MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

DISERTAČNÍ PRÁCE

Brno 2004 Jan Koláček

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

Jádrové odhady regresní funkce

DISERTAČNÍ PRÁCE

Brno 2004 Jan Koláček

Děkuji touto cestou prof. RNDr. Ivaně Horové, CSc. za poskytnutou odbornou pomoc při vypracování této práce.

Obsah

Seznam použitého značení

Číslo v závorce označuje stránku, kde je symbol poprvé použit nebo definován.

$\mathbb{N}$	množina všech přirozených čísel (7)
$\mathbb{N}_0$	množina všech přirozených čísel a nula (18)
$\mathbb{R}$	množina všech reálných čísel (14)
$\mathbb{C}^T$	T - rozměrný komplexní vektorový prostor (43)
$\boldsymbol{c}'$	transpozice vektoru $\boldsymbol{c}$ (11)
$m(x)$	regresní funkce (3)
$\widehat{m}(x;h)$	odhad regresní funkce (3)
$Lip[a,b]$	třída spojitých funkcí splňujících Lipchitzovu podmínku (7)
$K(x)$	jádro (7)
$S_{\nu\kappa}$	třída všech jader řádu (ν,κ) (7)
$V(K)$	$\int_{-1}^{1} K^2(u)du$ (21)
β_κ	$\int_{-1}^{1} u^\kappa K(u)du$ (1)
A_κ	$\int_{0}^{1} m^{(\kappa)}(u)du$ (25)
$W_i(x)$	váhová funkce (10)
MSE	střední kvadratická chyba (15)
$AMSE$	průměrná střední kvadratická chyba (25)
ASE	průměrná kvadratická chyba (25)
$RSS_T(h)$	residuální součet čtverců (25)
$\widehat{R}_T(h)$	odhad průměrné střední kvadratické chyby (27)
$\boldsymbol{x}^{\pm}$	diskrétní Fourierova transformace vektoru $\boldsymbol{x}$ (43)
$\boldsymbol{I}_Y$	periodogram vektoru $\boldsymbol{Y}$ (43)
$\circledast$	diskrétní cyklická konvoluce (43)
$\varphi(x) = O(\psi(x))$	značí $\limsup_{x\to\infty} \frac{\lvert\varphi(x)\rvert}{\psi(x)} < \infty$ (18)
$\varphi(x) = o(\psi(x))$	značí $\lim_{x\to\infty} \frac{\lvert\varphi(x)\rvert}{\psi(x)} = 0$ (20)
$\varphi(x) \approx \psi(x)$	značí $\lim_{x\to\infty} \frac{\varphi(x)}{\psi(x)} = c + o(1)$ pro libovolnou konstantu $c \neq 0$ (20)

Předmluva

V oblasti neparametrických metod odhadu regresní funkce představují metody jádrového vyhlazování jednu z nejúčinnějších vyhlazovacích technik. Tyto metody je možné snadno matematicky analyzovat a výsledky vyhlazení lze následně snadno interpretovat. Z výpočetního hlediska zaujímá velmi důležitou úlohu skutečnost, že jádrové odhady jsou za poměrně obecných předpokladů asymptoticky ekvivalentní jiným vyhlazovacím metodám, např. splajnům.

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 $[x_t, Y_t],\ t = 0, \dots T-1$, 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* a příslušný regresní vztah zapisujeme do tvaru

$$Y_t = m(x_t) + \varepsilon_t, \qquad t = 0, \dots, T-1,$$

kde m je neznámá regresní funkce a $\varepsilon_t,\ t = 0, \dots T-1$, jsou chyby měření. 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 následujícími způsoby:

1. *Parametrický přístup* – předpokládáme, že neznámá regresní křivka má jistý předepsaný tvar (známá je např. regresní přímka vyjadřující lineární závislost). Obecně tento přístup znamená, že neznámá funkce náleží do třídy funkcí popsaných množinou parametrů.

2. *Neparametrický přístup* – nepředpokládáme, že hledaná funkce má stanovený tvar. V tomto případě předpokládáme pouze jistou hladkost hledané funkce.

V první polovině dvacátého století byla věnována pozornost zejména parametrickým metodám. V posledních letech však zaznamenaly značný rozvoj metody neparametrické. Tento rozvoj souvisí s rostoucími požadavky na zpracování dat, ať již jde o rozsah souborů či rozmanitost těchto dat apod. Čistě parametrický přístup však nevyhovuje vždy dostatečně potřebám flexibility. Nebývalý rozmach výpočetní techniky vytvořil dobré předpoklady pro rozvoj již zmíněných neparametrických metod.

Úvod

Kvalita jádrových odhadů regresní funkce závisí na mnoha faktorech. Nejdůležitějším z nich je šířka vyhlazovacího okna, která řídí hladkost odhadu. Tento parametr nejvíce ovlivňuje výsledný odhad a jeho volba je zásadním problémem ve vyhlazovacích metodách. Hlavním cílem této disertační práce bylo vytvořit ucelený souhrn poznatků z teorie hledání optimální šířky okna.

V první kapitole jsou uvedeny základní pojmy a definice. Jsou zde formulovány předpoklady pro regresní model, definována jádra třídy $S_{0\kappa}$ a také základní typy jádrových odhadů, které se nejčastěji používají.

Ve druhé kapitole jsou nejprve odvozeny lokálně polynomiální odhady, které se používají především při odhadech regresní funkce. Dále je věnována pozornost jejich střední kvadratické chybě, která lokálně popisuje kvalitu jádrových odhadů. Situace je podrobně popsána pro Nadarayovy – Watsonovy estimátory, které jsou speciálním typem lokálně polynomiálních odhadů. V závěru kapitoly je demonstrován vliv vyhlazovacího parametru na kvalitu odhadu.

Třetí kapitola se již zabývá vlastním hledáním vyhlazovacího parametru. Globálním kritériem pro tuto volbu je průměrná střední kvadratická chyba, jejíž minimum je definováno jako teoretická optimální šířka okna h. Tato hodnota však závisí na neznámých parametrech, především na samotné regresní funkci, a proto má pouze teoretický význam. Dále je uveden souhrn dosud známých metod používaných pro odhad optimální šířky okna. U každé z nich je na simulovaných datech provedeno srovnání výsledků získaných těmito metodami s teoretickou hodnotou optimálního vyhlazovacího parametru. Z uvedených srovnání je patrné, že často dochází k nalezení menších hodnot než je teoretické optimum.

Ve čtvrté kapitole je stávající regresní model periodicky rozšířen na tzv. „cyklický model“. Nejprve jsou zkoumány vlastnosti tohoto modelu, kterých se využije v dalších úvahách. Dále jsou uvedeny základní pojmy z Fourierovy analýzy. V souvislosti s těmito poznatky lze nahlížet na jádrové odhady regresní funkce v cyklickém modelu jako na diskrétní cyklickou konvoluci váhového vektoru s vektorem pozorování. S využitím znalostí z Fourierovy teorie tak lze nalézt nové přístupy k problematice hledání optimální šířky okna. Jedním z nich je metoda Fourierovy transformace, která je popsána také v této kapitole. Základní myšlenka této metody je pak aplikována i pro odhady neznámých parametrů při konstrukci plug-in metody, jež je prezentována v závěru kapitoly. U obou metod opět nechybí simulační studie a srovnání získaných výsledků s teoretickou optimální šířkou okna.

V páté kapitole je provedeno vzájemné porovnání všech uvedených metod na simulovaných datech. Při simulacích bylo vygenerováno 200 řad se stejnou regresní funkcí a pro každou nalezeny odhady optimální šířky okna pomocí porovnávaných metod. Rozložení výsledků všech metod ilustrují histogramy. V tabulkách jsou uvedeny střední hodnoty a směrodatné odchylky všech získaných hodnot spolu s teoretickou hodnotou optimální šířky okna pro jednotlivá κ.

V poslední kapitole jsou porovnávány zmíněné postupy aplikací na reálných datech. Při odhadech optimální šířky okna bylo použito týchž šesti metod jako v předchozí kapitole při simulacích. V prvních dvou příkladech jsou použita data většího rozsahu, a přestože všechny srovnávané metody jsou asymptoticky ekvivalentní, výsledné odhady jsou rozdílné. V posledním příkladě jsou naopak analyzována data poměrně malého rozsahu. I zde jsou podle očekávání výsledky zcela rozdílné.

Záměry a cíle práce

Jádrové vyhlazování má široké využití v mnoha matematických odvětvích. Kromě regresní funkce se často také odhaduje její derivace, velmi rozšířené jsou též jádrové odhady hustoty. V této práci se budeme zabývat pouze odhady regresní funkce. Budeme navíc předpokládat, že body plánu x_t jsou pevné veličiny rovnoměrně rozložené na daném intervalu (bez újmy na obecnosti budeme uvažovat interval $[0, 1]$).
Kvalita jádrových odhadů regresní funkce závisí na těchto třech parametrech:

– na typu jádrového odhadu, který má význam váhové funkce. Základní poznatky týkající se této oblasti jsou uvedeny např. v [19].

– na řádu jádra, které použijeme ke konstrukci odhadu. Podrobněji je o volbě optimálního řádu pojednáno v [8].

– na šířce vyhlazovacího okna, která řídí hladkost odhadu.

Poslední ze jmenovaných faktorů nejvíce ovlivňuje výsledný odhad a jeho volba je zásadním problémem ve vyhlazovacích metodách. Tato problematika je popsána v mnoha publikacích, např. v [5], [6], [10], [18]. Konečný odhad regresní funkce je ovlivněn ještě dalšími skutečnostmi. Například na okrajích intervalu se mohou objevit tzv. hraniční efekty a odhadům v těchto bodech je třeba věnovat zvláštní pozornost. Tento problém je možno řešit např. použitím hraničních jader (viz [7]).

Hlavním cílem této disertační práce je vytvořit ucelený souhrn poznatků z teorie hledání optimální šířky okna. Jelikož se jedná o velmi rozsáhlou problematiku, je celá práce zaměřena pouze na model s ekvidistantními body plánu a na odhad globálního vyhlazovacího parametru.
Dílčí cíle této disertační práce jsou:

- Shrnout dosud známé teoretické výsledky v dané oblasti a popsat metody pro odhad optimální šířky okna.
- Zkoumat problémy, ke kterým dochází při hledání optimálního vyhlazovacího parametru a pokusit se zobecnit některé ze studovaných technik pro jádra obecného řádu κ.
- Hledat a rozvíjet nové přístupy k problematice odhadů optimální šířky okna.
- Porovnávat dosud známé i nově navrhované metody a testovat dosažené výsledky na simulovaných datech.
- Demonstrovat využití studovaných postupů na reálných datech a analyzovat kvalitu jádrových odhadů regresní funkce v závislosti na vyhlazovacích parametrech získaných zkoumanými metodami.

1 Základní pojmy a definice

Regresní model s pevným plánem

Uvažujme standardní regresní model s pevným plánem

$$Y_t = m(x_t) + \varepsilon_t, \qquad t = 0, \dots, T-1, \qquad T \in \mathbb{N}, \tag{1}$$

kde x_t, $t = 0, \dots T-1$, jsou uspořádané „pevné“ body měření a ε_t, $t = 0, \dots T-1$, jsou chyby měření, o nichž se předpokládá, že jsou nezávislé náhodné veličiny mající stejné rozdělení a splňující podmínky

$$E(\varepsilon_t) = 0, \qquad var(\varepsilon_t) = \sigma^2 > 0, \qquad t = 0, \dots, T-1.$$

Funkce m se nazývá *regresní funkce*, neboť $E(Y_t) = m(x_t)$, $t = 0, \dots T-1$. Odhad regresní funkce budeme značit $\widehat{m}$.

Pro jednoduchost budeme v dalším předpokládat, že body x_t jsou ekvidistantně rozloženy na intervalu $[0, 1]$, tj.

$$x_t = t/T, \quad t = 0, \dots, T-1.$$

Označme $Lip[a, b]$ třídu spojitých funkcí na intervalu $[a, b]$ splňujících nerovnost

$$|g(x) - g(y)| \le L|x - y|, \quad \forall x, y \in [a, b],\ L > 0,\ L \text{ je konstanta.}$$

Definice 1.1. Nechť ν, κ jsou celá nezáporná čísla taková, že platí $0 \le \nu \le \kappa - 2$, ν a κ mají stejnou paritu. Funkci $K \in Lip[-1, 1]$, nosič$(K) = [-1, 1]$, splňující podmínky

$$(i) \qquad K(-1) = K(1) = 0$$

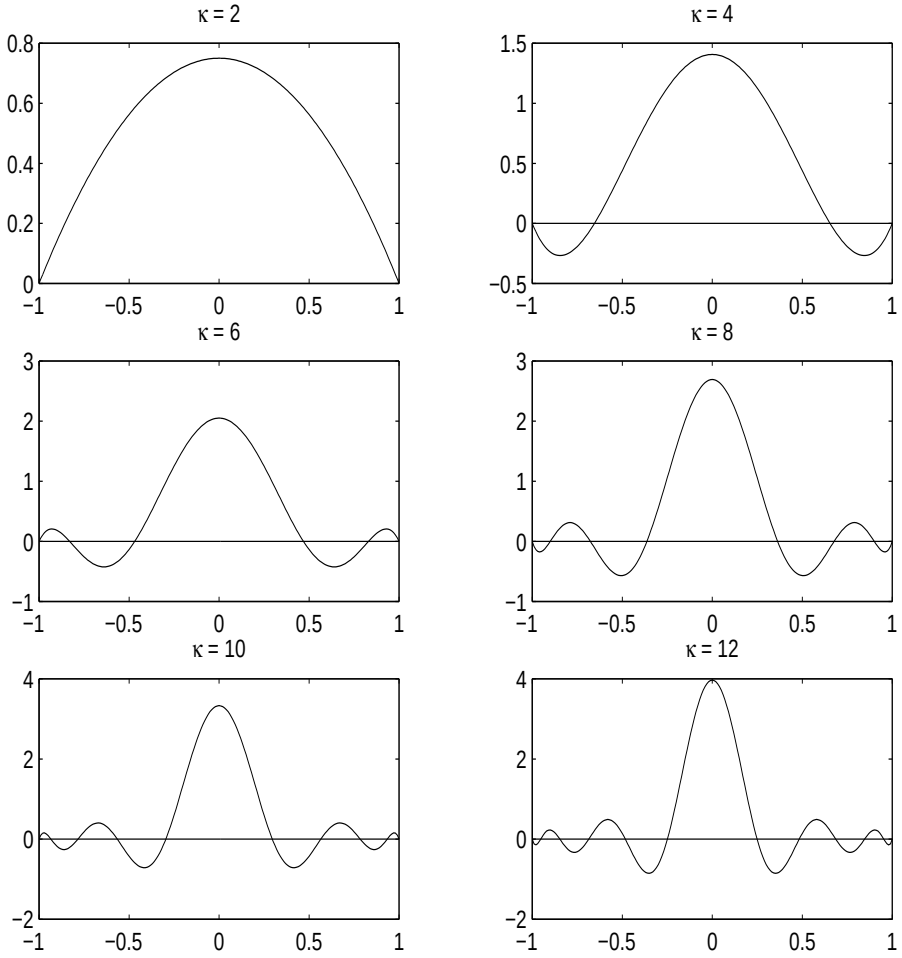
$$(ii) \qquad \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 \le j < \kappa, j \ne \nu \\ (-1)^\nu \nu!, & j = \nu \\ \beta_\kappa \ne 0, & j = \kappa, \end{cases}$$

nazýváme *jádrem* řádu (ν, κ) a třídu všech takových jader značíme $S_{\nu\kappa}$.

