Úvod

Vyhlazování je statistická technika pro rekonstrukci reálné funkce na základě pozorovaných nebo naměřených dat. Cílem vyhlazování je nalezení takového odhadu neznámé funkce, aby byly odfiltrovány náhodné výkyvy a bylo možné lépe poznat strukturu dat. K tomuto úkolu lze přistoupit dvěma způsoby – parametricky a neparametricky:

- *Parametrické odhady* jsou založeny na předpokladu, že neznámá funkce patří do třídy funkcí závislých na parametrech, a cílem je odhadnout tyto parametry.
- *Neparametrické odhady* nepředepisují datům „Prokrustovo lože" parametrizace, ale nechávají „hovořit samotná data".

V tomto učebním textu se zaměříme na neparametrické odhady, a to zejména na jádrové odhady, které patří mezi efektivní neparametrické odhady. Budeme se zabývat jádrovými odhady regresní funkce, hustoty, distribuční funkce a také odhadem dvourozměrné hustoty. Všechny jádrové odhady závisí na jádře, které má roli vahové funkce, a na šířce vyhlazovacího okna, která řídí hladkost odhadu.

Budeme zabývat následujícími otázkami:

- Jaké jsou statistické vlastnosti jádrových odhadů.
- Jaký vliv má tvar jádra na odhad.
- Jaký vliv má šířka vyhlazovacího okna na odhad.
- Jak lze tuto šířku stanovit v praxi.

Volba vhodného vyhlazovacího parametru je zásadním problémem ve všech typech jádrových odhadů a tomuto problému budeme věnovat značnou pozornost.

Příslušný toolbox pro program Matlab je dostupný na adrese:

`https://www.math.muni.cz/veda-a-vyzkum/vyvijeny-software/274-matlab-toolbox.html`

V příloze (kapitola 6) jsou uvedeny soubory dat pro samostatnou práci studentů. Tyto soubory již byly zpracovány v příslušných kapitolách a studenti si tak mohou ověřit správnost svých výsledků.

Definice základních statistických pojmů a jejich vlastností lze najít např. v elektronických skriptech Pravděpodobnost a statistika I autorů M. Forbelské a J. Koláčka (jsou dostupná na Elportálu Informačního systému).

Na tomto místě bych ráda poděkovala Mgr. Kamile Hasilové, Ph.D., za pomoc při sazbě tohoto textu a za příspěvek ke kapitole 5 a kapitolám o reálných datech.

Obsah

Seznam použitého značení

h	vyhlazovací parametr
$\mathbf{H}$	matice vyhlazovacích parametrů
$m(\cdot)$	regresní funkce
$f(\cdot)$	hustota pravděpodobnosti
$F(\cdot)$	distribuční funkce
$\hat{h}$	odhad vyhlazovacího parametru h
$\widehat{\sigma}$	odhad směrodatné odchylky σ
$\widehat{m}$	odhad regresní funkce m
$\hat{f}$	odhad hustoty f
$\widehat{F}$	odhad distribuční funkce F
$\int$	značí integrál $\int_{-\infty}^{\infty}$, pokud není uvedeno jinak
$K(\cdot)$	jádrová funkce (jádro)
$V(K)$	$V(K) = \int K^2(x)\,\mathrm{d}x$
$\beta_k(K)$	$\beta_k(K) = \int x^k K(x)\,\mathrm{d}x$
$f * g$	konvoluce funkcí f a g, $(f * g)(x) = \int f(t)g(x-t)\,\mathrm{d}t$
$W(\cdot)$	integrál z jádra, $W(x) = \int_{-\infty}^{x} K(t)\,\mathrm{d}t$
$C^k[0,1]$	prostor funkcí, které mají spojité derivace až do řádu k včetně na intervalu $[0,1]$
NW	Nadarayovy-Watsonovy odhady
PC	Priestleyovy-Chaovy odhady
LL	lokálně lineární odhady
GM	Gasserovy-Müllerovy odhady

MSE	střední kvadratická chyba (mean square error)
MISE	střední intergální kvadratická chyba (mean integrated square error)
AMISE	asymptotická střední integrální kvadratická chyba (asymptotic mean integrated square error)
AIV	asymptotický tvar integrálu rozptylu (asymptotic intergated variance)
AISB	asymptotický tvar integrálu druhé mocniny vychýlení (asymptotic integrated square bias)
AMSE	střední průměrná kvadratická chyba (average mean square error)

CV	metoda křížového ověřování
REF	metoda referenční hustoty
MS	metoda maximálního vyhlazení
PI	plug-in metoda

Kapitola 1

Jádrové funkce a jejich vlastnosti

Výstupy z výukové jednotky

Student

- bude znát základní třídy jádrových funkcí, jejich vlastnosti a metody jejich konstrukce.

1 Základní pojmy a definice

V úvodu bylo uvedeno, že všechny jádrové odhady závisí na jádrové funkci (jádře), a proto se v této kapitole budeme zabývat jádrovými funkcemi. Nyní uvedeme definici jádra a jeho vlastnosti.

Definice 1.1. Nechť ν, k jsou nezáporná celá čísla, $0 \leq \nu < k$, nechť K je reálná funkce s těmito vlastnostmi

1. K splňuje Lipschitzovu podmínku na intervalu $[-1,1]$, tj. $|K(x) - K(y)| \leq L|x-y|$ pro $\forall x, y \in [-1,1]$, $L > 0$,
2. nosič$(K) = [-1,1]$, tj. $K = 0$ vně intervalu $[-1,1]$,
3. K splňuje momentové podmínky:

$$\int_{-1}^{1} x^j K(x)\,\mathrm{d}x = \begin{cases} 0 & 0 \leq j < k, j \neq \nu, \\ (-1)^\nu \nu! & j = \nu \end{cases} \tag{1.1}$$

   a $\int_{-1}^{1} x^k K(x)\,\mathrm{d}x \neq 0$, tuto hodnotu označíme $\beta_k(K)$.

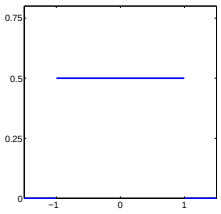
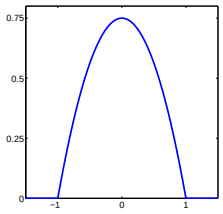
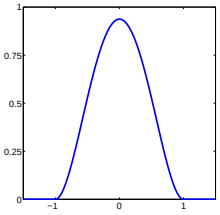
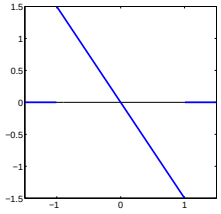
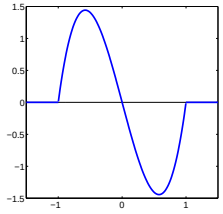
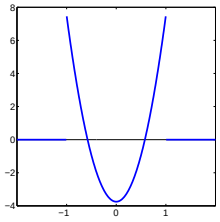
Taková funkce K se nazývá *jádro řádu* k a třída všech těchto funkcí se označuje $S_{\nu k}$.

Příklad 1.1. V tabulce 1.1 jsou uvedeny příklady několika jader společně s jejich grafy. Přitom funkce $I_{[-1,1]}(x)$ je indikátorová funkce intervalu $[-1,1]$, tj.

$$I_{[-1,1]}(x) = \begin{cases} 1 & \text{pro } x \in [-1,1], \\ 0 & \text{jinak.} \end{cases}$$

Nyní uvedeme dva typy jader – jádra s minimálním rozptylem a optimální jádra.

Tabulka 1.1: Jádra

<table>
<tr><td colspan="3">S_{02}</td></tr>
<tr><td>$K(x) = \frac{1}{2} I_{[-1,1]}(x)$
obdélníkové jádro
  </td><td>$K(x) = \frac{3}{4}(1-x^2) I_{[-1,1]}(x)$
Epanečnikovo jádro
  </td><td>$K(x) = \frac{15}{16}(1-x^2)^2 I_{[-1,1]}(x)$
kvartické jádro
  </td></tr>
<tr><td colspan="2">S_{13}</td><td>S_{24}</td></tr>
<tr><td>$K(x) = -\frac{3}{2} x I_{[-1,1]}(x)$
  </td><td>$K(x) = -\frac{15}{4} x(1-x^2) I_{[-1,1]}(x)$
  </td><td>$K(x) = -\frac{15}{4}(1-3x^2) I_{[-1,1]}(x)$
  </td></tr>
</table>

1.1 Jádra s minimálním rozptylem

Předpokládejme, že $K \in S_{\nu k}$, $0 \le \nu \le k-2$, ν a k jsou obě sudá nebo lichá [1]. Uvažujme funkcionál $V(K) = \int_{-1}^{1} K^2(x)\,\mathrm{d}x$ a zabývejme se problémem najít takové jádro $K \in S_{\nu k}$, pro které tento funkcionál nabývá minimální hodnoty, tj. řešíme variační úlohu

$$\min V(K) \quad \text{za předpokladu } K \in S_{\nu k}.$$

Řešení této úlohy se nazývají *jádra s minimálním rozptylem*, což jsou polynomy stupně $k-2$ na intervalu $[-1,1]$. Tyto polynomy jsou sudé funkce pro k sudé a liché funkce pro k liché a mají $k-2$ různých kořenů v intervalu $(-1,1)$. Obecný vztah pro jádra s minimálním rozptylem lze nalézt ve [3].

