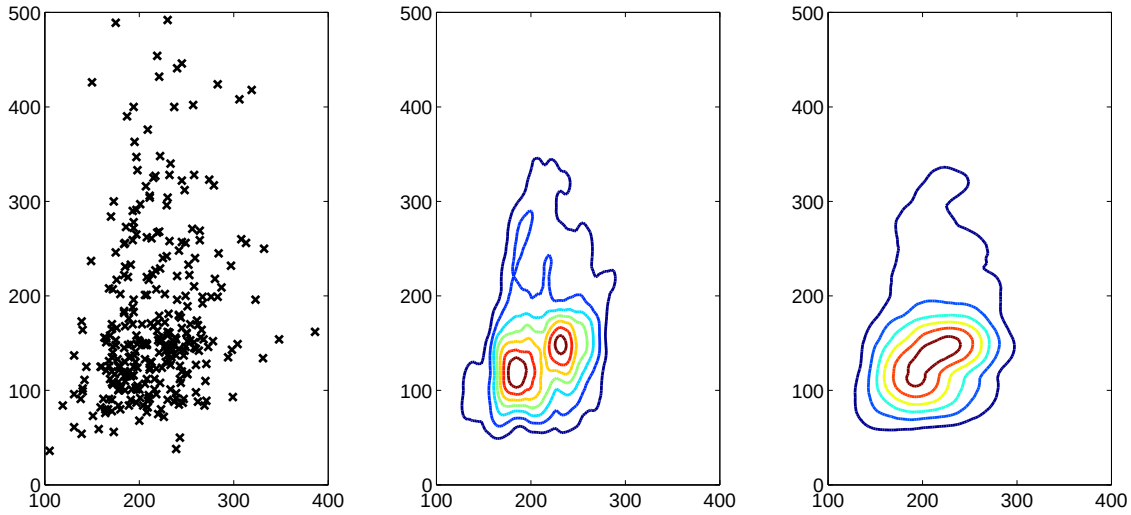
(a) Skutečná hustota (b) Data (c) Odhad s $\mathbf{H}_{\mathrm{CV}}$ (d) Porovnání odhadu (—) se skutečnou hustotou (—)

Obrázek 5.6: Jádrový odhad dvourozměrné hustoty – metoda křížového ověřování

Odhad hustoty na obrázku 5.6 se jeví podhlazený, zejména ve směru osy y. Metoda křížového ověřování nedává dobré výsledky pro odhad hustoty vícerozměrných dat, částečně je to způsobeno problémy při minimalizaci funkce křížového ověřování.

6 Aplikace na reálná data

Tento datový soubor pochází ze studie koncentrace lipidů v krevní plazmě, která vyšla v časopise Circulation v roce 1980. Výběrový soubor, který jsme převzali z [11] a s nímž zde pracujeme, obsahuje měření množství cholesterolu a triglyceridu v krevní plazmě u 320 pacientů, kteří si stěžovali na bolest v hrudníku. Data jsou shrnuta v tabulkách 6.9 a 6.10.



(a) Data (b) REF (c) CV

Obrázek 5.7: Vrstevnicové grafy odhadnutých hustot pro koncentraci lipidů – na ose x je vyneseno množství cholesterolu (v miligramech na 100 ml plazmy) a na ose y množství triglyceridu v krevní plazmě (mg/100 ml)

Matice vyhlazovacích parametrů určené podle metody referenční hustoty a metody křížového ověřování jsou následující:

$$\mathbf{H}_{\mathrm{REF}} = \begin{pmatrix} 16{,}45 & 0 \\ 0 & 38{,}94 \end{pmatrix} \qquad \mathbf{H}_{\mathrm{CV}} = \begin{pmatrix} 42{,}39 & 0 \\ 0 & 29{,}88 \end{pmatrix}$$

Na obrázku 5.7 jsou znázorněna data a vrstevnice jádrového odhadu s Epanečnikovým součinovým jádrem.

Shrnutí

Odhad dvourozměrné hustoty pravděpodobnosti $f(x, y)$ v bodě $[x, y]$ je tvaru

$$\hat{f}(x, y; \mathbf{H}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K\Big(\mathbf{H}^{-1}(x - X_i, y - Y_i)^T\Big)$$

Dva typy jader:

- součinové jádro: $K^P(x, y) = K_1(x) \cdot K_1(y)$,
- sféricky symetrické jádro: $K^S(x, y) = c_k^{-1} K_1\big(\sqrt{x^2 + y^2}\big)$, $c_k = \iint K_1\big(\sqrt{x^2 + y^2}\big)\,\mathrm{d}x\,\mathrm{d}y$.

Asymptotická střední integrální kvadratická chyba dvourozměrného jádrového odhadu

$$\mathrm{AMISE}(\mathbf{H}) = \frac{V(K)}{n h_1 h_2} + \frac{1}{4}\beta_2^2(K)\big(h_1^4 V(f_{xx}) + 2h_1^2 h_2^2 V(f_x f_y) + h_2^2 V(f_{yy})\big).$$

Optimální vyhlazovací parametry vzhledem k AMISE

$$h_{1,opt} = \left(\frac{V^{3/4}(f_{yy})V(K)}{n\beta_2^2(K)V^{3/4}(f_{xx})\big[V(f_x f_y) + \sqrt{V(f_{xx})V(f_{yy})}\big]}\right)^{1/6},$$

$$h_{2,opt} = \left(\frac{V^{3/4}(f_{xx})V(K)}{n\beta_2^2(K)V^{3/4}(f_{yy})\big[V(f_x f_y) + \sqrt{V(f_{xx})V(f_{yy})}\big]}\right)^{1/6}.$$

Metody pro odhad optimálních hodnot matice vyhlazovacích parametrů $\mathbf{H} = \mathrm{diag}(h_1, h_2)$

- metoda referenční hustoty

$$h_{i,\mathrm{REF}} = \widehat{\sigma}_i n^{-1/6}, \quad i = 1, 2,$$

- metoda křížového ověřování

$$\mathbf{H}_{\mathrm{CV}} = \arg\min_{\mathbf{H}\in\mathcal{D}} \mathrm{CV}(\mathbf{H}), \quad \mathrm{CV}(\mathbf{H}) = \iint \big(\hat{f}(x, y, \mathbf{H})\big)^2\,\mathrm{d}x\,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i, Y_i, \mathbf{H}).$$

Dodatky a cvičení

1. Určete součinové a sféricky symetrické dvourozměrné jádro odvozené z kvartického jádra $K(x) = \frac{15}{16}(1 - x^2)^2$.
2. Odvoďte vztahy (5.3) a (5.4) pro optimální vyhlazovací parametry.
3. Aplikujte metodu referenční hustoty a metodu křížového ověřování na simulovaná i reálná data.