Poznámka. Takto definovaná jádra jsou řešením jisté optimalizační úlohy [19]. Pro odhad regresní funkce používáme jader třídy $S_{0\kappa}$ (viz obr.1), pro odhad ν-té derivace regresní funkce je vhodné volit jádra třídy $S_{\nu\kappa}$. Tabulka 1 udává explicitní tvar jádrových funkcí pro $\nu = 0$ a různé hodnoty κ na obr.1.

Tabulka 1: *Jádra třídy $S_{0\kappa}$, explicitní tvar*

κ	$K(x)$
2	$-\frac{3}{4}(x^2-1)$
4	$\frac{15}{32}(x^2-1)(7x^2-3)$
6	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$
8	$\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$
10	$-\frac{3465}{65536}(x^2-1)(4199x^8-7956x^6+4914x^4-1092x^2+63)$
12	$\frac{9009}{524288}(x^2-1)(52003x^{10}-124355x^8+106590x^6-39270x^4+5775x^2-231)$



Obrázek 1: *Jádra třídy $S_{0\kappa}$, $\kappa = 2, 4, 6, 8, 10, 12$.*

Označení. Nechť $K \in S_{0\kappa}, \kappa$ – sudé, označme

$$K_h(.) = \frac{1}{h} K(\frac{\cdot}{h}),$$

kde parametr $h > 0$ je tzv. „šířka vyhlazovacího okna“. Snadno je vidět, že nosičem jádra $K_h(.)$ je interval $[-h, h]$ a na něm také splňuje podmínky $(i), (ii)$.

Poznámka. Dále budeme uvažovat pouze jádra třídy $S_{0\kappa}, \kappa$ – sudé, a $h \in (0, 1)$. Pro vyhlazovací parametr $h \geq 1$ nemají smysl další úvahy, neboť body měření x_t jsou v intervalu $[0, 1]$.

Uveďme nyní nejznámější a nejčastěji používané typy jádrových odhadů regresní funkce. První dvě formule patří mezi tzv. „lokálně polynomiální“ odhady, kterým se budeme věnovat v odstavci 2.1. Poslední vzorec je konvolučním typem odhadu a je vhodný především k odhadům derivací regresní funkce.
Mezi nejznámější typy jádrových odhadů tedy patří:

1. **Nadarayovy – Watsonovy odhady** (1964)

$$\widehat{m}_{NW}(x; h) = \frac{\sum_{i=0}^{T-1} K_h(x_i - x) Y_i}{\sum_{i=0}^{T-1} K_h(x - x_i)}$$

2. **Lokální lineární estimátory** (Stone 1977, Cleveland 1979)

$$\widehat{m}_{LL}(x; h) = \frac{1}{T} \sum_{i=0}^{T-1} \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(x_i - x)\} K_h(x_i - x) Y_i}{\hat{s}_2(x; h)\hat{s}_0(x; h) - \hat{s}_1(x; h)^2},$$

   kde

$$\hat{s}_r(x; h) = \frac{1}{T} \sum_{i=0}^{T-1} (x_i - x)^r K_h(x_i - x)$$

3. **Pristleyho – Chaovy odhady** (1972)

$$\widehat{m}_{PCH}(x; h) = \frac{1}{T} \sum_{i=0}^{T-1} K_h(x_i - x) Y_i$$

4. **Gasserovy – Müllerovy odhady** (1979)

$$\widehat{m}_{GM}(x; h) = \sum_{i=0}^{T-1} Y_i \int_{s_{i-1}}^{s_i} K_h(t - x) dt,$$

   kde

$$s_i = \frac{x_i + x_{i+1}}{2},\ i = 0, \ldots, T - 2, \quad s_{-1} = 0, \quad s_{T-1} = 1.$$

Poznámka. V případě, že by v některém z prvních dvou odhadů nastalo dělení nulou, dodefinujeme příslušný odhad jako nulový.

Jádrové odhady tedy můžeme zapsat ve tvaru

$$\widehat{m}(x;h) = \sum_{i=0}^{T-1} W_i(x;h) Y_i, \tag{2}$$

kde váhy W_i odpovídají postupně odhadům $\widehat{m}_{NW}, \widehat{m}_{LL}, \widehat{m}_{PCH}, \widehat{m}_{GM}$. Můžeme tedy konstatovat, že jádrový odhad funkce m v bodě x je vážený průměr těch pozorování, pro která odpovídající body plánu leží v symetrickém okolí $[x-h, x+h]$ bodu x. Váhy $W_i(x;h)$ závisí na bodu x, na vyhlazovacím parametru h a na jádře K. Pro lepší přehlednost budeme v dalším tyto váhy zapisovat pouze $W_i(x)$. Vyhlazovací parametr řídí hladkost odhadu a jeho volba je zásadním problémem ve vyhlazovacích metodách.

2 Jádrové odhady

2.1 Lokálně polynomiální jádrové odhady

V tomto odstavci se budeme věnovat speciálnímu typu jádrových odhadů, který se nazývá *lokálně polynomiální*. Používá se především při odhadech regresní funkce. Hodnota neznámé regresní funkce v libovolném bodě plánu se získá tak, že proložíme dané body polynomem stupně p váženou metodou nejmenších čtverců. Ve speciálním případě pro $p = 0$, tj. prokládáním konstantou, obdržíme tzv. Nadarayovy – Watsonovy estimátory. Podobně, pro $p = 1$, prokládáme-li naměřená data přímkou, získáme lokálně lineární estimátory. Naším úkolem bude odvodit formální vzorec pro obecný stupeň polynomu.

Označme $\widehat{m}(x; p, h)$ odhad regresní funkce m v bodě x proložením polynomu stupně p váženou metodou nejmenších čtverců. Nechť tento polynom má tvar

$$P(u) = \beta_0 + \beta_1(u - x) + \cdots + \beta_p(u - x)^p.$$

Nechť K je nezáporné jádro na intervalu $[-1, 1]$, tj. $K(x) \geq 0,\ \forall x \in [-1, 1]$. Hodnotu $\widehat{m}(x; p, h)$ získáme váženou metodou nejmenších čtverců, tj. minimalizujeme funkcionál Φ v závislosti na vektoru parametrů $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$

$$\Phi = \sum_{i=0}^{T-1} \{Y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p\}^2 K_h(x_i - x).$$

Označme $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots \hat{\beta}_p)'$ vektor, pro který Φ nabývá minimální hodnoty. Odhad regresní funkce v bodě x získaný výše popsanou metodou je tedy hodnota parametru $\hat{\beta}_0$, tj.

$$\widehat{m}(x; p, h) = \hat{\beta}_0.$$

Naším cílem je najít explicitní vyjádření pro $\hat{\beta}_0$. To je popsáno v níže uvedené větě. Nejprve však předchází pomocné tvrzení, které bude potřeba při důkazu této věty.

Nechť $g(\boldsymbol{\beta})$ je skalární funkce vektoru $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. Derivací funkce g podle vektoru $\boldsymbol{\beta}$ rozumíme vektor

$$\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\boldsymbol{\beta})}{\partial \beta_0}, \ldots, \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_p} \right)'.$$

Lemma 2.1.1. *Nechť $\boldsymbol{c} = (c_0, \ldots, c_p)'$ je vektor délky $p+1$ a $\boldsymbol{A} = (a_{ij})_{i,j=0}^{p}$ je symetrická čtvercová matice řádu $p + 1$. Pak*

1. $\dfrac{\partial \boldsymbol{c}'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{c}$

2. $\dfrac{\partial \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{A}\boldsymbol{\beta}.$

Důkaz. Nejdřív dokážeme první vztah. Nechť $0 \leq k \leq p$, derivací výrazu $\boldsymbol{c}'\boldsymbol{\beta}$ podle β_k dostáváme k-tou složku vektoru $\frac{\partial \boldsymbol{c}'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$. Rozepsáním dostáváme $\boldsymbol{c}'\boldsymbol{\beta} = \sum_{i=0}^{p} c_i\beta_i$, derivujeme-li tento součet podle β_k, obdržíme právě c_k.
Druhý vztah se dokáže podobně. Nejprve podrobněji rozepíšeme $\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta}$

$$\begin{aligned}\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} &= \sum_{i=0}^{p}\sum_{j=0}^{p} a_{ij}\beta_i\beta_j \\ &= \sum_{\substack{i=0\\ i\neq k}}^{p}\sum_{\substack{j=0\\ j\neq k}}^{p} a_{ij}\beta_i\beta_j + \sum_{\substack{i=0\\ i\neq k}}^{p} a_{ik}\beta_i\beta_k + \sum_{\substack{j=0\\ j\neq k}}^{p} a_{kj}\beta_k\beta_j + a_{kk}\beta_k^2\end{aligned}$$

a následně zderivujeme podle β_k

$$\frac{\partial \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta}}{\partial \beta_k} = \sum_{\substack{i=0\\ i\neq k}}^{p} a_{ik}\beta_i + \sum_{\substack{j=0\\ j\neq k}}^{p} a_{kj}\beta_j + 2a_{kk}\beta_k = \sum_{\substack{i=0\\ i\neq k}}^{p}(a_{ik}+a_{ki})\beta_i + 2a_{kk}\beta_k.$$

Podle předpokladu je $\boldsymbol{A}$ symetrická matice, tj. $a_{ij} = a_{ji}$, pro všechna $0 \leq i, j \leq p$, a tedy

$$\frac{\partial \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta}}{\partial \beta_k} = 2\sum_{i=0}^{p} a_{ki}\beta_i,$$

což je k-tá složka vektoru $2\boldsymbol{A}\boldsymbol{\beta}$. □

V dalším budeme používat následující označení

$$\boldsymbol{Y} = \begin{pmatrix} Y_0 \\ \vdots \\ Y_{T-1} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_0 - x & \dots & (x_0-x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T-1} - x & \dots & (x_{T-1}-x)^p \end{pmatrix},$$

$$\boldsymbol{W} = \begin{pmatrix} K_h(x_0-x) & 0 & \dots & 0 \\ 0 & K_h(x_1-x) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & K_h(x_{T-1}-x) \end{pmatrix},$$

$$\boldsymbol{e}_1 = (1, 0 \dots 0)' \quad \text{vektor délky } p+1.$$

Věta 2.1.2. *Nechť je matice $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ regulární. Pro odhad regresní funkce platí*

$$\widehat{m}(x; p, h) = \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}. \tag{3}$$

Důkaz. Hodnotu $\widehat{m}(x;p,h)$ chceme získat váženou metodou nejmenších čtverců, tj. minimalizací funkce

$$\Phi = \sum_{i=0}^{T-1} \left\{ Y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p \right\}^2 K_h(x_i - x)$$

v závislosti na vektoru parametrů $\boldsymbol{\beta} = (\beta_0,\ \ldots,\ \beta_p)'$. Tuto funkci lze zapsat vektorově

$$\Phi = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Roznásobením dostáváme

$$\Phi = \boldsymbol{Y}'\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}.$$

Nyní zderivujeme Φ podle $\boldsymbol{\beta}$ a výraz položíme roven nule. Protože matice $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ je symetrická, můžeme při derivování použít předchozí lemma

$$\frac{\partial \Phi}{\partial \boldsymbol{\beta}} = \boldsymbol{0} - \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}.$$

Po jednoduché úpravě dostáváme tzv. *vážený systém normálních rovnic*

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}.$$

Podle předpokladu je matice $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ regulární, a tedy řešení existuje a je rovno

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}.$$

Nakonec stačí vyjádřit první složku $\hat{\beta}_0$, která je odhadem regresní funkce m, tj.

$$\widehat{m}(x;p,h) = \hat{\beta}_0 = \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}.$$

□

Příklad. Uvažujme speciální případ, kdy stupeň polynomu je $p = 0$, tj. naměřená data prokládáme lokálně konstantou. Jednotlivé matice jsou tvaru

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} K_h(x_0 - x) & 0 & \ldots & 0 \\ 0 & K_h(x_1 - x) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & K_h(x_{T-1} - x) \end{pmatrix}, \quad \boldsymbol{e}_1 = 1.$$

Spočítáme jejich součin $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} = \sum_{i=0}^{T-1} K_h(x_i - x).$$

Za předpokladu, že tento součet je nenulový, můžeme vypočítat inverzi

$$(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} = \frac{1}{\sum_{i=0}^{T-1} K_h(x_i - x)}.$$

Nakonec vyjádříme součin $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} = \sum_{i=0}^{T-1} K_h(x_i - x) Y_i.$$

Dosazením do (3) dostáváme

$$\widehat{m}(x; 0, h) = \frac{\sum_{i=0}^{T-1} K_h(x_i - x) Y_i}{\sum_{i=0}^{T-1} K_h(x - x_i)}.$$

Takto sestrojené odhady regresní funkce se nazývají *Nadarayovy – Watsonovy odhady.*

Poznámka. V tomto odstavci jsme užili předpokladu, že jádro K je nezáporné. I když jádra vyšších řádů tuto podmínku nesplňují, používají se ke konstrukci lokálně polynomiálních odhadů pro jejich dobré asymptotické vlastnosti (viz [19]).

2.2 Statistické vlastnosti odhadů

Kvalitu jádrového odhadu lze lokálně popsat pomocí střední kvadratické chyby. Budeme se zabývat asymptotickým tvarem této chyby, neboť pro neparametrické odhady, narozdíl od odhadů parametrických, neexistuje nevychýlený odhad, tj. takový odhad, že $E\widehat{m} = m$ pro s.v. $x \in \mathbb{R}$ (Collomb 1976).

Věta 2.2.1. *Nechť $K \in S_{0\kappa}, E\boldsymbol{Y}^2 < \infty$ a nechť posloupnost vyhlazovacích parametrů $h = h_T,\ T = 1, 2, \ldots,$ splňuje podmínky: $h_T \to 0,\ Th_T \to \infty$ pro $T \to \infty$. Pak v každém bodě spojitosti funkce m platí*

$$\sum_{k=0}^{T-1} W_k(x) Y_k \xrightarrow{p} m(x),$$

kde W_k jsou váhové funkce odpovídající postupně odhadům $\widehat{m}_{PCH},\ \widehat{m}_{NW},\ \widehat{m}_{LL},\ \widehat{m}_{GM}$.