Příklad 1.2. Jádra s minimálním rozptylem:

$$\begin{aligned} S_{02}&: \quad K(x) = \tfrac{1}{2} I_{[-1,1]}(x) \\ S_{13}&: \quad K(x) = -\tfrac{3}{2} x I_{[-1,1]}(x) \\ S_{24}&: \quad K(x) = -\tfrac{15}{4}(1-3x^2) I_{[-1,1]}(x) \end{aligned}$$

Poznámka 1.1. Jádra s minimálním rozptylem mají skoky v koncových bodech intervalu $[-1,1]$, což negativně ovlivňuje hladkost výsledného odhadu.

[1] Obvykle se používá pojem *parita*, tedy ν a k mají stejnou paritu.

Tabulka 1.2: Optimální jádra pro $\nu = 0, 1, 2$

$\nu = 0$		
k	δ_{0k}	$K_{opt,0,k}$
2	1,7188	$-\frac{3}{4}(x^2-1)$
4	2,0165	$\frac{15}{32}(x^2-1)(7x^2-3)$
6	2,0834	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$
8	2,1021	$\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$
$\nu = 1$		
k	δ_{1k}	$K_{opt,1,k} = K^{(1)}$
3	1,4204	$\frac{15}{4}x(x^2-1)$
5	1,7656	$-\frac{105}{32}x(x^2-1)(9x^2-5)$
7	1,8931	$\frac{315}{32}x(x^2-1)(143x^4-154x^2+35)$
$\nu = 2$		
k	δ_{2k}	$K_{opt,2,k} = K^{(2)}$
4	1,3925	$-\frac{105}{16}(x^2-1)(5x^2-1)$
6	1,6964	$\frac{315}{64}(x^2-1)(77x^4-58x^2+5)$
8	1,8269	$-\frac{3465}{2048}(x^2-1)(1755x^6-2249x^4+721x^2-35)$

1.2 Optimální jádra

Při vyšetřování statistických vlastností se setkáváme s následujícím funkcionálem

$$T(K) = \Bigg(\Big| \underbrace{\int_{-1}^{1} x^k K(x)\,\mathrm{d}x}_{\beta_k(K)} \Big|^{2\nu+1} \Big(\underbrace{\int_{-1}^{1} K^2(x)\,\mathrm{d}x}_{V(K)} \Big)^{k-\nu} \Bigg)^{\frac{2}{2k+1}},$$

který lze zkráceně psát $T(K) = \left(|\beta_k(K)|^{2\nu+1} V(K)^{(k-\nu)}\right)^{2/(2k+1)}$. Jádra, pro která tento funkcionál nabývá minimální hodnoty, se nazývají *optimální jádra*. Jde o polynomy stupně k, které mají $k-2$ různých kořenů v intervalu $(-1,1)$ a body $-1, 1$ jsou rovněž kořeny těchto polynomů.

Příklad 1.3. Optimální jádra:

$$\begin{aligned}
S_{02}\text{:}&\quad K_{opt,0,2}(x) = \tfrac{3}{4}(1-x^2) I_{[-1,1]}(x) \\
S_{13}\text{:}&\quad K_{opt,1,3}(x) = \tfrac{15}{4} x(x^2-1) I_{[-1,1]}(x) \\
S_{24}\text{:}&\quad K_{opt,2,4}(x) = -\tfrac{105}{16}(x^2-1)(5x^2-1) I_{[-1,1]}(x)
\end{aligned}$$

Přehled vybraných optimálních jader je uveden v tabulce 1.2. Obecný vzorec pro tvar optimálních jader lze nalézt např. v [3].

Jádra $K_{opt,1,k}$ a $K_{opt,2,k}$ se používají pro odhad první a druhé derivace hustoty (viz kapitola 3.1) a z toho důvodu pro ně zavedeme dodatečné označení $K^{(1)}$, respektive $K^{(2)}$.

Užitečný při odvozování statistických vlastností odhadů bude i následující pojem.

Definice 1.2. Pro jádro $K \in S_{\nu k}$ definujeme *kanonický faktor*

$$\delta_{\nu k} = \left(\frac{V(K)}{\beta_k^2(K)} \right)^{\frac{1}{2k+1}}.$$

Shrnutí
Funkcionály závisející na jádře K $\beta_k(K) = \int_{-1}^{1} x^k K(x)\,\mathrm{d}x \qquad V(K) = \int_{-1}^{1} K^2(x)\,\mathrm{d}x$ $T(K) = \left(\lvert\beta_k(K)\rvert^{2\nu+1} V(K)^{(k-\nu)} \right)^{\frac{2}{2k+1}} \qquad \delta_{\nu k} = \left(\frac{V(K)}{\beta_k^2(K)} \right)^{\frac{1}{2k+1}}$
Jádra s minimálním rozptylem minimalizují funkcionál $V(K)$. Optimální jádra minimalizují funkcionál $T(K)$.
Podrobnější popis jader a vzorců s nimi souvisejících lze nalézt v toolboxu Matlabu.

Dodatky a cvičení

1. Tvrzení: Nechť $K \in S_{\nu+1,k+1}$ je jádro s minimálním rozptylem a $K_{opt,\nu,k} \in S_{\nu k}$ je optimální jádro. Pak platí
$$K'_{opt,\nu,k}(x) = K(x), \quad x \in [-1, 1]$$
Důkaz viz [3]. Jako příklad lze uvést jádro $K_{opt,0,2}(x) = \frac{3}{4}(1 - x^2) \in S_{02}$ a $K'_{opt,0,2}(x) = K_{1,3}(x) = -\frac{3}{2}x \in S_{13}$.

2. Nechť $K \in S_{\nu k}$, $\nu \in \mathbb{N}$, pro $\delta > 0$ položíme
$$K_\delta(\cdot) = \frac{1}{\delta^{\nu+1}} K\left(\frac{\cdot}{\delta}\right).$$
Jádra K, K_δ nazýváme *ekvivalentní*. Dokažte: Funkcionál $T(K)$ je invariantní vzhledem k této transformaci, tj. $T(K) = T(K_\delta)$.

3. Nechť jsou dány funkce f a g, pro které platí $\int f^2(x)\,\mathrm{d}x < \infty$ a $\int g^2(x)\,\mathrm{d}x < \infty$. *Konvoluci* $f * g$ definujeme vztahem
$$(f * g)(x) = \int f(t) g(x - t)\,\mathrm{d}t.$$
Vlastnosti konvoluce
   - $f * g = g * f$,
   - $f * (g * h) = (f * g) * h$,
   - $f * (g + h) = f * g + f * h$.

   Ukažte, že pro jádrovou funkci $K \in S_{02}$ platí vztah
$$\int (K * K)(x) x^j\,\mathrm{d}x = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \\ 2\beta_2(K) & j = 2. \end{cases}$$

Kapitola 2

Jádrové odhady regresní funkce

Výstupy z výukové jednotky

Student

- bude znát základní typy jádrových odhadů regresní funkce a jejích derivací.
- bude schopen analyzovat statistické vlastnosti odhadů.
- bude mít přehled o metodách pro volbu vyhlazovacího parametru.
- se seznámí s automatickou procedurou pro simultánní volbu vyhlazovacího parametru, jádra a jeho řádu.
- bude schopen analyzovat daný soubor dat a aplikovat uvedenou proceduru na tento soubor.
- bude schopen použít příslušný toolbox v Matlabu a zkonstruovat odhad regresní funkce.

1 Motivace

Předpokládejme, že pro pevné nebo náhodné hodnoty nezávisle proměnné X máme k dispozici naměřené hodnoty závisle proměnné Y. Chceme-li tato data analyzovat, musíme nalézt vhodný funkční vztah mezi těmito proměnnými.

Jestliže dvojice bodů $(x_i, Y_i), i = 1, \ldots, n$, znázorníme graficky, pak pouhý pohled na takový dvourozměrný bodový diagram obvykle nestačí k tomu, abychom určili tento funkční vztah. *Statistická úloha, kterou se budeme zabývat, spočívá v proložení vhodné křivky těmito body tak, aby byly odfiltrovány náhodné výkyvy a bylo možné lépe poznat strukturu dat. Tuto křivku nazýváme* regresní křivkou.

Formalizujme nyní tuto úlohu: Uvažujme standardní regresní model

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

kde m je neznámá regresní funkce, x_i, $i = 1, \ldots, n$, jsou body plánu a ε_i, $i = 1, \ldots, n$, jsou chyby měření, o nichž se předpokládá, že jsou nezávislými, identicky rozdělenými náhodnými veličinami splňujícími podmínky

$$E\varepsilon_i = 0, \quad \operatorname{var} \varepsilon_i = \sigma^2, \quad i = 1, \ldots, n. \tag{2.2}$$

Poznámka 1.1. Jsou-li body plánu uspořádaná nenáhodná čísla, mluvíme o regresním modelu s pevným plánem. V případě, že body plánu $X_1, \ldots, X_n$ jsou náhodné veličiny se stejnou hustotou f, jedná se o regresní model s náhodným plánem (podrobněji např. [14]). Budeme se dále zabývat modelem s pevným plánem.