Důkaz. Důkaz můžeme najít např. v [6]. □

Poznámka. Uvedené odhady jsou tedy *konzistentními* odhady m.

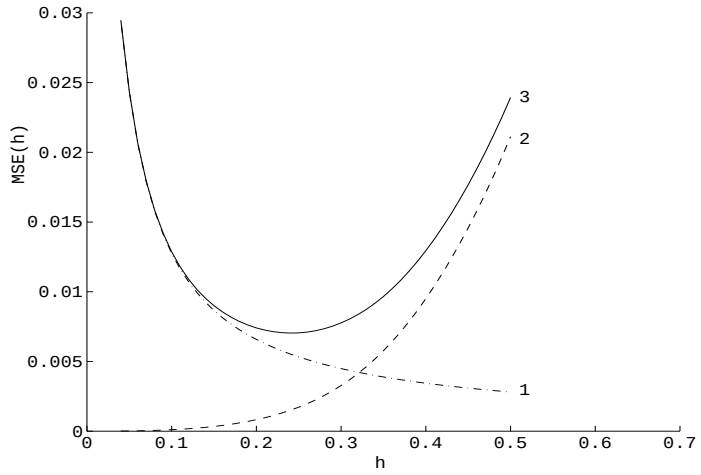
Střední kvadratická chyba MSE odhadu $\widehat{m}$ v bodě x je obecně dána vztahem

$$MSE(\widehat{m}(x;h)) = E(\widehat{m}(x;h) - m(x))^2. \tag{4}$$

Tento vztah lze dále zapsat ve tvaru

$$MSE(\widehat{m}(x;h)) = var\ \widehat{m}(x;h) + (E\widehat{m}(x;h) - m(x))^2 \tag{5}$$

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



Obrázek 2: *Střední kvadratická chyba* $MSE(h)$ *(křivka 3) jako součet rozptylu (křivka 1) a vychýlení*^{2} *(křivka 2) pro regresní funkci* $m(x) = \sin(2\pi x),\ \sigma^2 = 0.15$.

Budeme se zabývat chováním asymptotické střední kvadratické chyby ve vnitřních bodech intervalu $[0, 1]$. Jelikož v tomto případě jsou výše uvedené odhady asymptoticky ekvivalentní, budeme většinou vynechávat indexy PCH, NW, LL, GM označující jednotlivé odhady a odhad označíme pouze $\widehat{m}$. Situaci podrobně probereme pro lokálně polynomiální odhady, případ $p = 0$, tj. pro Nadarayovy – Watsonovy odhady. V blízkosti hraničních bodů se mohou objevit tzv. „hraniční efekty“ a odhadům v těchto bodech je třeba věnovat zvláštní pozornost.

Formulujme nyní předpoklady, za kterých budeme zkoumat střední kvadratickou chybu

1. Šířka okna $h = h_T$ splňuje podmínky

$$\lim_{T\to\infty} h_T = 0, \quad \lim_{T\to\infty} Th_T = \infty.$$

2. Bod x je vnitřním bodem intervalu $[0,1]$, tj. existuje T_0 tak, že

$$h < x < 1-h, \qquad \forall T \geq T_0.$$

3. Body plánu jsou ekvidistantní, tj.

$$x_t = \frac{t}{T}, \qquad t = 0, \ \ldots, T-1.$$

4. Jádro $K \in S_{0\kappa}, \ \kappa$ – sudé a první derivace $K^{'}$ je omezená.

5. Nechť $m \in C^{\kappa_0}[0,1], \ \kappa_0 > \kappa$.

Lokálně polynomiální odhady pro $p = 0$, známé jako Nadarayovy – Watsonovy odhady, jsou podle věty 2.1.2 tvaru

$$\widehat{m}(x;0,h) = \boldsymbol{e}_1'(\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{Y},$$

kde

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} K_h(x_0-x) & 0 & \ldots & 0 \\ 0 & K_h(x_1-x) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & K_h(x_{T-1}-x) \end{pmatrix}, \quad \boldsymbol{e}_1 = 1.$$

Střední kvadratická chyba pro tyto odhady je dána vztahem (5)

$$MSE(\widehat{m}(x;0,h)) = var\ \widehat{m}(x;0,h) + (E\widehat{m}(x;0,h) - m(x))^2,$$

kde první člen je *rozptyl* odhadu a druhý (*vychýlení*)^2. Nejprve se zaměříme na **vychýlení** Nadarayových – Watsonových odhadů.
Označme $\boldsymbol{M} = (m(x_0), \ \ldots, m(x_{T-1}))'$. Vyjádříme střední hodnotu uvažovaného odhadu

$$E\widehat{m}(x;0,h) = \boldsymbol{e}_1'(\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{W}\boldsymbol{M}.$$

Hodnoty vektoru $\boldsymbol{M}$ nahradíme Taylorovým rozvojem řádu κ funkce m se středem v bodě x

$$\boldsymbol{M} = m(x)\boldsymbol{X} + m^{'}(x)\begin{pmatrix} x_0-x \\ \vdots \\ x_{T-1}-x \end{pmatrix} + \frac{1}{2!}m^{''}(x)\begin{pmatrix} (x_0-x)^2 \\ \vdots \\ (x_{T-1}-x)^2 \end{pmatrix} + \ldots$$
$$+ \frac{1}{\kappa!}m^{(\kappa)}(x)\begin{pmatrix} (x_0-x)^\kappa \\ \vdots \\ (x_{T-1}-x)^\kappa \end{pmatrix} + R_\kappa(x),$$

kde

$$R_\kappa(x) = \frac{1}{(\kappa+1)!} m^{(\kappa+1)}(\xi) \begin{pmatrix} (x_0 - x)^{\kappa+1} \\ \vdots \\ (x_{T-1} - x)^{\kappa+1} \end{pmatrix}, \qquad \xi \in (0,1)$$

je chyba této aproximace. Po dosazení dostáváme

$$\begin{aligned} E\widehat{m}(x;0,h) = {} & \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} m(x) + \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m'(x) \begin{pmatrix} x_0 - x \\ \vdots \\ x_{T-1} - x \end{pmatrix} \\ & + \frac{1}{2!}\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m''(x) \begin{pmatrix} (x_0 - x)^2 \\ \vdots \\ (x_{T-1} - x)^2 \end{pmatrix} + \dots \\ & + \frac{1}{\kappa!}\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m^{(\kappa)}(x) \begin{pmatrix} (x_0 - x)^\kappa \\ \vdots \\ (x_{T-1} - x)^\kappa \end{pmatrix} \\ & + \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} R_\kappa(x). \end{aligned}$$

Odtud můžeme vyjádřit vychýlení

$$\begin{aligned} E\widehat{m}(x;0,h) - m(x) = {} & \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m'(x) \begin{pmatrix} x_0 - x \\ \vdots \\ x_{T-1} - x \end{pmatrix} \\ & + \frac{1}{2!}\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m''(x) \begin{pmatrix} (x_0 - x)^2 \\ \vdots \\ (x_{T-1} - x)^2 \end{pmatrix} + \dots \\ & + \frac{1}{\kappa!}\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} m^{(\kappa)}(x) \begin{pmatrix} (x_0 - x)^\kappa \\ \vdots \\ (x_{T-1} - x)^\kappa \end{pmatrix} \\ & + \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} R_\kappa(x). \end{aligned}$$

Užitím výpočtů z příkladu v odstavci 2.1 dostáváme

$$\begin{aligned} E\widehat{m}(x;0,h) - m(x) = {} & \frac{\hat{s}_1(x;h)}{\hat{s}_0(x;h)} m'(x) + \frac{1}{2!}\frac{\hat{s}_2(x;h)}{\hat{s}_0(x;h)} m''(x) + \dots \\ & + \frac{1}{\kappa!}\frac{\hat{s}_\kappa(x;h)}{\hat{s}_0(x;h)} m^{(\kappa)}(x) + \frac{1}{(\kappa+1)!}\frac{\hat{s}_{\kappa+1}(x;h)}{\hat{s}_0(x;h)} m^{(\kappa+1)}(\xi), \end{aligned} \tag{6}$$

kde

$$\hat{s}_r(x;h) = \frac{1}{T}\sum_{i=0}^{T-1}(x_i - x)^r K_h(x_i - x), \quad r \in \mathbb{N}_0.$$

Lemma 2.2.2. *Pro všechna $r \in \mathbb{N}_0$ platí*

$$\hat{s}_r(x;h) = h^r \int_{-1}^{1} u^r K(u)du + O(T^{-1}).$$

Důkaz. Položme

$$\varphi(x) = \int_0^1 (y-x)^r K_h(y-x)dy = \sum_{i=0}^{T-1}\int_{x_i}^{x_{i+1}} (y-x)^r K_h(y-x)dy.$$

Funkci $F(y) = (y-x)^r K_h(y-x)$ aproximujeme na intervalu $[x_i, x_{i+1}]$ konstantou – funkční hodnotou v bodě x_i a hodnoty příslušných integrálů aproximujeme obsahy vzniklých obdélníků. Dostáváme tedy

$$\varphi(x) = \sum_{i=0}^{T-1}\frac{1}{T}(x_i - x)^r K_h(x_i - x) + R(F),$$

kde $R(F)$ je chyba aproximace. Analyzujme nyní podrobněji tuto chybu. Chceme-li nahradit funkci $F(y)$ na intervalu $[x_i, x_{i+1}]$ funkční hodnotou $F(x_i)$, jedná se v podstatě o aproximaci Taylorovým rozvojem nultého řádu se středem v bodě x_i. Označme tento rozvoj $F_i(y)$, pak podle Taylorovy věty

$$F_i(y) = (x_i - x)^r K_h(x_i - x) + R(F_i),$$

kde $R(F_i) = F'(\xi_i)(y - x_i),\ \xi_i \in [x_i, y]$ je chyba tohoto rozvoje. Je tedy jasné, že

$$R(F) = \sum_{i=0}^{T-1}\int_{x_i}^{x_{i+1}} R(F_i)dy = \sum_{i=0}^{T-1}\int_{x_i}^{x_{i+1}} F'(\xi_i)(y - x_i)dy.$$

Zderivujeme-li funkci $F(y)$, dostáváme

$$F'(y) = r(y-x)^{r-1}K_h(y-x) + (y-x)^r\frac{\partial}{\partial y}K_h(y-x).$$

Podle předpokladu 4 je $\frac{\partial}{\partial y}K_h(y-x)$ omezená, a tedy i $F'(y)$ je omezená, tj.

$$\exists L > 0, \quad |F'(y)| \le L \quad \forall y \in [0,1].$$

Můžeme tedy odhadnout velikost chyby

$$|R(F)| \le \sum_{i=0}^{T-1}|F'(\xi_i)|\left|\int_{x_i}^{x_{i+1}}(y - x_i)dy\right| \le L\sum_{i=0}^{T-1}\frac{1}{2T^2} = \frac{L}{2T} = O(T^{-1}).$$

Celkem získáváme rovnost

$$\varphi(x) = \int_0^1 (y-x)^r K_h(y-x)dy = \hat{s}_r(x;h) + O(T^{-1}).$$

Odtud vyjádříme $\hat{s}_r(x;h)$ a upravíme integrál použitím substituce $u = \frac{y-x}{h}$

$$\begin{aligned}\hat{s}_r(x;h) &= \int_0^1 (y-x)^r K_h(y-x)dy + O(T^{-1}) \\ &= h^r \int_{\frac{-x}{h}}^{\frac{1-x}{h}} u^r K(u)du + O(T^{-1}).\end{aligned}$$

Podle předpokladu 2 je $h < x < 1-h$, odtud můžeme vyjádřit nerovnosti pro meze $\frac{-x}{h} < -1$ a $1 < \frac{1-x}{h}$. Podle definice je $K(u)$ funkce nulová vně intervalu $[-1,1]$, a proto můžeme uvažovat integrál pouze na tomto intervalu

$$\hat{s}_r(x;h) = h^r \int_{-1}^{1} u^r K(u)du + O(T^{-1}).$$

□

Poznámka. Protože uvažujeme jádra třídy $S_{0\kappa}$, můžeme aplikovat jejich vlastnosti na výsledek předchozího lemmatu. Obdržíme vyjádření pro $\hat{s}_r(x;h)$, $r \in \mathbb{N}_0$, která využijeme při hledání asymptotického tvaru vychýlení

$$\hat{s}_r(x;h) = \begin{cases} 1 + O(T^{-1}), & r = 0 \\ O(T^{-1}), & 0 < r < \kappa \\ h^\kappa \beta_\kappa + O(T^{-1}), & r = \kappa \\ O(h^r) + O(T^{-1}), & r > \kappa. \end{cases}$$

Nyní použijeme těchto výsledků a po dosazení do vztahu (6) dostáváme

$$\begin{aligned}E\widehat{m}(x;0,h) - m(x) &= \frac{O(T^{-1})}{1+O(T^{-1})} m'(x) + \frac{1}{2!}\frac{O(T^{-1})}{1+O(T^{-1})} m''(x) + \dots \\ &\quad + \frac{1}{\kappa!}\frac{h^\kappa \beta_\kappa + O(T^{-1})}{1+O(T^{-1})} m^{(\kappa)}(x) \\ &\quad + \frac{1}{(\kappa+1)!}\frac{O(h^{\kappa+1}) + O(T^{-1})}{1+O(T^{-1})} m^{(\kappa+1)}(\xi) \\ &= O(T^{-1}) m'(x) + \frac{1}{2!} O(T^{-1}) m''(x) + \dots \\ &\quad + \frac{1}{\kappa!}(h^\kappa \beta_\kappa + O(T^{-1}) + o(h^\kappa)) m^{(\kappa)}(x) \\ &\quad + \frac{1}{(\kappa+1)!}(o(h^\kappa) + O(T^{-1})) m^{(\kappa+1)}(\xi).\end{aligned}$$

Odtud jednoduchými úpravami dostáváme

$$E\widehat{m}(x;0,h) - m(x) = \frac{h^\kappa}{\kappa!}\beta_\kappa m^{(\kappa)}(x) + O(T^{-1}) + o(h^\kappa).$$

Využijeme-li asymptotických vlastností funkcí $O(T^{-1})$ a $o(h^\kappa)$, tj.

$$O(T^{-1}) \to 0, \text{ pro } T \to \infty$$

$$o(h^\kappa) \to 0, \text{ pro } T \to \infty,$$

získáme hlavní člen pro vychýlení Nadarayových – Watsonových odhadů

$$E\widehat{m}(x;0,h) - m(x) \approx \frac{h^\kappa}{\kappa!}\beta_\kappa m^{(\kappa)}(x). \tag{7}$$

V dalším se budeme zabývat prvním členem formule (5), tj. **rozptylem** $var\ \widehat{m}(x;0,h)$. Připomeňme si jeho obecný tvar

$$var\ \widehat{m}(x;0,h) = E(\widehat{m}(x;0,h) - E\widehat{m}(x;0,h))^2.$$

Podobně jako v případě vychýlení zapíšeme tento vztah vektorově. Označme $\boldsymbol{\varepsilon} = (\varepsilon_0,\ \ldots, \varepsilon_{T-1})'$, pak

$$\widehat{m}(x;0,h) - E\widehat{m}(x;0,h) = \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{M}) = \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{\varepsilon},$$

umocníme

$$(\widehat{m}(x;0,h) - E\widehat{m}(x;0,h))^2 = \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{e}_1$$

a vyjádříme střední hodnotu

$$\begin{aligned} E(\widehat{m}(x;0,h) - E\widehat{m}(x;0,h))^2 &= \boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{e}_1 \\ &= \sigma^2\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{e}_1. \end{aligned}$$

Odtud je jasné, že

$$\boldsymbol{X}'\boldsymbol{W}^2\boldsymbol{X} = \sum_{i=0}^{T-1} K_h^2(x_i - x),$$

a tedy

$$var\ \widehat{m}(x;0,h) = \frac{\frac{\sigma^2}{T^2}\sum\limits_{i=0}^{T-1} K_h^2(x_i - x)}{\hat{s}_0^2(x;h)}.$$

Lemma 2.2.3. *Platí následující vztah*

$$\frac{1}{T}\sum_{i=0}^{T-1} K_h^2(x_i - x) = \frac{1}{h}\int_{-1}^{1} K^2(u)du + O(T^{-1}).$$

Důkaz. Tvrzení by se dokázalo podobným způsobem jako v předchozím lemmatu. □

Označme

$$V(K) = \int_{-1}^{1} K^2(u)du,$$

pak analogickým postupem jako u vychýlení lze dojít k formuli pro rozptyl

$$var\ \widehat{m}(x;0,h) = \frac{\frac{\sigma^2}{T}(\frac{1}{h}V(K) + O(T^{-1}))}{(1+O(T^{-1}))^2} = \frac{\sigma^2 V(K)}{Th} + o(T^{-1}h^{-1}).$$

Limitním přechodem pro $T \to \infty$ obdržíme hlavní člen rozptylu

$$var\ \widehat{m}(x;0,h) \approx \frac{\sigma^2}{Th}V(K). \tag{8}$$

Na závěr lze použít vzorců (7) a (8) pro vychýlení a rozptyl a formulovat asymptotický tvar střední kvadratické chyby v bodě x pro Nadarayovy – Watsonovy odhady

$$MSE(\widehat{m}(x;0,h)) \approx \frac{\sigma^2}{Th}V(K) + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2(m^{(\kappa)}(x))^2. \tag{9}$$

Pro ostatní uvažované typy odhadů bychom obdrželi stejný výsledek, neboť všechny jsou asymptoticky ekvivalentní.

2.3 Vliv vyhlazovacího parametru na kvalitu odhadu

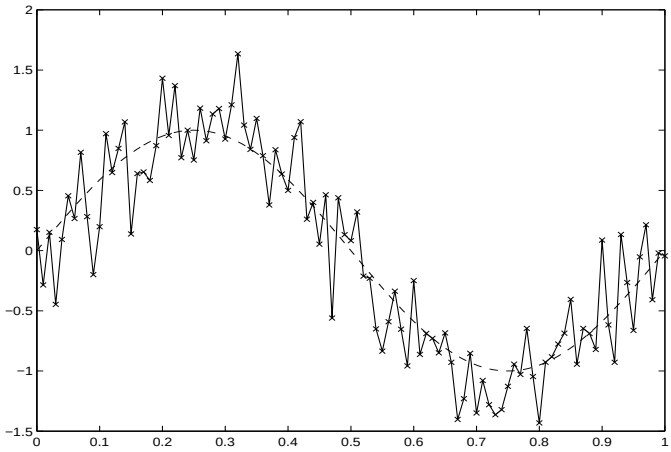
V mnoha aplikacích je užitečný zejména $\widehat{m}_{NW}$ odhad. Tohoto odhadu použijeme nyní k ilustraci vlivu vyhlazovacího parametru na kvalitu odhadu. V tomto případě jsou váhové funkce tvaru

$$W_t(x) = \frac{K_h(x-x_t)}{\sum_{i=0}^{T-1} K_h(x-x_i)}, \qquad t = 0 \dots T-1.$$

Jádrový odhad není definován pro $\sum_{i=0}^{T-1} K_h(x-x_i) = 0$. Jestliže nastane případ "0/0", pak klademe $\widehat{m}_{NW}(x;h) = 0$. Omezíme se nyní na odhady m v bodech plánu x_t, $t = 0, \dots T-1$.
Pro $h \to 0$ platí

$$\widehat{m}_{NW}(x_t;h) \to \frac{K(0)Y_t}{K(0)} = Y_t.$$

To znamená, že při malé šířce vyhlazovacího okna odhad reprodukuje data (viz obr.3).

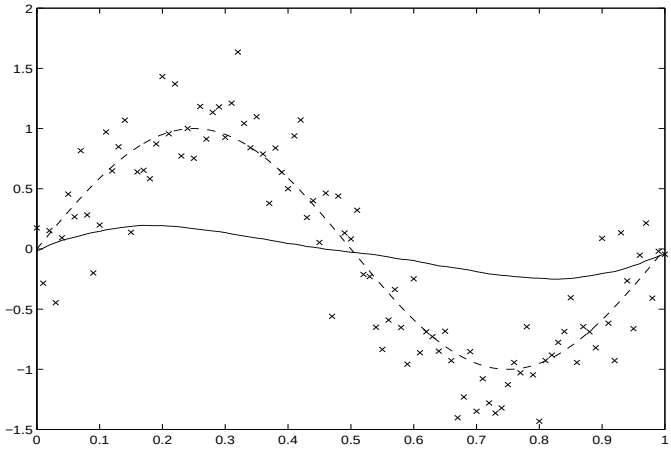


Obrázek 3: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.15$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = \sin(2\pi x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s parametrem* $h = 0.005$.

Pro $h \to \infty$ platí

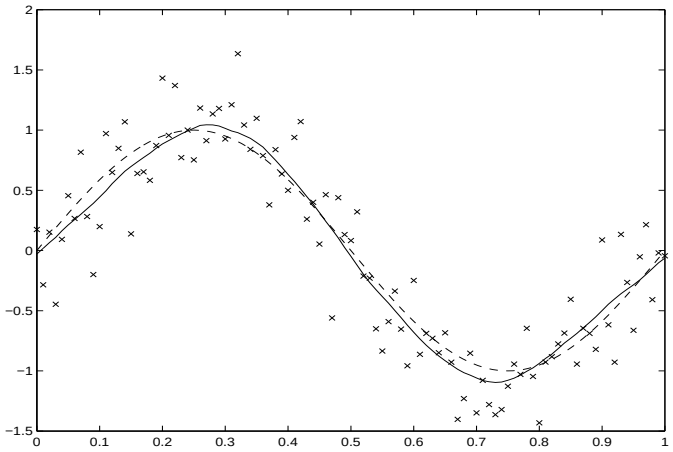
$$\widehat{m}_{NW}(x_t; h) \to \frac{\sum_{j=0}^{T-1} K(0) Y_j}{\sum_{j=0}^{T-1} K(0)} = \frac{K(0) \sum_{j=0}^{T-1} Y_j}{TK(0)} = \frac{1}{T} \sum_{j=0}^{T-1} Y_j.$$

Tedy velká šířka okna vede k přehlazení a to k průměru dat (viz obr.4). Na obrázku 5



Obrázek 4: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.15$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = \sin(2\pi x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s parametrem* $h = 0.6$.

je znázorněn odhad s optimální hodnotou h. 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, což ponechává prostor pro subjektivní posouzení.



Obrázek 5: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.15$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = \sin(2\pi x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s optimální šířkou okna* $h = 0.11$.

3 Volba šířky okna

3.1 Teoretické odhady vyhlazovacího parametru

Jak bylo uvedeno v minulé kapitole, hodnota vyhlazovacího parametru h, který nazýváme *šířka okna*, značně ovlivňuje výsledný odhad regresní funkce. Budeme se tedy zabývat hledáním optimální šířky okna, při níž bude jádrový odhad nejlepší. Kvalitu tohoto odhadu v bodě $x \in [0,1]$ teoreticky popisuje tzv. *střední kvadratická chyba.* Její asymptotický tvar (9) byl podrobněji popsán pro Nadarayovy – Watsonovy odhady v odstavci 2.2. Zaměříme se na odhady v bodech plánu. Odhad regresní funkce m na celém intervalu $[0,1]$ budeme charakterizovat pomocí globální chyby, tzv. *průměrné střední kvadratické chyby* (13). Optimální šířka okna je definována vztahem (14) jako minimum této funkce. Jedná se však pouze o teoretickou hodnotu, která závisí na neznámých parametrech. V praxi se postupuje tak, že minimalizujeme nějaký odhad průměrné střední kvadratické chyby. V této kapitole uvedeme několik klasických metod, které se používají k hledání optimální šířky okna. Všechny tyto metody jsou asymptoticky ekvivalentní a vycházejí z tzv. *residuálního součtu čtverců.*

Kvalitu jádrových odhadů popisuje střední kvadratická chyba MSE (4). Asymptotický tvar této chyby v bodech $x \in [0,1]$ můžeme vyjádřit vztahem

$$MSE(\widehat{m}(x;h)) = \overline{MSE}(\widehat{m}(x;h)) + o(1) \tag{10}$$

kde $\overline{MSE}(\widehat{m}(x;h))$ je hlavní člen, který byl v odstavci 2.2 podrobně odvozen pro Nadarayovy – Watsonovy odhady $\widehat{m}_{NW}$. Připomeňme si jeho tvar

$$\overline{MSE}(\widehat{m}(x;h)) = \underbrace{\frac{\sigma^2 V(K)}{Th}}_{\text{rozptyl}} + \underbrace{\frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 \left(m^{(\kappa)}(x)\right)^2}_{(\text{vychýlení})^2}, \tag{11}$$

kde

$$V(K) = \int_{-1}^{1} K^2(x)dx, \quad \beta_\kappa = \int_{-1}^{1} x^\kappa K(x)dx.$$

Soustřeďme se nyní na odhad funkce m v bodech plánu $x_i,\ i = 0, \dots T-1$. V tomto případě je vhodné odhad charakterizovat pomocí globální chyby, a to *průměrné střední kvadratické chyby AMSE* (Average Mean Square Error)

$$R_T(h) = \frac{1}{T}\sum_{i=0}^{T-1} E(\widehat{m}(x_i;h) - m(x_i))^2. \tag{12}$$

Hlavní člen této chyby lze na základě vztahů (10) a (11) vypočítat takto

$$\overline{R_T}(h) = \frac{1}{T}\sum_{i=0}^{T-1}\left(\frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 \left(m^{(\kappa)}(x_i)\right)^2\right).$$

Označíme-li

$$\left(\overline{m}^{(\kappa)}\right)^2 = \frac{1}{T}\sum_{i=0}^{T-1}\left(m^{(\kappa)}(x_i)\right)^2,$$

pak

$$\overline{R_T}(h) = \frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2\left(\overline{m}^{(\kappa)}(x)\right)^2.$$

V literatuře se často místo průměrné střední kvadratické chyby *AMSE* minimalizuje *integrální střední kvadratická chyba IMSE* (Integral Mean Square Error). V tomto případě se pak místo $\left(\overline{m}^{(\kappa)}\right)^2$ aplikuje výraz

$$A_\kappa = \int_0^1 \left(m^{(\kappa)}(x)\right)^2 dx.$$

My ho též užijeme v našich úvahách, $\overline{R_T}(h)$ má pak tvar

$$\overline{R_T}(h) = \frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa. \tag{13}$$

Hodnota h, pro kterou $\overline{R_T}(h)$ nabývá minimální hodnoty je určena vztahem

$$\frac{\partial \overline{R_T}(h)}{\partial h} = 0.$$

Odtud získáme teoretickou hodnotu optimální šířky vyhlazovacího okna

$$h_{opt} = \left(\frac{\sigma^2 V(K)(\kappa!)^2}{2\kappa T\beta_\kappa^2 A_\kappa}\right)^{\frac{1}{2\kappa+1}}. \tag{14}$$

Tato hodnota h_{opt} závisí na neznámých veličinách σ^2, $m^{(\kappa)}(x)$, a není tedy užitečná pro praktické účely. Má ovšem teoretický význam a umožní nám např. posoudit asymptotickou rychlost konvergence *AMSE*.

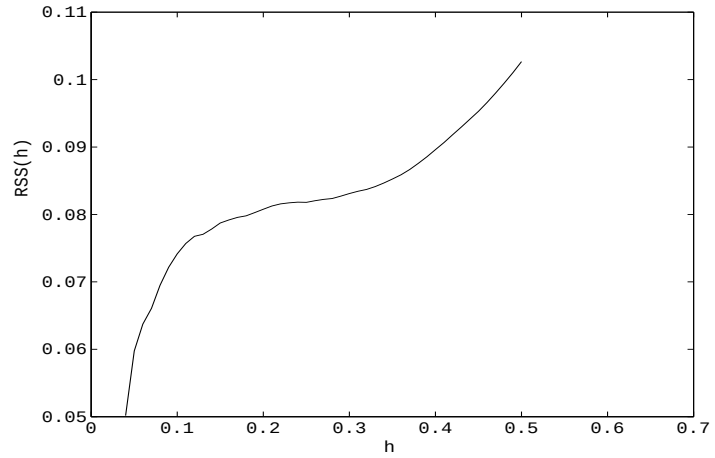
V praxi se také často minimalizuje tzv. *průměrná kvadratická chyba ASE* (Average Square Error), která je asymptoticky ekvivalentní s *AMSE* (viz např. [6])

$$ASE(h) = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - m(x_i)]^2. \tag{15}$$

Tato chybová funkce závisí na neznámých hodnotách regresní funkce $m(x_i)$. Nahrazením teoretických hodnot $m(x_i)$ naměřenými hodnotami Y_i získáme odhad $R_T(h)$, tzv. *residuální součet čtverců*

$$RSS_T(h) = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - Y_i]^2. \tag{16}$$

Avšak $RSS_T(h)$ je bohužel vychýlený odhad funkce $R_T(h)$. Jak si můžeme na obrázku 6 povšimnout, $RSS_T(h)$ je rostoucí funkce proměnné h. Minimalizace této funkce by tedy vedla k příliš malým hodnotám optimální šířky okna h (viz např. [12]).