Bez újmy na obecnosti budeme v dalším předpokládat, že pro body x_i, $i = 1, \ldots, n$, platí

$$0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1.$$

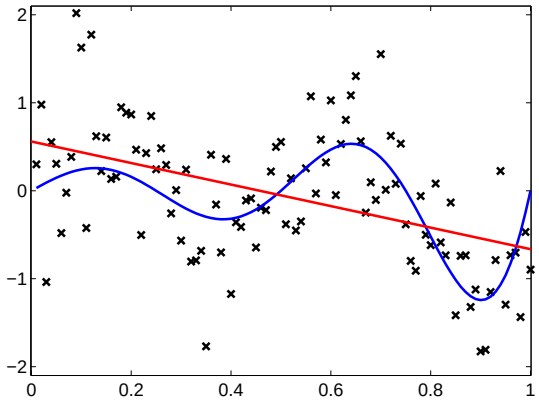
Cílem regresní analýzy je nalézt vhodnou aproximaci $\widehat{m}$ neznámé funkce m. Tento proces odhadu regresní funkce se obvykle nazývá *vyhlazování*. K tomuto úkolu lze přistoupit dvěma způsoby – parametricky a neparametricky. Příkladem parametrického odhadu regresní funkce je regresní přímka vyjadřující lineární závislost. Naopak u neparametrického přístupu nepředpokládáme, že funkce má nějaký předepsaný tvar, pouze předpokládáme jistou hladkost odhadované funkce (tj. dostatečný počet spojitých derivací).

V první polovině dvacátého století byla věnována pozornost zejména parametrickým metodám. V posledních letech však zaznamenaly značný rozvoj neparametrické metody. Tento vývoj souvisí s rostoucími požadavky na zpracování dat, ať už jde o rozsah souborů, rozmanitost těchto dat apod. Čistě parametrický přístup nevyhovuje vždy potřebám flexibility a nebývalý rozmach výpočetní techniky vytvořil dobré předpoklady pro rozvoj neparametrických metod. I přes tento vývoj si oba způsoby zachovávají své výhody a nijak si nekonkurují. Někdy je vhodné užít neparametrické metody a pak na výsledný odhad použít parametrickou metodu.

Příklad 1.1. Obr. 2.1 ilustruje na simulovaných datech nevhodnost aplikace parametrického přístupu. V tomto případě byla data generována podle vztahu

$$Y_i = \frac{\sin 4\pi x_i}{(1 + \cos 0{,}6\pi x_i)^2} + \varepsilon_i,$$

kde body $x_i = i/100$, $i = 1, \ldots, 100$, a chyby ε_i, $i = 1, \ldots, 100$, mají normální rozdělení $N(0; 0{,}25)$. (Data jsou v tabulce 6.1.)



Obrázek 2.1: Simulovaná data (×) s regresní přímkou (—) a původní funkcí (—)

Předpokládejme, že hledaná křivka je přímka a známou metodou nejmenších čtverců určíme rovnici této přímky. Obr. 2.1 znázorňuje přesnou funkci, generovaná data a výslednou přímku. Je zřejmé, že náš předpoklad, že hledaná funkce je přímka, není správný.

2 Základní typy neparametrických odhadů

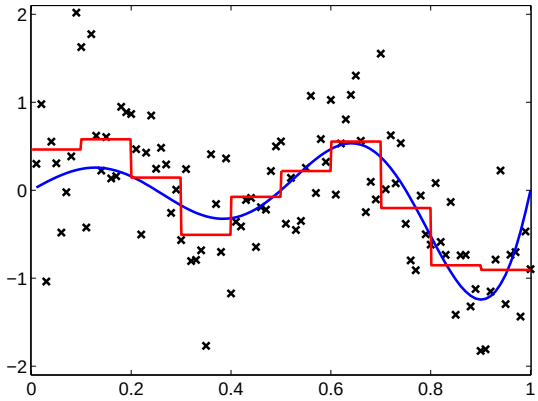
Pokud jde o historii neparametrických metod, připomeňme, že v r. 1857 saský ekonom Engel analyzoval data týkající se nákladů na domácnost a pro vyjádření závislosti použil schodovitou (tj. po částech konstantní funkcí), kterou dnes nazýváme *regresogram*. Regresogram užívá stejné základní myšlenky jako histogram pro odhad hustoty. Tato myšlenka spočívá v rozdělení

množiny hodnot proměnné X na intervaly B_j, $j = 1, \ldots, J$, a za odhad v bodě $x \in B_j$ se vezme průměr hodnot Y na tomto subintervalu, tj.

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i I_{[B_j]}(x_i)}{\sum_{i=1}^{n} I_{[B_j]}(x_i)},$$

kde $I_{[B_j]}$ je indikátorová funkce subintervalu B_j.

Výsledek aplikace regresogramu na simulovaná data příkladu 1.1 je znázorněn na obr. 2.2. Vidíme, že tento odhad „vhodně“ vystihuje tvar funkce, ale výsledný odhad je příliš hrubý.

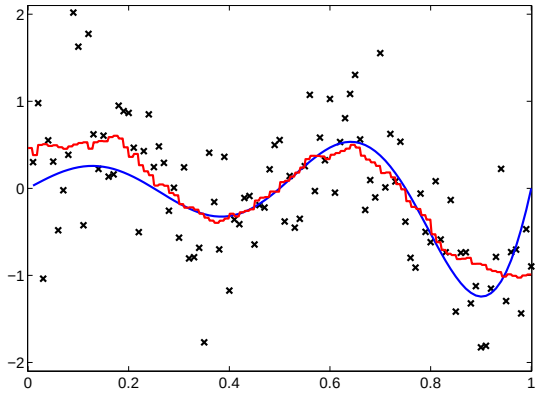


Obrázek 2.2: Regresogram (—) pro simulovaná data (×) z příkladu 1.1 s původní funkcí (—)

Přirozeným zobecněním regresogramu je *metoda klouzavých průměrů*. Tato metoda používá lokálních průměrů hodnot Y, ale odhad v bodě x je založen na centrovaném okolí bodu x: $[x - h, x + h]$, $h > 0$, přesněji

$$\widehat{m}(x, h) = \frac{\sum_{i=1}^{n} Y_i I_{[x-h,x+h]}(x_i)}{\sum_{i=1}^{n} I_{[x-h,x+h]}(x_i)}. \tag{2.3}$$

Obr. 2.3 ilustruje aplikaci této metody na simulovaných datech příkladu 1.1. Uvedené metody



Obrázek 2.3: Klouzavý průměr (—) pro simulovaná data z příkladu 1.1

patří mezi nejjednodušší neparametrické vyhlazovací metody. Jádrové odhady lze považovat za zobecnění těchto metod.

Připomeňme zde základní myšlenku vyhlazování tak, jak ji formuloval R. Eubank v r. 1988:

> *Jestliže předpokládáme, že m je hladká funkce, pak pozorování v bodech x_i blízko bodu x obsahují informace o hodnotě m v bodě x. Bylo by tedy vhodné užít lokálních průměrů dat blízko bodu x, abychom získali odhad $m(x)$.*

Obecně lze jádrové odhady regresní funkce m v bodě x definovat takto

$$\widehat{m}(x,h) = \sum_{i=1}^{n} W_i(x,h) Y_i, \tag{2.4}$$

kde funkce W_i, $i = 1, \ldots, n$, se nazývají váhy, nezávisí na Y, ale závisí na kladném čísle h, které se nazývá *vyhlazovací parametr* (nebo také šířka vyhlazovacího okna). Speciální, velmi užitečný typ W, závisí na jádrové funkci K.

Nechť $K \in S_{0k}$, k je sudé číslo, položme $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$. Mezi nejznámější typy jádrových odhadů regresní funkce patří:

1. Nadarayovy-Watsonovy odhady (1964)

$$\widehat{m}_{NW}(x,h) = \frac{\sum_{i=1}^{n} K_h(x - x_i) Y_i}{\sum_{i=1}^{n} K_h(x - x_i)},$$

2. Priestleyovy-Chaovy odhady (1972)

$$\widehat{m}_{PC}(x,h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) Y_i,$$

3. lokálně lineární odhady (Stone 1977, Cleveland 1979)

$$\widehat{m}_{LL}(x,h) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\hat{s}_2(x,h) - \hat{s}_1(x,h)(x - x_i)\} K_h(x - x_i) Y_i}{\hat{s}_2(x,h)\hat{s}_0(x,h) - \hat{s}_1(x,h)^2},$$

   kde

$$\hat{s}_r(x) = \frac{1}{n} \sum_{i=1}^{n} (x - x_i)^r K_h(x - x_i), \quad r = 0, 1, 2,$$

4. Gasserovy-Müllerovy odhady (1979)

$$\widehat{m}_{GM}(x,h) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(x - t)\, \mathrm{d}t,$$

   kde $s_0 = 0$, $s_i = (x_i + x_{i+1})/2$, $s_n = 1$. Tento odhad je konvolučním typem odhadu.

Úmluva. Uvedené jádrové odhady budeme zapisovat ve tvaru

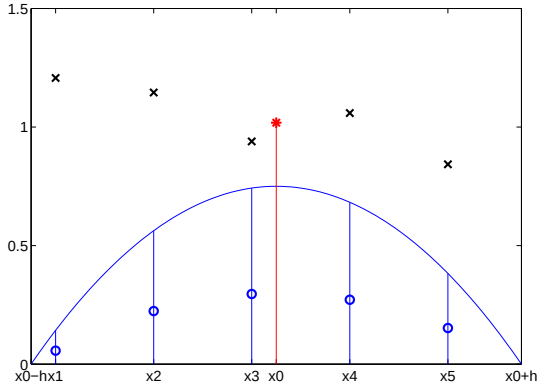
$$\widehat{m}_j(x,h) = \sum_{i=1}^{n} W_i^{(A)}(x,h) Y_i,$$

kde váhy $W_i^{(A)}$ a A značí příslušný typ odhadu NW, PC, LL, GM. V dalším textu budeme zkráceně psát W_i.

V mnoha aplikacích je užitečný zejména Nadarayův-Watsonův odhad $\widehat{m}_{NW}$. Popíšeme nyní jeho konstrukci a budeme ilustrovat vliv vyhlazovacího parametru na kvalitu odhadu. Pro daný bod x_0, $h < x_0 < 1 - h$, jsou váhy Nadarayova-Watsonova odhadu dány vztahem

$$W_i(x_0, h) = \frac{K_h(x_0 - x_i)}{\sum_{j=1}^{n} K_h(x_0 - x_j)}, \quad \sum_{j=1}^{n} W_j(x_0, h) = 1.$$

Obrázek 2.4 ilustruje konstrukci odhadu v bodě x_0, který je založen na pěti pozorováních (x_1, Y_1), $\ldots$, (x_5, Y_5) (černé křížky). Parabola reprezentuje Epanečnikovo jádro K_h a kroužky znázorňují hodnoty vah $W_i = K_h(x_0 - x_i) / \sum_{i=1}^{5} K_h(x_0 - x_i)$ pro $i = 1, \ldots, 5$. Výsledný odhad regresní funkce $\widehat{m}$ v bodě x_0 je označen hvězdičkou.



Obrázek 2.4: Ilustrace Nadarayova-Watsonova odhadu v bodě x_0

Jádrový odhad není definován pro $\sum_{i=1}^{n} K_h(x - x_i) = 0$. Jestliže nastane případ „0/0“, pak klademe $\widehat{m}_{NW}(x, h) = 0$. Omezíme se nyní na odhady funkce m v bodech plánu x_i, $i = 1, \ldots, n$.

Pro malé hodnoty h je výraz $\left|\frac{x_i - x_j}{h}\right| > 1$ pro $x_i \neq x_j$, a tedy hodnota jádra v těchto bodech je rovna nule, s výjimkou bodu x_i, v němž dostáváme odhad

$$\widehat{m}_{NW}(x_i, h) \to \frac{K(0) Y_i}{K(0)} = Y_i.$$

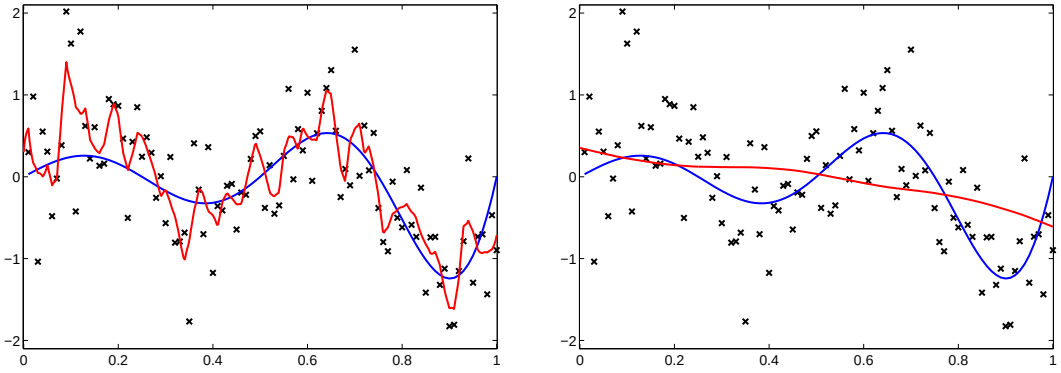
To znamená, že při malé šířce vyhlazovacího okna ($h \to 0$) odhad reprodukuje data (viz obr. 2.5(a)).

Podobně pro velké hodnoty h je výraz $\left|\frac{x_i - x_j}{h}\right| \approx 0$, tedy pro všechny body plánu máme stejnou hodnotu jádrové funkce $K\left(\frac{x_i - x_j}{h}\right) \approx K(0)$ a dostaneme tak průměr dat

$$\widehat{m}_{NW}(x_i, h) \to \frac{\sum_{j=1}^{n} K(0) Y_j}{\sum_{j=1}^{n} K(0)} = \frac{K(0) \sum_{j=1}^{n} Y_j}{n K(0)} = \frac{1}{n} \sum_{j=1}^{n} Y_j$$

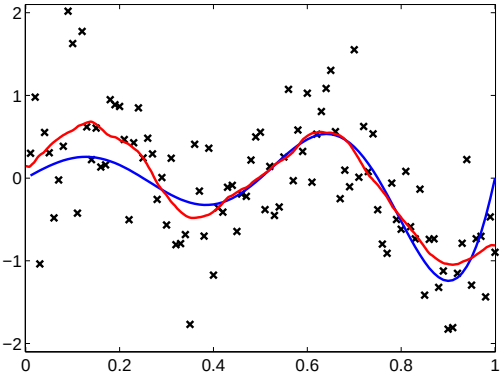
Tedy velká šířka okna ($h \to \infty$) vede k přehlazení, a to k průměru dat (viz obr. 2.5(b)).

Na obrázku 2.6 je znázorněn odhad s Epanečnikovým jádrem. Tento odhad se nejvíce blíží skutečné regresní funkci. Pokud jde o volbu vyhlazovacího parametru, je třeba si uvědomit, že konečné rozhodnutí o odhadované křivce je částečně subjektivní, neboť i asymptoticky optimální odhady obsahují poměrně značné „množství šumu“ a to nechává prostor pro subjektivní posouzení.



(a) Podhlazený odhad, $h = 0{,}02$ (b) Přehlazený odhad, $h = 0{,}4$

Obrázek 2.5: Podhlazený a přehlazený odhad regresní funkce z příkladu 1.1



Obrázek 2.6: Optimální odhad, $h = 0{,}0816$

Poznámka 2.1. *Intervaly spolehlivosti* pro hodnotu regresní funkce m v bodě x jsou užitečné v mnoha aplikacích. Bodový interval spolehlivosti udává interval, v němž s pravděpodobností $1 - \alpha$ leží hodnota funkce m v bodě x. Jsou definovány takto

$$\left[\widehat{m}(x,h) - u_{1-\frac{\alpha}{2}}\sqrt{\frac{V(K)\widehat{\sigma}^2(x)}{nh}}, \widehat{m}(x,h) + u_{1-\frac{\alpha}{2}}\sqrt{\frac{V(K)\widehat{\sigma}^2(x)}{nh}}\right],$$

kde $u_{1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$-kvantil standardního normálního rozdělení a odhad rozptylu v bodě x je dán vztahem

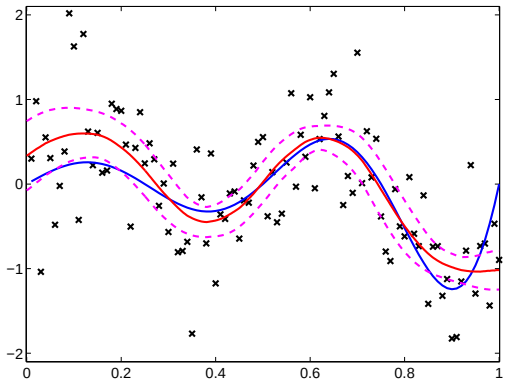
$$\widehat{\sigma}^2(x) = \sum_{i=1}^{n} W_i(x,h)\big(Y_i - \widehat{m}(x,h)\big)^2.$$

Ukázka intervalu spolehlivosti pro $\alpha = 0{,}05$ je na obrázku 2.7.

3 Statistické vlastnosti jádrových odhadů

Kvalitu jádrového odhadu lze lokálně popsat pomocí střední kvadratické chyby (MSE) odhadu $\widehat{m}$ v bodě x, která je obecně dána vztahem

$$\operatorname{MSE}\widehat{m}(x,h) = E\big(\widehat{m}(x,h) - m(x)\big)^2.$$

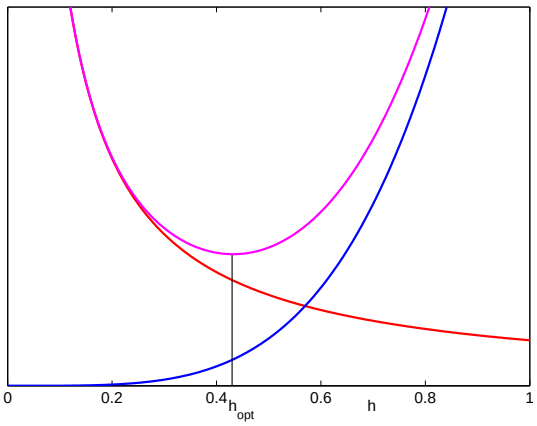


Obrázek 2.7: Interval spolehlivosti pro data z příkladu 1.1 při $\alpha = 0{,}05$

Upravíme tento vztah

$$\begin{aligned} &= E\widehat{m}^2(x,h) - 2m(x)E\widehat{m}(x,h) + m^2(x) \\ &= \underbrace{\big(E\widehat{m}(x,h) - m(x)\big)^2}_{\text{bias}^2} + \underbrace{E\widehat{m}^2(x,h) - \big(E\widehat{m}(x,h)\big)^2}_{\text{var}}, \end{aligned} \tag{2.5}$$

což znamená, že střední kvadratická chyba může být vyjádřena jako součet *rozptylu odhadu* $\operatorname{var}\widehat{m}(x,h)$ a čtverce *vychýlení* $\operatorname{bias}^2\widehat{m}(x,h)$ (viz obr. 2.8). Tento rozklad *rozptyl-vychýlení* usnadňuje analýzu vlastností odhadu.