Obrázek 6: *Chybová funkce $RSS_T(h)$ pro simulovaná data z obr.3.*

Podrobněji popisují vychýlení residuálního součtu čtverců následující lemma a věta.

Lemma 3.1.1. *$RSS_T(h)$ lze psát ve tvaru*

$$RSS_T(h) = ASE(h) + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right]. \tag{17}$$

Důkaz.

$$\begin{aligned} RSS_T(h) &= \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - Y_i]^2 = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - m(x_i) - \varepsilon_i]^2 \\ &= \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - m(x_i)]^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i(\widehat{m}(x_i;h) - m(x_i)) + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 \\ &= ASE(h) + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right]. \end{aligned}$$

□

Poznámka. Označme poslední člen B_{1T},tj.

$$B_{1T} := -\frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right].$$

Věta 3.1.2. *$RSS_T(h)$ je vychýlený odhad $R_T(h)$, neboť střední hodnota tohoto odhadu je*

$$E(RSS_T(h)) = R_T(h) + \sigma^2 - \frac{2\sigma^2}{T}\sum_{i=0}^{T-1} W_i(x_i).$$

Důkaz. Připomeňme základní předpoklady našeho modelu, tj. ε_i jsou nezávislé náhodné veličiny splňující podmínky

$$E(\varepsilon_i) = 0, \qquad var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2, \qquad i = 0, \ldots, T-1.$$

Počítejme střední hodnotu $E(RSS_T(h))$

$$\begin{aligned}
E(RSS_T(h)) &= E(ASE(h)) + E\left(\frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2\right) - E\left(\frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1} W_j(x_i)Y_j - m(x_i)\right]\right) \\
&= R_T(h) + \frac{1}{T}\sum_{i=0}^{T-1} E(\varepsilon_i^2) - \frac{2}{T}\sum_{i=0}^{T-1}\sum_{j=0}^{T-1}[W_j(x_i)E(\varepsilon_i Y_j) - m(x_i)E(\varepsilon_i)] \\
&= R_T(h) + \frac{1}{T}\sum_{i=0}^{T-1}\sigma^2 - \frac{2}{T}\sum_{i=0}^{T-1}\sum_{j=0}^{T-1} W_j(x_i)E(\varepsilon_i[m(x_j) + \varepsilon_j]) \\
&= R_T(h) + \sigma^2 - \frac{2}{T}\sum_{i=0}^{T-1}\sum_{j=0}^{T-1} W_j(x_i)E(\varepsilon_i\varepsilon_j) \\
&= R_T(h) + \sigma^2 - \frac{2}{T}\sum_{i=0}^{T-1} W_i(x_i)E(\varepsilon_i^2) \\
&= R_T(h) + \sigma^2 - \frac{2\sigma^2}{T}\sum_{i=0}^{T-1} W_i(x_i).
\end{aligned}$$

□

V dalších úvahách se budeme snažit „upravit“ residuální součet čtverců $RSS_T(h)$ tak, aby se stal nevychýleným, případně alespoň asymptoticky nevychýleným, odhadem chybové funkce $R_T(h)$.
Například Rice [17] uvažuje odhad

$$\widehat{R}_T(h) = RSS_T(h) - \hat{\sigma}^2 + \frac{2\hat{\sigma}^2}{T}\sum_{i=0}^{T-1} W_i(x_i), \tag{18}$$

kde $\hat{\sigma}^2$ je odhad σ^2

$$\hat{\sigma}^2 = \frac{1}{2T-2}\sum_{t=1}^{T-1}(Y_t - Y_{t-1})^2.$$

Podobný typ poprvé navrhl Mallows [16] a Craven & Wahba [1].

3.2 Metoda křížového ověřování

Jednou z nejznámějších metod pro hledání optimální šířky okna je tzv. *metoda křížového ověřování*. V literatuře ([4], [6], [11], [18]) se vyskytuje velmi často, a to nejen v souvislosti s jádrovými odhady, ale také např. v teorii vyhlazovacích splajnů. Hlavní myšlenka této metody spočívá v tom, že odhadneme hodnotu $\widehat{m}$ v bodě x_j bez použití tohoto bodu, tj. pomocí zbývajících $T-1$ bodů. Takto definované odhady pak použijeme při výpočtu residuálního součtu čtverců $RSS_T(h)$. Obdržíme tzv. *funkci křížového ověřování*, která bude již asymptoticky nevychýleným odhadem chybové funkce $R_T(h)$. Minimalizací této funkce získáme odhad optimální šířky okna $\hat{h}_{opt}$.

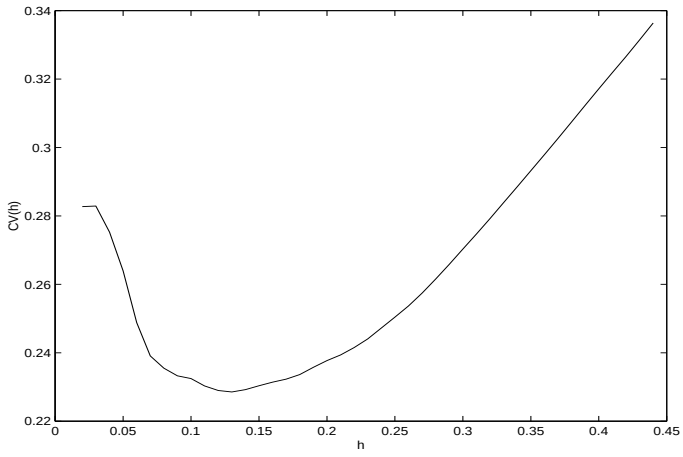
Označme $\widehat{m}_j(x_j;h)$ odhad hodnoty regresní funkce $\widehat{m}$ v bodě x_j bez použití tohoto bodu, tj.

$$\widehat{m}_j(x_j;h) = \sum_{\substack{i=0 \\ i\neq j}}^{T-1} W_i(x_j)Y_i.$$

S takto pozměněnými odhady má $RSS_T(h)$ tvar

$$CV(h) = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}_i(x_i;h) - Y_i]^2. \tag{19}$$

Funkce $CV(h)$ se nazývá *funkce křížového ověřování*. Tato funkce je znázorněna na obr.7.



Obrázek 7: *Funkce křížového ověřování $CV(h)$ pro simulovaná data z obr.3. Při jádrovém odhadu bylo použito jádra třídy S_{02} – viz tab.1.*

Odhad optimální šířky okna $\hat{h}_{opt}$ definujeme jako hodnotu, kde funkce $CV(h)$ nabývá svého minima, tj.

$$\hat{h}_{opt} = \underset{h\in(0,1)}{\arg\min}\ CV(h).$$

Soustřeďme se na statistické vlastnosti funkce křížového ověřování. V následujících úvahách ukážeme, že na rozdíl od residuálního součtu čtverců je asymptoticky nevychýleným odhadem funkce $R_T(h)$, případně $ASE(h)$.
Stejnými úpravami jako v odstavci 3.1 můžeme uvažovat funkci křížového ověřování v následujícím tvaru

$$CV(h) = ASE(h) + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i[\sum_{\substack{j=0\\j\neq i}}^{T-1} W_j(x_i)Y_j - m(x_i)].$$

Označme poslední člen tohoto vyjádření B_{2T}, tj.

$$B_{2T} = -\frac{2}{T}\sum_{i=0}^{T}\varepsilon_i[\sum_{\substack{j=0\\j\neq i}}^{T-1} W_j(x_i)Y_j - m(x_i)].$$

Tento výraz je podobný výrazu B_{1T}, který tvoří hlavní část vychýlení $RSS_T(h)$. Následující věta ukazuje, že střední hodnota B_{2T} je nulová, a tedy střední hodnota $CV(h)$ je hodnota funkce $R_T(h)$ posunutá o σ^2.

Věta 3.2.1. *Střední hodnota B_{2T} je nulová, tj.*

$$E(B_{2T}) = 0.$$

Důkaz. Při vyjádření střední hodnoty využijeme vlastností modelu $E(\varepsilon_i) = 0$ pro $i = 0, \ldots, T-1$ a také toho, že chyby jsou navzájem nezávislé, a proto $E(\varepsilon_i\varepsilon_j) = 0$ pro $i \neq j$.

$$\begin{aligned}
E(B_{2T}) &= E\left(-\frac{2}{T}\sum_{i=0}^{T}\varepsilon_i[\sum_{\substack{j=0\\j\neq i}}^{T-1} W_j(x_i)Y_j - m(x_i)]\right)\\
&= -\frac{2}{T}\sum_{i=0}^{T-1}\sum_{\substack{j=0\\j\neq i}}^{T-1}[W_j(x_i)E(\varepsilon_i Y_j) - m(x_i)E(\varepsilon_i)]\\
&= -\frac{2}{T}\sum_{i=0}^{T-1}\sum_{\substack{j=0\\j\neq i}}^{T-1} W_j(x_i)E(\varepsilon_i[m(x_j) + \varepsilon_j])\\
&= -\frac{2}{T}\sum_{i=0}^{T-1}\sum_{\substack{j=0\\j\neq i}}^{T-1} W_j(x_i)E(\varepsilon_i\varepsilon_j) = 0.
\end{aligned}$$

□

Poznamenejme, že samotný fakt, že výraz B_{2T} má nulovou střední hodnotu, ještě nezaručuje, že $\hat{h}_{opt}$ minimalizuje chybovou funkci $R_T(h)$, případně ekvivalentní $ASE(h)$.

Pro metodu křížového ověřování by měl být splněn také předpoklad, že výraz B_{2T} stejnoměrně konverguje k nule v závislosti na h. V praxi se však často stává, že člen B_{2T} nesplňuje tyto podmínky a ovlivňuje vychýlení funkce $CV(h)$. Její minimum je pak většinou menší než skutečná hodnota optimální šířky okna h_{opt}. K této situaci dochází zpravidla při malém rozsahu dat ($T < 50$).

Příklad.
Na simulovaných datech v systému MATLAB jsme porovnávali odhady získané metodou křížového ověřování s teoretickou optimální šířkou okna. Pozorování Y_t, pro $t = 0, \ldots, T-1 = 99$, byla vygenerována s náhodnými normálně rozloženými chybami s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.2$. Regresní funkce byla v našem případě

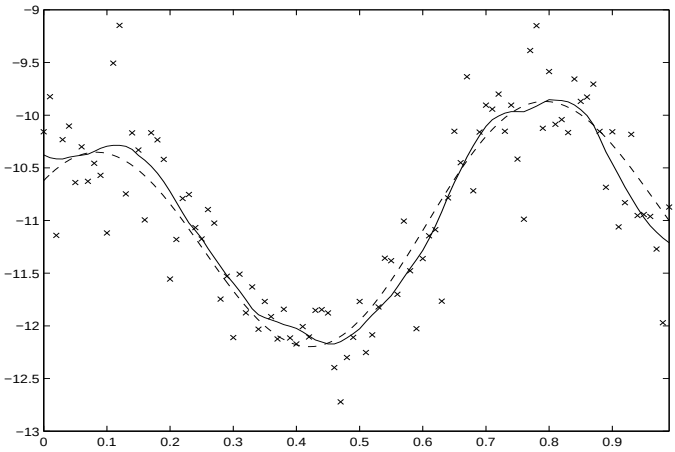
$$m(x) = \cos(9x - 7) - (3 + x^{12})/6 + 8^{x-1}.$$

Při výpočtech jsme použili Nadarayovy – Watsonovy estimátory a jádra třídy $S_{0\kappa}$ (viz tab.1) pro $\kappa = 2,\ 4,\ 6,\ 8$. Bylo vygenerováno 200 řad. U každé řady byly získány odhady optimální šířky okna tak, že jsme nejprve vypočítali hodnoty funkce křížového ověřování v 321 bodech ekvidistantně rozložených na intervalu $[0.01, 0.99]$ a pak z nich vybrali tu hodnotu, kde tato funkce nabývala svého minima. V tabulce 2 jsou uvedeny střední hodnoty a směrodatné odchylky všech odhadů, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot a $std(\hat{h}_{opt})$ je jejich směrodatná odchylka, h_{opt} označuje teoretickou optimální hodnotu spočtenou dle vzorce (14).

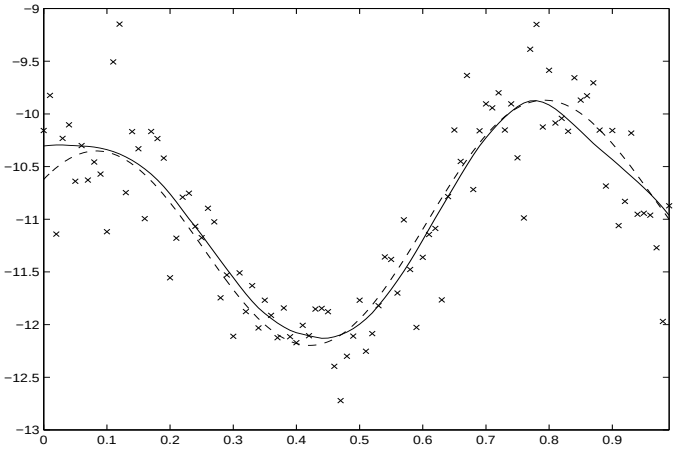
Tabulka 2: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných metodou křížového ověřování.*

κ	2	4	6	8
h_{opt}	0.0978	0.2488	0.4056	0.5684
$E(\hat{h}_{opt})$	0.0876	0.1942	0.3001	0.3948
$std(\hat{h}_{opt})$	0.0235	0.0457	0.0782	0.1104

Vybereme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Pro odhad regresní funkce jsme použili jádro třídy S_{04}. V tomto případě byla vybrána optimální šířka okna $\hat{h}_{opt} = 0.1364$ jako minimum funkce $CV(h)$. Na obr.8 jsou vykreslena simulovaná data, regresní funkce $m(x)$ a její odhad s tímto parametrem. Obr.9 znázorňuje jádrový odhad s použitím teoretické optimální šířky okna $h_{opt} = 0.2488$.



Obrázek 8: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1364$.



Obrázek 9: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s optimální šířkou okna* $h_{opt} = 0.2488$.

3.3 Penalizační funkce

Druhým často používaným postupem při hledání optimální šířky okna je tzv. *metoda penalizačních funkcí*. Tato metoda také vychází z residuálního součtu čtverců $RSS_T(h)$ jako vychýleného odhadu funkce $R_T(h)$, případně $ASE(h)$. Její hlavní myšlenkou je vhodná „úprava“ funkce $RSS_T(h)$, která vede k asymptotickému zanedbání jejího vychýlení. Na obr.6 je znázorněna funkce $RSS_T(h)$ jako rostoucí funkce proměnné h. Modifikace spočívá ve vynásobení této funkce určitou funkcí, která nabývá velkých hodnot pro malá h a naopak pro velké hodnoty h konverguje k nule. Takovou funkci nazýváme *penalizační funkce*, neboť penalizuje příliš malé hodnoty h. Vznikne tak nová chybová funkce. Odhad optimální šířky okna metodou penalizačních funkcí budeme definovat jako hodnotu, pro kterou tato funkce nabývá svého minima. Budeme-li zkoumat tuto funkci podrobněji, zjistíme, že její vychýlení obsahuje členy, které se asymptoticky vzájemně vyruší.