Obrázek 2.8: MSE (—) jako součet rozptylu (—) a vychýlení (—)

Všechny uvedené jádrové odhady regresní funkce $\widehat{m}_{NW}$, $\widehat{m}_{PC}$, $\widehat{m}_{LL}$, $\widehat{m}_{GM}$ jsou asymptoticky ekvivalentní (viz např. [8, 14]). Z tohoto důvodu budeme dále podrobněji zabývat *Priestleyovými-Chaovými odhady*, které budeme psát bez uvedení označení PC, tedy: $\widehat{m}(x,h)$ a W_i.

Připomeňme, že pro Priestleyovy-Chaovy odhady je váhová funkce tvaru

$$W_i(x,h) = K_h(x - x_i) = \frac{1}{h}K\Big(\frac{x - x_i}{h}\Big).$$

Pro další výpočty budeme předpokládat:

(i) Jádrová funkce K je sudou funkcí na intervalu $[-1,1]$, tj. $K \in S_{02}$,

(ii) vyhlazovací parametr $h = h(n)$ je posloupností splňující $h \to 0$ a $nh \to \infty$ pro $n \to \infty$,

(iii) bod x, v němž počítáme odhad, splňuje nerovnost $h < x < 1 - h$ pro všechna $n \geq n_0$, kde n_0 je pevné,

(iv) $m \in C^2[0,1]$,

(v) $x_i = \frac{i}{n}$, $i = 1, \ldots, n$.

Je zřejmé, že pro $n \to \infty$ platí[1]

$$E\widehat{m}(x,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)m(x_i) = \int_0^1 \frac{1}{h}K\Big(\frac{x-t}{h}\Big)m(t)\,\mathrm{d}t + O(n^{-1}). \tag{2.6}$$

Symbolika O a o viz příloha 1. Nechť $u = \frac{x-t}{h}$, odtud $\mathrm{d}t = -h\,\mathrm{d}u$ a tedy s využitím Taylorova rozvoje

$$\begin{aligned} E\widehat{m}(x,h) &= \int_{-(1-x)/h}^{x/h} K(u)m(x - hu)\,\mathrm{d}u + O(n^{-1}) \\ &= \int_{-(1-x)/h}^{x/h} K(u)\Big[m(x) - uhm'(x) + \frac{u^2h^2}{2}m''(x)\Big]\,\mathrm{d}u + o(h^2) + O(n^{-1}). \end{aligned} \tag{2.7}$$

Podle výše uvedených předpokladů platí $h < x < 1 - h$, tedy $x/h \to \infty$ a $-(1-x)/h \to -\infty$ pro $h \to 0$. Odtud, s využitím faktu, že nosičem funkce K je interval $[-1,1]$, plyne

$$E\widehat{m}(x,h) = m(x)\int_{-1}^{1} K(u)\,\mathrm{d}u - hm'(x)\int_{-1}^{1} uK(u)\,\mathrm{d}u + \frac{h^2}{2}m''(x)\int_{-1}^{1} u^2K(u)\,\mathrm{d}u + o(h^2) + O(n^{-1}).$$

Celkem dostaneme

$$\begin{aligned} E\widehat{m}(x,h) &= m(x) + \frac{h^2}{2}\beta_2(K)m''(x) + o(h^2) + O(n^{-1}) \\ &\approx m(x) + \frac{h^2}{2}\beta_2(K)m''(x). \end{aligned}$$

Podobně pro rozptyl platí

$$\begin{aligned} \operatorname{var}\widehat{m}(x,h) &= E\big(\widehat{m}(x,h) - E\widehat{m}(x,h)\big)^2 \\ &= E\bigg(\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)Y_i - \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)m(x_i)\bigg)^2 \\ &= \frac{1}{n^2}E\bigg(\sum_{i=1}^{n} K_h(x - x_i)\underbrace{(Y_i - m(x_i))}_{\varepsilon_i}\bigg)^2 \end{aligned}$$

z vlastností (2.2) plyne $EK_h(x_i)K_h(x_j)\varepsilon_i\varepsilon_j = 0$ pro $i \neq j$, tedy

$$\begin{aligned} &= \frac{1}{n^2}\sum_{i=1}^{n} K_h^2(x - x_i)E\varepsilon_i^2 \\ &= \frac{\sigma^2}{n^2h^2}\sum_{i=1}^{n} K^2\Big(\frac{x - x_i}{h}\Big) = \frac{\sigma^2}{nh}\left(\int_0^1 \frac{1}{h}K^2\Big(\frac{x-t}{h}\Big)\,\mathrm{d}t + O(n^{-1})\right). \end{aligned}$$

Opět s využitím substituce $u = \frac{x-t}{h}$ a vztahu $O(n^{-1}) = o((nh)^{-1})$ můžeme pro $n \to \infty$ psát

$$\operatorname{var}\widehat{m}(x,h) = \frac{\sigma^2}{nh}\int_{-1}^{1} K^2(u)\,\mathrm{d}u + o((nh)^{-1}) = \frac{\sigma^2}{nh}\int_{-1}^{1} K^2(u)\,\mathrm{d}u + o((nh)^{-1}) \approx \frac{\sigma^2}{nh}V(K).$$

Tímto jsme dokázali následující větu o tvaru střední kvadratické chyby.

[1] Jedná se o přibližný výpočet integrálu – viz Dodatek na str. 27

Věta 3.1. *Nechť jsou splněny předpoklady* $(i) - (iii)$*, pak střední kvadratická chyba nabývá tvaru*

$$\operatorname{MSE} \widehat{m}(x,h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{var}} + \underbrace{h^4 \beta_2^2(K) \frac{1}{4} \big(m''(x)\big)^2}_{\text{bias}^2} + o(h^4 + (nh)^{-1}). \tag{2.8}$$

Chyba MSE dává pouze lokální pohled na chybu odhadu, proto se častěji používá globální tvar chyby – AMISE – asymptotická střední intergální kvadratická chyba. AMISE je součástí střední integrální kvadratické chyby (MISE) a vztah mezi chybami MSE, MISE a AMISE je následující

$$\operatorname{MISE}(h) = \int_0^1 \operatorname{MSE}(x,h)\,\mathrm{d}x = \operatorname{AMISE}(h) + o(h^4 + (nh)^{-1}).$$

AMISE je tvaru

$$\operatorname{AMISE}(h) = \operatorname{AMISE} \widehat{m}(\cdot,h) = \underbrace{\frac{\sigma^2}{nh} V(K)}_{\text{AIV}} + \underbrace{\frac{h^4}{4} \beta_2^2(K) V(m'')}_{\text{AISB}}, \tag{2.9}$$

kde $V(m'') = \int_0^1 \big(m''(x)\big)^2 \,\mathrm{d}x$ a AIV značí asymptotický tvar rozptylu (*asymptotic integrated variance*) a AISB asymptotický tvar druhé mocniny vychýlení (*asymptotic integrated square bias*).

Naším cílem je minimalizovat $\operatorname{AMISE}(h)$, tzn. najít takovou hodnotu vyhlazovacího parametru h, pro kterou AMISE nabývá minimální hodnoty, a tedy odhad bude nejlepší ve smyslu AMISE. Užijeme metody matematické analýzy a spočítáme derivaci

$$\frac{\mathrm{d}\operatorname{AMISE}(h)}{\mathrm{d}h} = -\frac{\sigma^2 V(K)}{nh^2} + h^3 \beta_2^2(K) V(m''),$$

položíme ji rovnu nule a vyjádříme h

$$h_{opt,0,2}^5 = \frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')}. \tag{2.10}$$

Poznámka 3.1. Tento výpočet vede k nalezení minima AMISE, protože platí

$$\frac{\mathrm{d}^2 \operatorname{AMISE}}{\mathrm{d}h^2} > 0.$$

Poznámka 3.2. Jestliže jádro K náleží do třídy S_{0k}, pak AMISE je tvaru

$$\operatorname{AMISE}(h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(m^{(k)}), \tag{2.11}$$

a pro optimální vyhlazovací parametr h platí

$$h_{opt,0,k}^{2k+1} = \frac{\sigma^2 V(K) (k!)^2}{2kn \beta_k^2(K) V(m^{(k)})}, \tag{2.12}$$

kde $V(m^{(k)}) = \int_0^1 \big(m^{(k)}(x)\big)^2 \,\mathrm{d}x$, podrobněji např. [7].