Definice 3.1. Libovolnou funkci $\Xi(u)$, jejíž Taylorův rozvoj 1. řádu se středem v nule je tvaru

$$\Xi(u) = 1 + 2u + O(u^2),$$

nazýváme *penalizační funkce.*

Příklady některých penalizačních funkcí jsou uvedeny v následujícím přehledu. Jejich průběh je znázorněn na obr.10.

Příklady penalizačních funkcí:

1. *Generalized cross-validation* (Craven, Wahba 1979; Li 1985)

$$\Xi_{GCV}(u) = \frac{1}{(1-u)^2}$$

2. *Akaike's Information Criterion* (Akaike 1970)

$$\Xi_{AIC}(u) = e^{2u}$$

3. *Finite Prediction Error* (Akaike 1974)

$$\Xi_{FPE}(u) = \frac{1+u}{1-u}$$

4. *Shibata's model selector* (Shibata 1981)

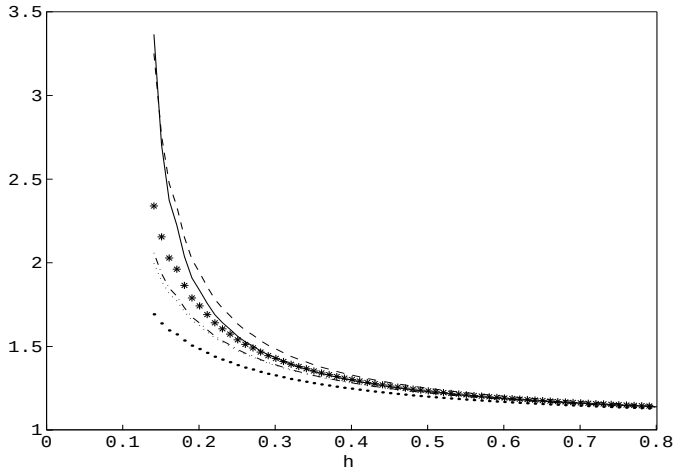
$$\Xi_S(u) = 1 + 2u$$

5. *Rice's bandwidth selector* (Rice 1984)

$$\Xi_R(u) = \frac{1}{1-2u}$$

6. *ET bandwidth selector* (Koláček 2001)

$$\Xi_{ET}(u) = e^{\frac{4}{\pi}\tan\frac{\pi}{2}u}$$



Obrázek 10: *Graf 6 penalizačních funkcí v závislosti na h: - - Rice, – ET, ** Generalized, -.- FPE, .. Akaike, ●● Shibata.*

Myšlenka metody penalizačních funkcí spočívá v následujícím. Nechť $\Xi(u)$ je penalizační funkce. Každý člen residuálního součtu čtverců (16) $RSS_T(h)$ vynásobíme výrazem $\Xi(W_i(x_i))$. Důvodem pro tuto úpravu je fakt, že $\Xi(W_i(x_i))$ nabývá velkých hodnot pro malá h. Připomeňme, že funkce $RSS_T(h)$ je rostoucí, a tedy její minimalizace vedla právě k příliš malým hodnotám h. Vynásobením výrazem $\Xi(W_i(x_i))$ penalizujeme tyto hodnoty. Dostáváme tedy novou chybovou funkci

$$G(h) = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - Y_i]^2\Xi(W_i(x_i)). \tag{20}$$

Hodnotu, pro kterou tato funkce nabývá svého minima, definujeme jako odhad optimální šířky okna

$$\hat{h}_{opt} = \underset{h\in(0,1)}{\arg\min}\ G(h).$$

Průběh funkce $G(h)$ a její minima pro různé penalizační funkce znázorňuje obr.11. Nyní podrobněji rozebereme asymptotické chování této funkce. Následující věta ukazuje, že střední hodnota $G(h)$ je hodnota funkce $R_T(h)$ posunutá o σ^2.

Věta 3.3.1. *Střední hodnota funkce $G(h)$ je rovna*

$$E(G(h)) = R_T(h) + \sigma^2.$$

Důkaz. Penalizační funkci $\Xi(W_i(x_i))$ nahradíme v (20) jejím Taylorovým rozvojem, tj.

$$G(h) = \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - Y_i]^2(1 + 2W_i(x_i)) + O(T^{-3}h^{-2}).$$

Poslední člen můžeme zanedbat, funkci $RSS_T(h)$ vyjádříme podle vztahu (17)

$$G(h) = \frac{1}{T}\sum_{i=0}^{T-1}\left([\widehat{m}(x_i;h) - m(x_i)]^2 + \varepsilon_i^2 - 2\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right]\right)(1 + 2W_i(x_i)),$$

roznásobením dostáváme

$$\begin{aligned} G(h) = {} & \frac{1}{T}\sum_{i=0}^{T-1}[\widehat{m}(x_i;h) - m(x_i)]^2 + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right] \\ & + \frac{2}{T}\sum_{i=0}^{T-1}W_i(x_i)[\widehat{m}(x_i;h) - m(x_i)]^2 + \frac{2}{T}\sum_{i=0}^{T-1}W_i(x_i)\varepsilon_i^2 \\ & - \frac{4}{T}\sum_{i=0}^{T-1}W_i(x_i)\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right]. \end{aligned}$$

Čtvrtý a šestý člen můžeme opět zanedbat, neboť jsou řádu $O(T^{-2}h^{-1})$, tj.

$$G(h) = ASE(h) + \frac{1}{T}\sum_{i=0}^{T-1}\varepsilon_i^2 - \frac{2}{T}\sum_{i=0}^{T-1}\varepsilon_i\left[\sum_{j=0}^{T-1}W_j(x_i)Y_j - m(x_i)\right] + \frac{2}{T}\sum_{i=0}^{T-1}W_i(x_i)\varepsilon_i^2.$$

Spočteme-li střední hodnotu $G(h)$, poslední dva členy se vyruší

$$E(G(h)) = R_T(h) + \sigma^2 - \frac{2\sigma^2}{T}\sum_{i=0}^{T-1}W_i(x_i) + \frac{2\sigma^2}{T}\sum_{i=0}^{T-1}W_i(x_i).$$

□

Poznámka
Metoda křížového ověřování je také způsob penalizace funkce $RSS(h)$, neboť

$$\frac{CV(h)}{RSS(h)} = 1 + 2W_i(x_i) + O(T^{-2}h^{-2}).$$

Podrobnější důkaz můžeme najít v [5].

Příklad.
Na simulovaných datech v systému MATLAB jsme porovnávali odhady získané pomocí uvedených penalizačních funkcí mezi sebou a také s teoretickou optimální šířkou okna. Pozorování Y_t, pro $t = 0, \ldots, T-1 = 99$, byla vygenerována s náhodnými normálně

rozloženými chybami s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.1$. Regresní funkce byla v našem případě

$$m(x) = 11 - 1/3 \tan(5 + x^6) \sin(12\, x).$$

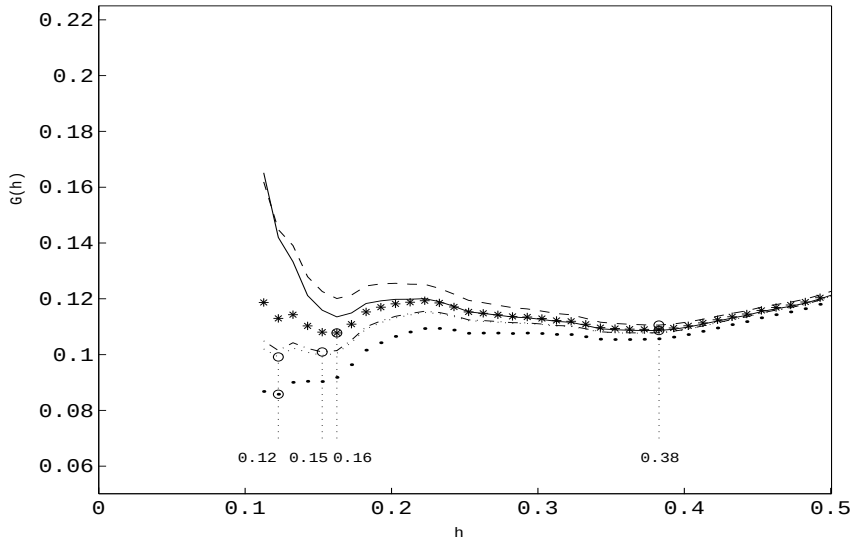
Při výpočtech jsme použili Nadarayovy – Watsonovy estimátory a jádra třídy $S_{0\kappa}$ (viz tab.1) pro $\kappa = 2,\ 4,\ 6,\ 8$. Bylo vygenerováno 200 řad. U každé řady byly získány odhady optimální šířky okna tak, že jsme nejprve spočítali hodnoty funkce $G(h)$ v 321 bodech ekvidistantně rozložených na intervalu $[0.01, 0.99]$ a pak z nich vybrali tu hodnotu, kde tato funkce nabývala svého minima. V tabulce 3 jsou uvedeny střední hodnoty všech odhadů. V prvním sloupci je označení penalizačních funkcí, které byly použity, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot, h_{opt} označuje teoretickou optimální hodnotu spočtenou dle vzorce (14).

Tabulka 3: *Střední hodnoty odhadů parametru h_{opt} získaných metodou penalizačních funkcí.*

	κ	2	4	6	8
	h_{opt}	0.0691	0.1739	0.2721	0.3742
GCV	$E(\hat{h}_{opt})$	0.0597	0.1317	0.2101	0.2818
AIC	$E(\hat{h}_{opt})$	0.0461	0.1115	0.1824	0.2549
FPE	$E(\hat{h}_{opt})$	0.0498	0.1151	0.1862	0.2569
S	$E(\hat{h}_{opt})$	0.0251	0.0646	0.1192	0.1868
R	$E(\hat{h}_{opt})$	0.0661	0.1432	0.2236	0.3023
ET	$E(\hat{h}_{opt})$	0.0623	0.1345	0.2131	0.2884

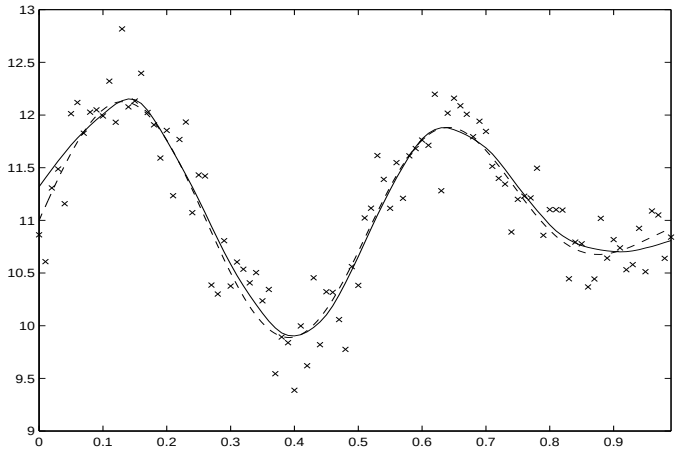
Porovnáme-li v tabulce hodnoty $\hat{h}_{opt}$ pro různé penalizační funkce s teoretickou optimální šířkou okna, je zřejmé, že jsou pro všechna κ výsledné odhady menší. S rostoucím κ jsou větší rozdíly mezi výsledky získanými jednotlivými metodami. To je způsobeno především tím, že odhady regresní funkce s jádry vyšších řádů jsou méně citlivé na malou změnu vyhlazovacího parametru. Je nezbytné si uvědomit, že všechny metody jsou pouze asymptoticky ekvivalentní a také $\hat{h}_{opt}$ jsou jen asymptoticky nevychýlenými odhady optimální šířky okna h_{opt}. Proto především pro menší rozsah dat dochází k rozdílům ve výsledcích.

Zvolíme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Průběh funkce $G(h)$ a její minima pro různé penalizační funkce a jádro třídy S_{08} znázorňuje obr.11. Nejblíže optimální šířce okna $h_{opt} = 0.3742$ byly v tomto případě výsledky získané pomocí ET a Riceho penalizační funkce. V této i celkově v dalších simulacích se jeví právě tyto dvě penalizační funkce jako nejvhodnější. Naopak, nejhorší výsledky



Obrázek 11: *Různě penalizovaná $RSS_T(h)$: - - Rice, – ET, ** Generalized, -.- FPE, .. Akaike, •• Shibata a její minima pro jádro třídy S_{08}. Teoretická hodnota $h_{opt} = 0.3742$.*

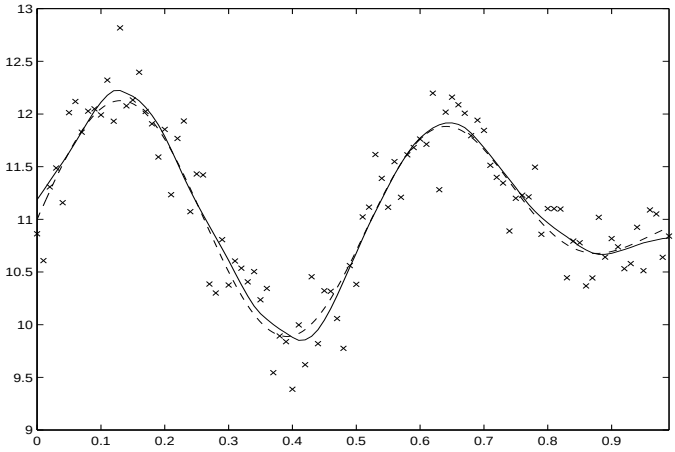
byly získány při použití AIC a Shibatovy penalizační funkce.
Pro odhad regresní funkce jsme použili jádro třídy S_{04}. Obr.12 znázorňuje jádrový odhad s použitím teoretické optimální šířky okna $h_{opt} = 0.1739$.



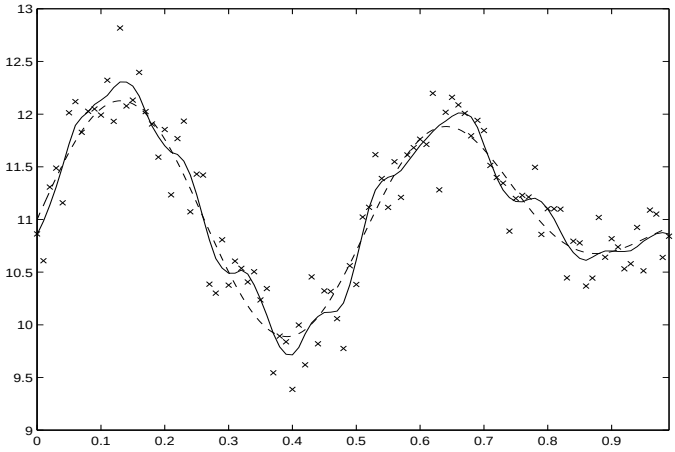
Obrázek 12: *Symboly × označují naměřené hodnoty Y s chybou $\sigma^2 = 0.2$. Čárkovaně je zobrazena skutečná regresní funkce $m(x) = 11 - 1/3 \tan(5 + x^6) \sin(12\,x)$. Plná čára znázorňuje jádrový odhad této regresní funkce s optimální šířkou okna $h_{opt} = 0.1739$. pro jádro $K \in S_{04}$*

Všimněme si ještě ET a Shibatovy penalizační funkce. V tomto případě byla vybrána optimální šířka okna pro ET $\hat{h}_{opt} = 0.1332$, odhad regresní funkce $m(x)$ s tímto parametrem je na obr. 13. Při použití Shibatovy penalizační funkce vyšel odhad optimální

šířky okna $\hat{h}_{opt} = 0.0672$. Na obr.14 je vidět, že je tato hodnota příliš malá a odhad podhlazený.



Obrázek 13: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = 11 - 1/3\,\tan(5 + x^6)\sin(12\,x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1332$. *pro jádro* $K \in S_{04}$

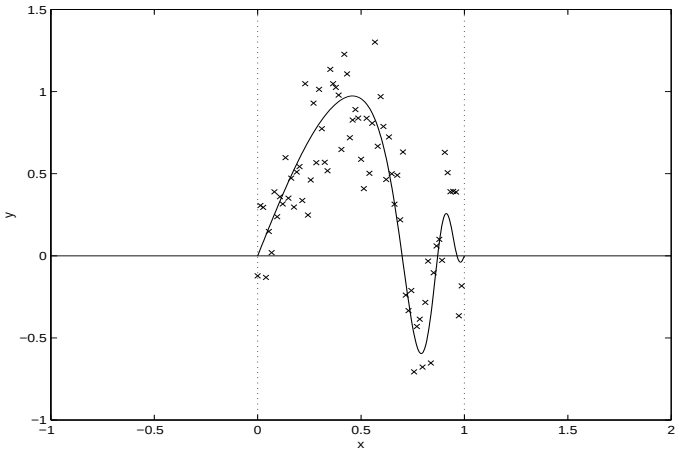


Obrázek 14: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = 11 - 1/3\,\tan(5 + x^6)\sin(12\,x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.0672$. *pro jádro* $K \in S_{04}$

4 Cyklický model

Jak jsme se již zmínili v odstavci 3.1, v blízkosti hraničních bodů se mohou objevit tzv. „hraniční efekty“ a odhadům v těchto bodech je nutno věnovat zvláštní pozornost. Cílem této práce je však zaměřit se především na hledání optimálního vyhlazovacího parametru h, a proto si situaci mírně zjednodušíme tím, že budeme uvažovat tzv. „cyklický model“. Cyklický model se často používá v teoretických studiích (např. [9], [10], [13], [14], [15]), právě za účelem odstranění problémů v krajních bodech intervalu. Tento model se od původního modelu s pevným plánem liší v následujícím:

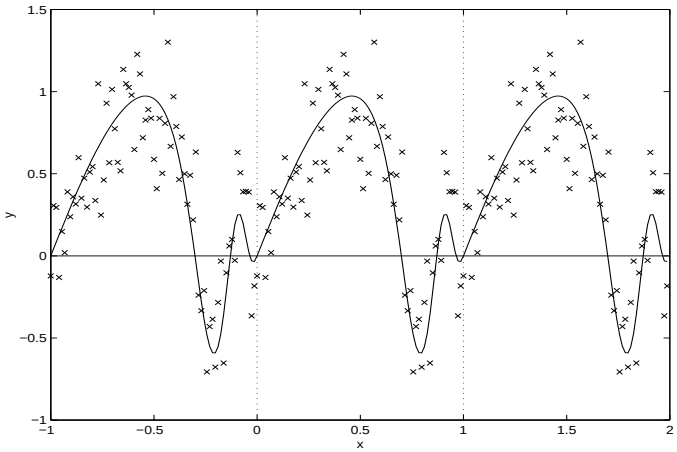
- Body plánu x_t rozšíříme na celou reálnou osu a zachováme jejich ekvidistantnost, tj. $x_t = t/T,\ t \in \mathbb{Z}$
- Předpokládáme, že $m(x)$ je hladká periodická funkce s periodou 1 a odhad je získán jádrovým vyhlazováním na rozšířené řadě $\widetilde{Y}_t$, kde $\widetilde{Y}_{t+kT} = Y_t$ pro $k = 0, \pm 1, \ldots$ (viz obr. 15, obr. 16).



Obrázek 15: *Původní model pro regresní funkci* $m(x) = \sin(\pi x)\cos(3\pi x^5)$. *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.05$. *Plnou čarou je zobrazena regresní funkce* $m(x)$.

4.1 Vlastnosti cyklického modelu

V tomto odstavci si všimneme některých pozoruhodných vlastností cyklického modelu, kterých budeme moci později využít. Budeme se zabývat především odhady v bodech plánu, neboť ty pak budou dále potřeba k odhadu optimální šířky okna. Získáme například zajímavý výsledek, že hodnoty Nadarayových – Watsonových a lokálně lineárních estimátorů v bodech plánu jsou totožné. To nám následně umožní zjednodušení zápisu těchto odhadů.



Obrázek 16: *Cyklický model pro regresní funkci* $m(x) = \sin(\pi x)\cos(3\pi x^5)$ *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.05$. *Plnou čarou je zobrazena regresní funkce* $m(x)$.

Připomeňme si nejznámější typy jádrových odhadů regresní funkce. Ty jsou v cyklickém modelu tvaru

1. **Nadarayovy – Watsonovy odhady** (1964)

$$\widehat{m}_{NW}(x;h) = \frac{\sum_{k=-T}^{2T-1} K_h(x_k - x)\widetilde{Y}_k}{\sum_{k=-T}^{2T-1} K_h(x_k - x)}$$

2. **Lokální lineární estimátory** (Stone 1977, Cleveland 1979)

$$\widehat{m}_{LL}(x;h) = \frac{1}{T}\sum_{k=-T}^{2T-1} \frac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)(x_k - x)\}K_h(x_k - x)\widetilde{Y}_k}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2},$$

   kde

$$\hat{s}_r(x;h) = \frac{1}{T}\sum_{k=-T}^{2T-1} (x_k - x)^r K_h(x_k - x)$$

3. **Pristleyho – Chaovy odhady** (1972)

$$\widehat{m}_{PCH}(x;h) = \frac{1}{T}\sum_{k=-T}^{2T-1} K_h(x_k - x)\widetilde{Y}_k$$

4. **Gasserovy – Müllerovy odhady** (1979)

$$\widehat{m}_{GM}(x;h) = \sum_{k=-T}^{2T-1} \widetilde{Y}_k \int_{s_{k-1}}^{s_k} K_h(t-x)dt,$$

kde

$$s_k = \frac{x_k + x_{k+1}}{2},\ k = -T, \ldots, 2T-2, \quad s_{-T-1} = -1, \quad s_{2T-1} = 2.$$

V cyklickém modelu lze jádrové odhady obecně zapsat ve tvaru

$$\widehat{m}(x;h) = \sum_{k=-T}^{2T-1} W_k^{(j)}(x)\widetilde{Y}_k, \tag{21}$$

kde váhy $W_k^{(j)}(x)$, $j \in \{NW, LL, PCH, GM\}$, odpovídají váhám u odhadů $\widehat{m}_{NW}$, $\widehat{m}_{LL}$, $\widehat{m}_{PCH}$, $\widehat{m}_{GM}$.

Věta 4.1.1. *Nechť $h \in (0,1), t \in \{0,1,\ldots,T-1\}$. Pak*

$$\frac{1}{T}\sum_{k=-T}^{2T-1} K_h(x_k - x_t) = \frac{1}{T}\sum_{k=-T+1}^{T-1} K_h(x_k),$$

tj. suma nezávisí na indexu t.
Důkaz.

$$\begin{aligned}
\frac{1}{T}\sum_{k=-T}^{2T-1} K_h(x_k - x_t) &= \frac{1}{T}\sum_{k=-T}^{2T-1} K_h(x_{k-t}) = \frac{1}{T}\sum_{k=-T-t}^{2T-t-1} K_h(x_k) \\
&= \frac{1}{T}\sum_{k=-T-t}^{-T} K_h(x_k) + \frac{1}{T}\sum_{k=-T+1}^{T-1} K_h(x_k) + \frac{1}{T}\sum_{k=T}^{2T-t-1} K_h(x_k) \\
&= \frac{1}{T}\sum_{k=-T+1}^{T-1} K_h(x_k).
\end{aligned}$$

První a třetí suma jsou rovny 0, neboť $|x_k| \geq 1$ pro $|k| \geq T$ a tedy $K_h(x_k) = 0$. □

Označme

$$C_T := \frac{1}{T}\sum_{k=-T+1}^{T-1} K_h(x_k).$$

Tuto konstantu budeme v dalším textu používat. Můžeme pomocí ní např. vyjádřit jednoduššeji hodnoty Nadarayových – Watsonových vah v bodech plánu

$$W_k^{(NW)}(x_t) = \frac{K_h(x_k - x_t)}{TC_T}.$$

Poznámka. V případě, že by se v některém z odhadů dělilo nulou, např. kdyby $C_T = 0$, dodefinujeme příslušný odhad jako nulový.

Lemma 4.1.2. *Nechť $h \in (0,1), t \in \{0,1,\dots,T-1\}$ a r je liché. Pak*

$$\hat{s}_r(x_t; h) = 0.$$

Důkaz.

$$\begin{aligned}
\hat{s}_r(x_t; h) &= \frac{1}{T}\sum_{k=-T}^{2T-1}(x_k - x_t)^r K_h(x_k - x_t) \\
&= \frac{1}{T}\sum_{k=-T}^{2T-1} x_{k-t}^r K_h(x_{k-t}) \\
&= \frac{1}{T}\sum_{k=-T-t}^{2T-1-t} x_k^r K_h(x_k) \\
&= \frac{1}{T}\sum_{k=-T-t}^{-T} x_k^r K_h(x_k) + \frac{1}{T}\sum_{k=-T+1}^{-1} x_k^r K_h(x_k) \\
&\quad + \frac{1}{T}\sum_{k=0}^{T-1} x_k^r K_h(x_k) + \frac{1}{T}\sum_{k=T}^{2T-1} x_k^r K_h(x_k)
\end{aligned}$$

První a poslední suma jsou rovny 0, neboť $|x_k| \geq 1$ pro $|k| \geq T$ a tedy $K_h(x_k) = 0$. Připomeňme ještě, že platí $x_{-k} = \frac{-k}{T} = -x_k$ a $K_h(x_k) = K_h(-x_k)$. Celkem

$$\begin{aligned}
\hat{s}_r(x_t; h) &= \frac{1}{T}\sum_{k=-T+1}^{-1} x_k^r K_h(x_k) + \frac{1}{T}\sum_{k=0}^{T-1} x_k^r K_h(x_k) \\
&= \frac{1}{T}\sum_{k=1}^{T-1} x_{-k}^r K_h(x_{-k}) + \frac{1}{T}\sum_{k=1}^{T-1} x_k^r K_h(x_k) + \frac{1}{T} x_0 K_h(x_0) \\
&= \frac{1}{T}\sum_{k=1}^{T-1}(-x_k)^r K_h(-x_k) + \frac{1}{T}\sum_{k=1}^{T-1} x_k^r K_h(x_k) \\
&= \frac{1}{T}\sum_{k=1}^{T-1}(x_k^r - x_k^r) K_h(x_k) = 0.
\end{aligned}$$

□

Důsledek 4.1.3. *Nechť $h \in (0,1), t \in \{0,1,\dots,T-1\}$. Pak*

$$W_k^{(LL)}(x_t) = W_k^{(NW)}(x_t),$$

pro $k \in \{-T, -T+1, \dots, 2T-1\}$.

Důkaz.

$$W_k^{(LL)}(x_t) = \frac{1}{T}\frac{\{\hat{s}_2(x_t;h) - \hat{s}_1(x_t;h)(x_k - x_t)\}K_h(x_k - x_t)}{\hat{s}_2(x_t;h)\hat{s}_0(x_t;h) - \hat{s}_1(x_t;h)^2}.$$

Podle předchozího lemmatu je $\hat{s}_1(x_t;h) = 0$, a proto

$$\begin{aligned} W_k^{(LL)}(x_t) &= \frac{1}{T}\frac{\hat{s}_2(x_t;h)K_h(x_k - x_t)}{\hat{s}_2(x_t;h)\hat{s}_0(x_t;h)} \\ &= \frac{1}{T}\frac{K_h(x_k - x_t)}{\hat{s}_0(x_t;h)} \\ &= \frac{1}{T}\frac{K_h(x_k - x_t)}{\frac{1}{T}\sum_{k=-T}^{2T-1} K_h(x_k - x_t)} = W_k^{(NW)}(x_t) \end{aligned}$$

□

Poznámka. Výše uvedený důsledek tedy ukazuje, že hodnoty Nadarayových – Watsonových a lokálně lineárních odhadů v bodech plánu jsou stejné. Využijeme této zajímavé skutečnosti a pro lepší přehlednost nebudeme index j u $W_k(x)$ v dalším textu uvádět.