Nyní uvedeme důležité lemma, které ukazuje zajímavou vlastnost vyhlazovacího parametru.

Lemma 3.1. *Pro* $h_{opt,0,k}$ *platí*

$$\operatorname{AIV}(h_{opt,0,k}) = 2k \operatorname{AISB}(h_{opt,0,k}).$$

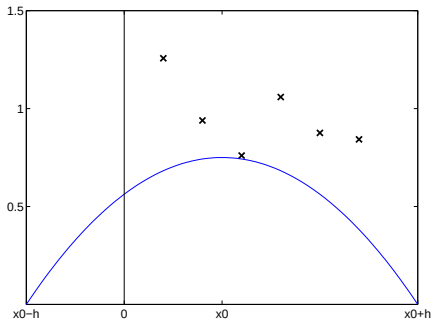
Důkaz. Viz cvičení. □

Vztah (2.12) pro optimální šířku vyhlazovacího okna ukazuje, že řád konvergence optimální šířky vyhlazovacího okna $h_{opt,0,2}$ závisí na řádu jádra k, tedy pro $k = 2$ je $O\big(n^{-1/5}\big)$. Rovnice (2.12) má pouze teoretický charakter, protože hodnota $h_{opt,0,2}$ závisí na neznámých veličinách σ^2 a $m''(x)$, a tedy není užitečná pro praktické účely. Na druhou stranu umožní vyjádřit asymptotickou rychlost konvergence $\mathrm{AMISE}(h)$. Dosadíme-li (2.10) do vztahu (2.9) pro AMISE, dostaneme

$$\mathrm{AMISE}(h_{opt,0,2}) = \frac{5}{4}\big(\sigma^2 V(K)\big)^{4/5}\big(\beta_2^2(K)V(m'')\big)^{1/5} n^{-4/5}. \tag{2.13}$$

Lze ukázat (viz cvičení), že pro jádra $K \in S_{0k}$ je $\mathrm{AMISE}(h) = O\left(n^{-\frac{2k}{2k+1}}\right)$. To znamená, že s rostoucím k se zvyšuje asymptotická rychlost konvergence. Ale není zcela jasné, zda tato zvýšená rychlost konvergence vede již k zlepšení pro konečné rozsahy výběrů, neboť ostatní veličiny se rovněž mění s k. O těchto problémech pojednáme podrobněji v dalším odstavci. Nevýhodou jader vyšších řádů je fakt, že pro tato jádra je optimální šířka okna větší, což může mít negativní dopad na hraniční efekty. Na druhé straně, chování jádrových odhadů s jádry vyšších řádů je méně citlivé na volbu šířky okna, není-li určena zcela optimálně, neboť křivka $\mathrm{AMISE}(h)$ je plošší.

Poznámka 3.3. Vyšetřování kvality odhadu obvykle probíhá za předpokladu, že pracujeme s vnitřními body intervalu $[0, 1]$. V hraničních oblastech, tj. v intervalech $[0, h) \cup (1 - h, 1]$, je kvalita odhadu ovlivněna negativně skutečností, že jádro K zde nesplňuje momentové podmínky (1.1). To je způsobeno tím, že blízko krajních bodů nosič jádra K zasahuje do oblasti, kde nejsou žádná data, což zhoršuje odhad – viz obr. 2.9. Hraniční efekty jsou také patrné na obrázcích 2.6 a 2.12,



Obrázek 2.9: Hraniční efekt

zejména u pravého okraje intervalu. Problém okrajových efektů lze překonat např. použitím hraničních jader (viz [9]) nebo reflexní metodou (viz [3]).

4 Volba jádra

Volba jádra není z asymptotického hlediska podstatná, jak je zřejmé z faktu (2.13). Je vhodné zvolit optimální jádro, které minimalizuje fukcionál $T(K)$, neboť tato jádra jsou spojitá na $\mathbb{R}$ a odhadovaná regresní funkce tak „zdědí" hladkost jádra. Vhodná jsou jádra třídy S_{02} a S_{04}, lze je vybrat z tabulek pro S_{0k}.

5 Volba vyhlazovacího parametru

5.1 Metoda křížového ověřování

Jednou z nejrozšířenějších a nejpoužívanějších metod pro určení optimální hodnoty parametru h je metoda křížového ověřování (*cross-validation method*). Tato metoda je založena na odhadu

regresní funkce (2.4), v němž vynecháme i-té pozorování:

$$\widehat{m}_{-i}(x_i, h) = \sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} W_\ell(x_i, h) Y_\ell, \quad i = 1, \ldots, n.$$

Funkce křížového ověřování je definována takto

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - Y_i\big)^2 \tag{2.14}$$

a odhadem optimální hodnoty vyhlazovacího parametru je bod, v němž nastává minimum této funkce, tj.

$$\hat{h}_{opt,0,k} = h_{\mathrm{CV}} = \arg\min_{h \in H_n} \mathrm{CV}(h).$$

Při praktických úlohách hledáme minimum na intervalu $H_n = [a_k n^{-1/(2k+1)}, b_k n^{-1/(2k+1)}]$, jehož tvar plyne ze vztahu (2.12), přičemž a_k, b_k jsou konstanty ($0 < a_k < b_k < \infty$), které ovšem neznáme. A proto pro ekvidistantní body plánu byl na základě zkušeností doporučen interval $[\frac{1}{n}, 2]$.

Poznámka 5.1. Někdy se místo chyby MISE používá průměrná střední kvadratická chyba AMSE (*average mean square error*)

$$\mathrm{AMSE}(h) = \frac{1}{n} E \sum_{i=1}^{n} \big(\widehat{m}(x_i, h) - m(x_i)\big)^2$$

Využívá se zejména v případech, kdy je nevyhovující použít numerické integrování související s chybou MISE.

Věta 5.1. *Pro střední hodnotu funkce* $\mathrm{CV}(h)$ *platí*

$$E\,\mathrm{CV}(h) = \underbrace{\frac{1}{n} E \sum_{i=1}^{n} \big(\widehat{m}(x_i, h) - m(x_i)\big)^2}_{\mathrm{AMSE}} + \sigma^2.$$

Důkaz. Funkci křížového ověřování lze rozepsat

$$\begin{aligned}
\mathrm{CV}(h) &= \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i) - (\underbrace{Y_i - m(x_i)}_{\varepsilon_i})\big)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)^2 - 2\frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)\varepsilon_i + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)^2 - \frac{2}{n} \sum_{i=1}^{n} \varepsilon_i \Big(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{n} W_\ell(x_i, h) Y_\ell - m(x_i)\Big) + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2.
\end{aligned}$$

Střední hodnota $E\,\mathrm{CV}(h)$ je rovna součtu tří hodnot. Předpokládejme, že $\widehat{m}_{-i}(x_i, h) \approx \widehat{m}(x_i, h)$, pak první ze sčítanců je roven přímo $\mathrm{AMSE}(h)$:

$$\frac{1}{n} E \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i, h) - m(x_i)\big)^2 = \mathrm{AMSE}(h).$$

Dále víme, že $Y_\ell = m(x_\ell) + \varepsilon_\ell$, tedy můžeme psát:

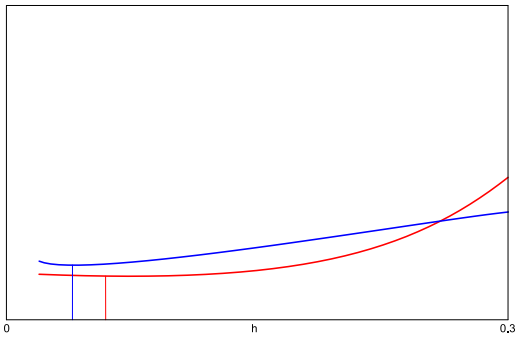
$$
\begin{aligned}
&-\frac{2}{n}E\sum_{i=1}^{n}\varepsilon_i\Big(\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} W_\ell(x_i,h)Y_\ell - m(x_i)\Big)\\
&= -\frac{2}{n}E\sum_{i=1}^{n}\varepsilon_i\Big(\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} W_\ell(x_i,h)(m(x_\ell+\varepsilon_\ell) - m(x_i)\Big)\\
&= -\frac{2}{n}E\left[\sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\varepsilon_i W_\ell(x_i,h)m(x_\ell) + \sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\varepsilon_i W_\ell(x_i,h)\varepsilon_\ell - \sum_{i=1}^{n}\varepsilon_i m(x_i)\right]\\
&= -\frac{2}{n}\left[\sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} W_\ell(x_i,h)m(x_\ell)\underbrace{E\varepsilon_i}_{=0} + \sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n} W_\ell(x_i,h)\underbrace{E\varepsilon_i\varepsilon_\ell}_{=0} - \sum_{i=1}^{n} m(x_i)\underbrace{E\varepsilon_i}_{=0}\right]\\
&= 0,
\end{aligned}
$$

Stejně jako pro druhý sčítanec, i pro třetí sčítanec využijeme vlastnosti (2.2):

$$\frac{1}{n}E\sum_{i=1}^{n}\varepsilon_i^2 = \sigma^2.$$

□

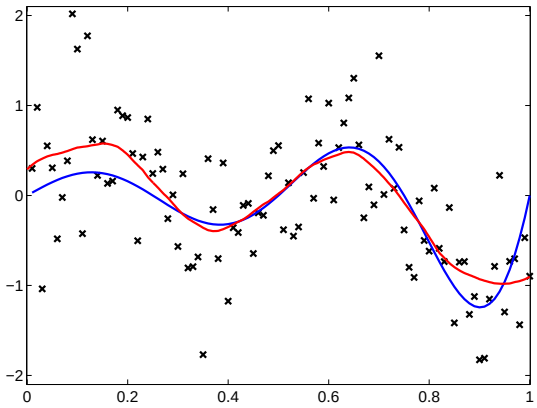
Tento výsledek znamená, že minimalizace $E\,\mathrm{CV}(h)$ odpovídá minimalizaci AMSE. Jestliže tedy předpokládáme, že minimum $\mathrm{CV}(h)$ je blízko minima $E\,\mathrm{CV}(h)$, pak tato minimalizace dává dobrou volbu vyhlazovacího parametru – viz ilustrace na obr. 2.10.



Obrázek 2.10: Porovnání minima AMSE (červenou) a minima funkce křížového ověřování CV (modrou) pro simulovaná data

Příklad 5.1. Použijeme metodu křížového ověřování pro nalezení vyhlazovacího parametru pro data z příkladu 1.1. Při použití Epanečnikova jádra získáme vyhlazovací parametr $h_{\mathrm{CV}} = 0{,}1158$. Na obrázku 2.11 je zobrazen odhad regresní funkce s tímto parametrem.

Kromě metody křížového ověřování se také pro odhad optimálního vyhlazovacího parametru používají metody založené na ASE (*average square error*), metody plug-in, metody odvozené z Fourierovy transformace a bootstrapové metody (podrobněji např. [3, 6]).



Obrázek 2.11: Simulovaná data (×) s jádrovým odhadem regresní funkce ($h_{\mathrm{CV}} = 0{,}1158$) (—) a původní funkcí (—)

6 Automatická procedura

Z dříve uvedených odhadů chyb je zřejmé, že kvalita jádrového odhadu závisí na šířce okna h, na jádře K a na řádu jádra k, což je číslo, které odpovídá předpokládanému počtu derivací v odhadovaném modelu. Je zřejmé, že všechny tyto tři veličiny se vzájemně ovlivňují, a proto je třeba zabývat se jejich volbou současně.

Pro simultánní volbu jádra, optimálního vyhlazovacího parametru a řádu jádra byla navržena automatická procedura (viz [3]), která odhadne všechny parametry tak, aby byla minimalizována AMISE. Procedura byla původně odvozena pro odhad hustoty pravděpodobnosti ([4]), ale lze ji aplikovat i pro odhad regresní funkce. Uvedeme zde její zjednodušenou verzi.

Podle poznámky 3.2 je známo, že AMISE a $h_{opt,0,k}$ jsou tvaru

$$\mathrm{AMISE}(h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{(k!)^2} h^{2k} \beta_k^2(K) V(m^{(k)}), \qquad h_{opt,0,k}^{2k+1} = \frac{\sigma^2 \delta_{0k}^{2k+1} (k!)^2}{2knV(m^{(k)})}.$$

Ze vztahu pro $h_{opt,0,k}$ vypočteme $V(m^{(k)})$ a dosadíme do vztahu pro AMISE, použijeme vztahy

$$\delta_{0k} = \left(\frac{V(K)}{\beta_k^2(K)}\right)^{\frac{1}{2k+1}} \qquad T(K) = \left(\beta_k(K) V^k(K)\right)^{\frac{2}{2k+1}} \qquad T(K) = \delta_{0k}^{2k} \beta_k^2(K)$$

dostaneme vyjádření AMISE

$$\mathrm{AMISE}(h_{opt,0,k}) = \frac{\sigma^2 (2k+1) \delta_{0k}}{2nkh_{opt,0,k}} T(K),$$

ve kterém jsou parametry K, h a k separovány, což umožňuje vybrat tyto parametry simultánně, Právě tento vztah je základem automatické procedury. Označme

$$L(k) = \frac{(2k+1)\delta_{0k}}{2nkh_{opt,0,k}} T(K).$$

a množinu vhodných řádů k označme

$$I(k_0) = \left\{2j, j = 0, \ldots, \left[\frac{k_0}{2}\right]\right\},$$

kde $[z]$ značí celou část čísla z. Procedura pak probíhá v pěti krocích:

Krok 1 Pro $k \in I(k_0)$ najděte optimální jádro $K_{opt,0,k} \in S_{0k}$, které je dáno tabulkou 1.2 a vypočtěte kanonický faktor δ_{0k}.

Krok 2 Pro $k \in I(k_0)$ a $K_{opt,0,k} \in S_{0k}$ najděte optimální vyhlazovací parametr $\hat{h}_{opt,0,k}$.

Krok 3 Pro $k \in I(k_0)$ vypočtěte hodnotu výběrového kriteria $L(k)$ s využitím hodnot získaných v krocích 1 a 2.

Krok 4 Vypočtěte optimální hodnotu řádu $\hat{k}$, které minimalizuje funkcionál $L(k)$.

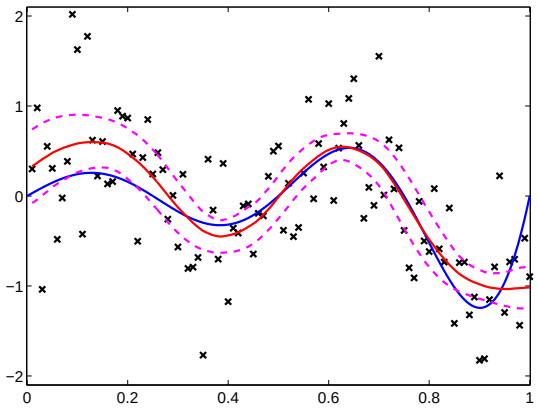
Krok 5 Použijte parametry z předchozích kroků k získání optimálního jádrového odhadu regresní funkce, tj.

$$\widehat{m}(x, \hat{h}_{opt,0,\hat{k}}) = \sum_{i=1}^{n} W_i(x, \hat{h}_{opt,0,\hat{k}}) Y_i.$$

Příklad 6.1. Aplikace procedury na data z příkladu 1.1. Maximální řád jádra zvolme $k_0 = 8$, tedy množina možných řádů jader je $I(8) = \{0, 2, 4, 6, 8\}$. Pro tyto řády spočítejme hodnoty z kroků 1–3, v kroku 2 jsme použili metodu křížového ověřování pro nalezení optimálního vyhlazovacího parametru $\hat{h}_{opt,0,k}$.

k	$K_{opt,0,k}$	δ_{0k}	h	$L(k)$
2	$-\frac{3}{4}(x^2-1)$	1,7188	0,1158	0,0648
4	$\frac{15}{32}(x^2-1)(7x^2-3)$	2,0165	0,2446	0,0575
6	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$	2,0834	0,3575	0,0574
8	$\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$	2,1021	0,4543	0,0592

Z tabulky vidíme, že optimální řád jádra je $\hat{k} = 6$. Výsledný odhad je uveden na obrázku 2.12.



Obrázek 2.12: Simulovaná data (×) s jádrovým odhadem regresní funkce při použití procedury (—) a skutečnou funkcí (—) společně s 95% intervalem spolehlivosti

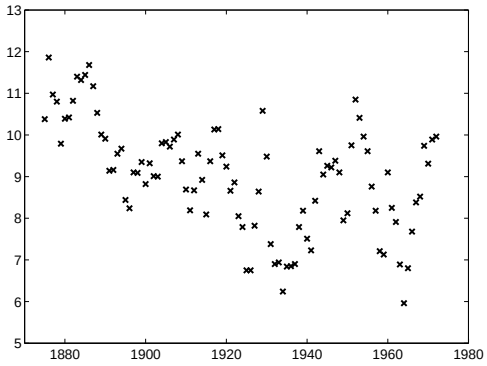
7 Aplikace na reálná data

První datový soubor obsahuje měření úrovně hladiny vody v Huronském jezeře. Měření byla prováděna ročně, v letech 1875 až 1972, tedy naměřené hodnoty jsou ekvidistantní. Data jsou shrnuta v tabulce 6.5 a na obrázku 2.13.

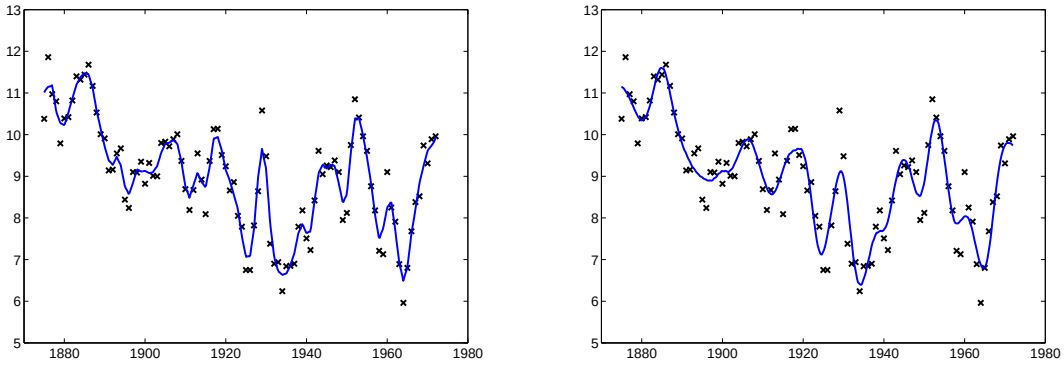
Vyhlazovací parametry jsme odhadli pomocí metody křížového ověřování (za použití Epanečnikova jádra) a také pomocí automatické procedury. Hodnoty vyhlazovacích parametrů jsou následující:

$$h_{\mathrm{CV}} = 0{,}0204 \quad (K_{opt,0,2}), \qquad h_{\mathrm{proc}} = 0{,}1525 \quad (K_{opt,0,12}).$$

Výsledné odhady na obrázku 2.14 ukazují, že metoda křížového ověřování spíše podhlazuje.



Obrázek 2.13: Úroveň hladiny Huronského jezera



(a) CV (b) procedura

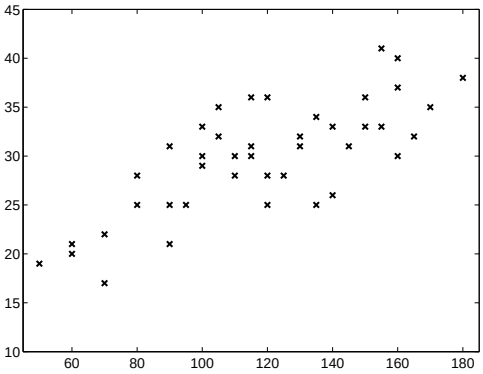
Obrázek 2.14: Odhadnuté regresní funkce – na ose x jsou roky a na ose y je hladina vody ve stopách (snížená o 570 stop – viz poznámka u dat)

Druhým datovým souborem jsou měření axiální délky krystalů ledu. Měření byla prováděna v místnosti s konstantní teplotou $-5\,^{\circ}\mathrm{C}$ v časových intervalech 50 až 180 vteřin po přinesení krystalu do místnosti. V tomto případě nejde o ekvidistantní data, protože hodnoty se liší o pět či deset vteřin. Data jsou v tabulce 6.6 a na obrázku 2.15.

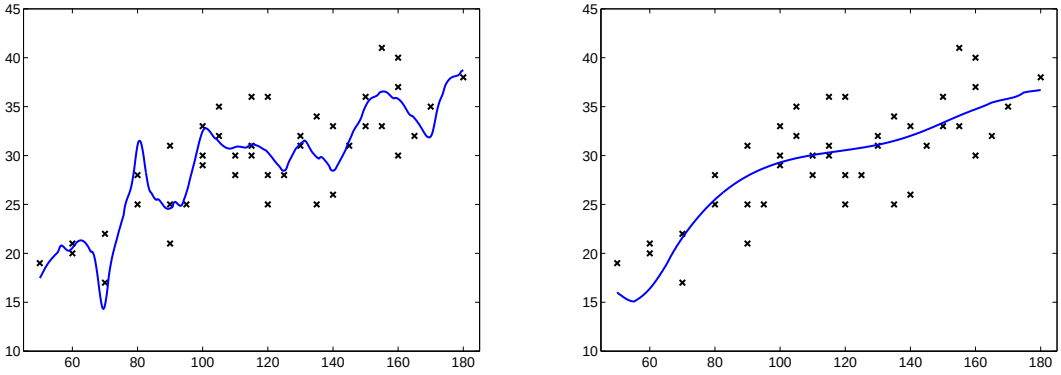
Odhady vyhlazovacích parametrů podle metody křížového ověřování a pomocí automatické procedury:

$$h_{\mathrm{CV}} = 0{,}1865 \quad (K_{opt,0,2}), \qquad h_{\mathrm{proc}} = 0{,}8826 \quad (K_{opt,0,10}).$$

Výsledné odhady regresní funkce jsou na obrázku 2.16. Je vidět, že CV metoda dává spíš podhlazený odhad. Na druhou stranu, odhad pomocí procedury se může zdát již přehlazený.



Obrázek 2.15: Axiální délka krystalů ledu



(a) CV (b) procedura

Obrázek 2.16: Odhadnuté regresní funkce – na ose x je vynesen čas ve vteřinách a na ose y délka krystalu v mikrometrech

Shrnutí

Odhad regresní funkce $m(x)$ v bodě x je tvaru

$$\widehat{m}(x,h) = \sum_{i=1}^{n} W_i(x,h) Y_i.$$

W_i závisí na K a h.

Vychýlení (bias) a rozptyl (var) odhadu s jádrem řádu 2 jsou

$$\operatorname{bias} \widehat{m}(x,h) \approx \frac{h^2}{2}\beta_2(K) m''(x), \qquad \operatorname{var} \widehat{m}(x,h) \approx \frac{\sigma^2}{nh} V(K).$$

Asymptotická střední kvadratická chyba jádrového odhadu regresní funkce s jádrem řádu 2 je

$$\operatorname{AMISE}(h) = \frac{\sigma^2}{nh} V(K) + \frac{1}{4} h^4 \beta_2^2(K) V(m'').$$

Optimální vyhlazovací parametr vzhledem k AMISE pro $k = 2$ je tvaru

$$h^5_{\operatorname{AMISE}} = \frac{\sigma^2 V(K)}{n \beta_2^2(K) V(m'')}$$

s řádem konvergence $n^{-1/5}$.

Metoda křížového ověřování pro odhad optimálního vyhlazovacího parametru

$$\operatorname{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{m}_{-i}(x_i,h) - Y_i\big)^2 \quad \Rightarrow \quad h_{\operatorname{CV}} = \arg\min \operatorname{CV}(h).$$

Automatická procedura pro simultánní volbu optimálního jádra, jeho řádu a vyhlazovacího parametru je dostupná v toolboxu Matlabu.

Dodatky a cvičení

1. Pro odhady $\widehat{m}_{NW}, \widehat{m}_{LL}, \widehat{m}_{GM}$ dokažte, že „množství“ vyhlazení pomocí jádra K s vyhlazovacím parametrem h je stejné, jako „množství“ vyhlazení jádrem K_δ s parametrem $h^* = h/\delta$, tj.
$$\widehat{m}(x,h,K) = \widehat{m}(x,h^*,K_\delta).$$

2. Dokažte vztah pro střední hodnotu odhadu (2.6).

3. Dokažte vztah (2.13) pro obecné k, tj. že platí
$$\operatorname{AMISE}(h_{opt,0,k}) = n^{-2k/(2k+1)} \Big(\sigma^2 V(K)\Big)^{2k/(2k+1)} \Big(\beta_k^2(K)(k!)^{-2} V(m^{(k)})\Big)^{1/(2k+1)} \frac{2k+1}{(2k)^{\frac{2k}{2k+1}}}.$$
Návod: Předpokládejme, že funkce m je spojitá a má spojité derivace až do řádu k včetně, tj. $m \in C^k[0,1]$. Užitím Taylorova (až do řádu k) rozvoje upravíme vztah pro střední hodnotu odhadu podobně jako ve vztahu (2.7). Užijte poznámky 3.2.

4. Dokažte vztah z lemmatu 3.1.

5. Dokažte, že příspěvek jádra $K_{\delta_{0k}}$, $\delta_{0k} = \left(\frac{V(K)}{\beta_k^2(K)}\right)^{1/(2k+1)}$, $K \in S_{0k}$, k oběma částem chyby AMISE je stejný, tj. platí rovnost

$$\int_{-\delta_{0k}}^{\delta_{0k}} K_{\delta_{0k}}^2(t)\,\mathrm{d}t = \left(\int_{-\delta_{0k}}^{\delta_{0k}} t^k K_{\delta_{0k}}(t)\,\mathrm{d}t\right)^2.$$

6. Aplikujte metodu křížového ověřování a automatickou proceduru na simulovaná i reálná data.

Dodatek

Výpočet integrálu, $t_i = \frac{i}{n}$, $i = 1, \ldots, n$,

$$\begin{aligned}
&\int_0^1 G(t)\,\mathrm{d}t \\
&= \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} G(t)\,\mathrm{d}t
\end{aligned}$$

s využitím Taylorova rozvoje funkce $G(t)$

$$\begin{aligned}
&= \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (G(t_{i+1}) + (t - t_{i+1})G'(t_{i+1}) + o(1))\,\mathrm{d}t \\
&= \sum_{i=0}^{n-1} G(t_{i+1})(\underbrace{t_{i+1} - t_i}_{=\frac{1}{n}}) + \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (t - t_{i+1})G'(t_{i+1})\,\mathrm{d}t + o(1) \\
&= \frac{1}{n}\sum_{i=1}^{n} G(t_i) + \sum_{i=0}^{n-1} \frac{(t_{i+1} - t_i)^2}{2} G'(t_{i+1}) + o(1) \\
&= \frac{1}{n}\sum_{i=1}^{n} G(t_i) + \frac{1}{2n^2}\sum_{i=0}^{n-1} G'(t_{i+1}) + o(1).
\end{aligned}$$

Za předpokladu $|G'(t_{i+1})| \leq M$, pak platí

$$\int_0^1 G(t)\,\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n} G(t_i) + O(n^{-1}).$$