Lemma 4.1.4. *Nechť $k, t \in \{0, 1, \ldots, T-1\}$, pro všechny uvažované typy odhadů platí*

$$W_k(x_t) = W_0(x_{t-k}).$$

Důkaz. Tvrzení nejprve dokážeme pro Nadarayovy – Watsonovy, tj. i pro lokálně lineární odhady. Je zřejmé

$$\begin{aligned} W_k(x_t) &= \frac{1}{TC_T}K_h(x_k - x_t) = \frac{1}{TC_T}K_h(0 - (x_t - x_k)) \\ &= \frac{1}{TC_T}K_h(x_0 - x_{t-k}) = W_0(x_{t-k}). \end{aligned}$$

V případě Pristleyho – Chaových odhadů je situace velmi podobná, neboť se liší pouze o konstantu C_T.
Závěrem dokážeme tvrzení pro Gasser – Müllerovy estimátory, které jsou tvaru

$$W_k(x_t) = \int_{s_{k-1}}^{s_k} K_h(u - x_t)du,$$

kde $s_k = \frac{x_k + x_{k+1}}{2} = \frac{\frac{2k+1}{T}}{2} = \frac{1}{2T} + \frac{k}{T} = s_0 + x_k$. Substitucí $v = u - x_k$ dostáváme

$$\int_{s_{k-1}}^{s_k} K_h(u - x_t)du = \int_{s_{-1}}^{s_0} K_h(v + x_k - x_t)dv = \int_{s_{-1}}^{s_0} K_h(v - (x_t - x_k))dv = W_0^{(GM)}(x_{t-k}).$$

Lemma 4.1.5. *Nechť $k, t \in \{0, 1, \ldots, T-1\}$, pro všechny uvažované typy odhadů platí*

$$W_{k\mp T}(x_t) = W_k(x_t \pm 1).$$

Důkaz. Tento vztah lze dokázat stejným způsobem jako v předchozím lemmatu. □

4.2 Využití Fourierovy analýzy

Teorie Fourierovy analýzy je velmi rozsáhlá a často se používá v různých odvětvích aplikované matematiky. I my využijeme v dalších úvahách některých jejích výsledků. Nejprve uvedeme základní definice a věty, které budeme dále potřebovat. Pak se zaměříme především na chybové funkce $R_T(h)$ a $RSS_T(h)$ a budeme zkoumat jejich tvar po Fourierově transformaci. Toho následně využijeme v dalších odstavcích k získání jiných metod pro odhad optimální šířky okna.

Definice 4.1. Nechť $\boldsymbol{x} = (x_0, \ldots, x_{T-1})' \in \mathbb{C}^T$. Vektor $\boldsymbol{x}^{\pm} = (x_0^{\pm}, \ldots, x_{T-1}^{\pm})' \in \mathbb{C}^T$, kde

$$x_t^{\pm} = \sum_{k=0}^{T-1} x_k e^{\pm \frac{i2\pi kt}{T}}, \quad t = 0, 1, \ldots, T-1,$$

se nazývá *diskrétní Fourierova transformace* vektoru $\boldsymbol{x}$. Píšeme $\boldsymbol{x}^{\pm} = DFT^{\pm}(\boldsymbol{x})$.

Poznámka. Z definice diskrétní Fourierovy transformace je vidět, že $x_{-t}^{\pm} = x_{T-t}^{\pm}$. Pokud $\boldsymbol{x} \in \mathbb{R}^T$ platí také $x_{-t}^{\pm} = \overline{x_t^{\pm}}$, celkem tedy dostáváme symetrii $x_{T-t}^{\pm} = \overline{x_t^{\pm}}$, kterou využijeme v pozdějších úvahách.

Definice 4.2. *Periodogram* vektoru $\boldsymbol{Y} = (Y_0, \ldots, Y_{T-1})'$ je vektor $\boldsymbol{I}_Y = (I_{Y_0}, \ldots, I_{Y_{T-1}})'$

$$I_{Y_t} = \frac{|Y_t^-|^2}{2\pi T}, \quad t = 0, \ldots T-1, \tag{22}$$

kde $\boldsymbol{Y}^- = DFT^-(\boldsymbol{Y})$.

Poznámka. Označme $S_t = m(x_t)$, $t = 0, \ldots, T-1$. Periodogramy a Fourierovy transformace pro vektory $\boldsymbol{S} = (S_0, \ldots, S_{T-1})'$ a $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_{T-1})'$ jsou definovány podobně jako pro $\boldsymbol{Y}$.

Věta 4.2.1. *(Parsevalova identita – diskrétní případ)*
Nechť $\boldsymbol{x} \in \mathbb{C}^T$, pak platí

$$\sum_{t=0}^{T-1} |x_t|^2 = \frac{1}{T} \sum_{t=0}^{T-1} |x_t^{\pm}|^2,$$

kde $\boldsymbol{x}^{\pm} = DFT^{\pm}(\boldsymbol{x})$.

Důkaz. Důkaz můžeme najít např. v [3]. □

Definice 4.3. Nechť $\boldsymbol{x} = (x_0, \ldots, x_{T-1})'$, $\boldsymbol{y} = (y_0, \ldots, y_{T-1})' \in \mathbb{C}^T$;

$$z_t = \sum_{k=0}^{T-1} x_{<t-k>_T} y_k,$$

kde $< t-k >_T$ označuje $(t-k)$ mod T. Pak $\boldsymbol{z} = (z_0, \ldots, z_{T-1})'$ nazýváme *diskrétní cyklickou konvolucí* vektorů $\boldsymbol{x}$ a $\boldsymbol{y}$; píšeme $\boldsymbol{z} = \boldsymbol{x} \circledast \boldsymbol{y}$.

Věta 4.2.2. *Nechť $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{C}^T$, pak pro $t \in \{0, \ldots, T-1\}$ platí*

$$(\boldsymbol{x} \circledast \boldsymbol{y})_t^{\pm} = x_t^{\pm} y_t^{\pm}.$$

Důkaz. Důkaz můžeme najít např. v [3]. □

Označení. Označme $\boldsymbol{w} := (w_0, w_1, \ldots, w_{T-1})'$, kde

$$w_t = W_0(x_t - 1) + W_0(x_t) + W_0(x_t + 1). \tag{23}$$

Připomeňme ještě, že v celé kapitole uvažujeme cyklický model. Díky tomuto předpokladu lze na jádrové odhady v bodech plánu pohlížet jako na diskrétní cyklickou konvoluci váhového vektoru $\boldsymbol{w}$ a vektoru naměřených hodnot $\boldsymbol{Y}$, tj. $\boldsymbol{w} \circledast \boldsymbol{Y}$. Tuto skutečnost popisuje následující věta.

Věta 4.2.3. *Nechť $h \in (0,1)$, $K \in S_{0\kappa}, t \in \{0, \ldots, T-1\}$. Pak*

$$\widehat{m}(x_t; h) = \sum_{k=0}^{T-1} w_{<t-k>_T} Y_k,$$

tj. označíme-li $\widehat{\boldsymbol{m}} = (\widehat{m}(x_0; h), \ \ldots, \widehat{m}(x_{T-1}; h))'$, pak

$$\widehat{\boldsymbol{m}} = \boldsymbol{w} \circledast \boldsymbol{Y}.$$

Důkaz. Rozepíšeme $\widehat{m}(x_t; h)$ podle definice a upravíme

$$\begin{aligned}
\widehat{m}(x_t; h) &= \sum_{k=-T}^{2T-1} W_k(x_t) \widetilde{Y}_k \\
&= \sum_{k=-T}^{-1} W_k(x_t) \widetilde{Y}_k + \sum_{k=0}^{T-1} W_k(x_t) \widetilde{Y}_k + \sum_{k=T}^{2T-1} W_k(x_t) \widetilde{Y}_k \\
&= \sum_{k=0}^{T-1} [W_{k-T}(x_t) + W_k(x_t) + W_{k+T}(x_t)] Y_k.
\end{aligned}$$

Využijeme posledních dvou pomocných tvrzení z předchozího odstavce, tj.

$$\begin{aligned}\widehat{m}(x_t;h) &= \sum_{k=0}^{T-1}[W_k(x_t+1)+W_k(x_t)+W_k(x_t-1)]Y_k \\ &= \sum_{k=0}^{T-1}[W_0(x_{t-k}+1)+W_0(x_{t-k})+W_0(x_{t-k}-1)]Y_k \\ &= \sum_{k=0}^{T-1} w_{<t-k>_T}Y_k.\end{aligned}$$

□

Další věta uvádí, jak vypadá diskrétní Fourierova transformace vektoru $\boldsymbol{w}$.

Věta 4.2.4. *Nechť* $t \in \{0,1,\ldots,T-1\}$, *pak*

$$w_t^- = \sum_{k=-T+1}^{T-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}}.$$

Důkaz. Rozepíšeme t-tou složku vektoru $\boldsymbol{w}^-$ podle definice

$$\begin{aligned}w_t^- &= \sum_{k=0}^{T-1} w_k e^{-\frac{i2\pi kt}{T}} \\ &= \sum_{k=0}^{T-1}[W_0(x_k-1)+W_0(x_k)+W_0(x_k+1)]e^{-\frac{i2\pi kt}{T}}.\end{aligned}$$

Třetí člen $W_0(x_k+1)$ v závorce můžeme vynechat, neboť funkce $W_0(x)$ je nulová vně intervalu $[-h,h],\ h<1$, a proto $W_0(x_k+1)=0$ pro $k=0,\ldots,T-1$.

$$\begin{aligned}w_t^- &= \sum_{k=0}^{T-1}[W_0(x_{k-T})+W_0(x_k)]e^{-\frac{i2\pi kt}{T}} \\ &= \sum_{k=-T}^{-1} W_0(x_k)e^{-\frac{i2\pi(k+T)t}{T}} + \sum_{k=0}^{T-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}} \\ &= \sum_{k=-T+1}^{-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}}\underbrace{e^{-i2\pi t}}_{1} + \sum_{k=0}^{T-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}} \\ &= \sum_{k=-T+1}^{T-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}}.\end{aligned}$$

□

Využitím symetrie funkce $W_0(x)$ a předchozí věty zjistíme zajímavý fakt, že diskrétní Fourierova transformace vektoru $\boldsymbol{w}$ je reálný vektor.

Důsledek 4.2.5. *Pro $t \in \{0, 1, \ldots, T-1\}$ je $w_t^- \in \mathbb{R}$.*

Důkaz. Postupnými úpravami dostáváme

$$\begin{aligned}
w_t^- &= \sum_{k=-T+1}^{T-1} W_0(x_k) e^{-\frac{i2\pi kt}{T}} \\
&= \sum_{k=-T+1}^{-1} W_0(x_k) e^{-\frac{i2\pi kt}{T}} + W_0(x_0) + \sum_{k=1}^{T-1} W_0(x_k) e^{-\frac{i2\pi kt}{T}} \\
&= \sum_{k=1}^{T-1} W_0(x_{-k}) e^{\frac{i2\pi kt}{T}} + W_0(x_0) + \sum_{k=1}^{T-1} W_0(x_k) e^{-\frac{i2\pi kt}{T}} \\
&= \sum_{k=1}^{T-1} W_0(x_k) \left[e^{\frac{i2\pi kt}{T}} + e^{-\frac{i2\pi kt}{T}} \right] + W_0(x_0) \\
&= 2 \sum_{k=0}^{T-1} W_0(x_k) \cos\left(\frac{2\pi kt}{T} \right).
\end{aligned}$$

□

Poznámka. Připomeňme nyní průměrnou střední kvadratickou chybu $R_T(h)$ (12) a residuální součet čtverců $RSS_T(h)$ (16)

$$R_T(h) = \frac{1}{T} E \sum_{t=0}^{T-1} (m(x_t) - \widehat{m}(x_t; h))^2,$$

$$RSS_T(h) = \frac{1}{T} \sum_{t=0}^{T-1} [Y_t - \widehat{m}(x_t; h)]^2.$$

Hlavní myšlenkou všech klasických metod pro hledání optimální šířky okna byla vhodná „úprava“ $RSS_T(h)$. Nyní využijeme teorie Fourierovy analýzy, především Parsevalovy identity, a budeme dále „upravovat“ Fourierovu transformaci $RSS_T(h)$. Jak tato transformace vypadá, charakterizuje následující věta.

Věta 4.2.6. *Nechť $N = \left[\frac{T-1}{2}\right]$, tj. N je celá část výrazu $\frac{T-1}{2}$. Pak*

$$RSS_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} I_{Y_t} \left\{ 1 - w_t^- \right\}^2. \tag{24}$$

Důkaz. Nejprve vyjádříme $RSS_T(h)$ podle definice, místo závorek můžeme psát absolutní hodnotu

$$\begin{aligned} RSS_T(h) &= \frac{1}{T}\sum_{t=0}^{T-1} |Y_t - \widehat{m}(x_t;h)|^2 \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \left| Y_t - \sum_{k=-T}^{2T-1} W_k(x_t)\widetilde{Y}_k \right|^2 . \end{aligned}$$

Jádrový odhad můžeme podle Věty 4.2.3 vyjádřit jako konvoluci vektorů $\boldsymbol{w}$ a $\boldsymbol{Y}$

$$\begin{aligned} RSS_T(h) &= \frac{1}{T}\sum_{t=0}^{T-1} \left| Y_t - \sum_{k=0}^{T-1} w_{<t-k>_T} Y_k \right|^2 \\ &= \frac{1}{T}\sum_{t=0}^{T-1} |Y_t - (\boldsymbol{w} \circledast \boldsymbol{Y})_t|^2 . \end{aligned}$$

Nyní můžeme použít Parsevalovu identitu (Věta 4.2.1) a Větu 4.2.2

$$RSS_T(h) = \frac{1}{T}\sum_{t=0}^{T-1} \left| Y_t^- - w_t^- Y_t^- \right|^2 .$$

Do součtu nemusíme zahrnout člen pro $t = 0$, neboť $w_0^- = \sum_{k=-T+1}^{T-1} W_0(x_k) = 1$. Využijeme též symetrie Fourierových obrazů reálného vektoru popsané v poznámce za definicí 4.1, stačí tedy sčítat pouze po index N a každý člen vzít dvakrát. Přesněji řečeno, tento postup platí pro T liché. Pro T sudé by bylo nutné ještě uvažovat příslušný člen pro $t = T/2$, avšak Fourierovy frekvence jsou v tomto případě celočíselné násobky π, proto je tento výraz zanedbatelný. Celkem tedy

$$\begin{aligned} RSS_T(h) &= \frac{2}{T^2}\sum_{t=1}^{N} \left| Y_t^- - w_t^- Y_t^- \right|^2 \\ &= \frac{2}{T^2}\sum_{t=1}^{N} |Y_t^-|^2 |1 - w_t^-|^2 . \end{aligned}$$

Využitím vztahu (22) pro I_{Y_t} dostáváme

$$RSS_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N} I_{Y_t} \left\{1 - w_t^-\right\}^2 .$$

V posledním výrazu není nutné uvádět absolutní hodnotu, neboť členy w_t^- jsou podle Důsledku 4.2.5 reálné. □

V další větě si všimneme, jak vypadá Fourierova transformace pro funkci $R_T(h)$. K důkazu této věty budeme potřebovat následující dvě pomocná tvrzení.

Lemma 4.2.7. *Nechť $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_{T-1})'$ je vektor chyb v regresním modelu (1) a $\boldsymbol{\varepsilon}^-$ je jeho diskrétní Fourierova transformace. Pak platí*

$$E|\varepsilon_t^-|^2 = T\sigma^2, \qquad t = 0, \ \ldots, T-1.$$

Důkaz. Nejprve vyjádříme ε_t^- podle definice a roznásobíme

$$\begin{aligned} E|\varepsilon_t^-|^2 &= E\left|\sum_{k=0}^{T-1} \varepsilon_k e^{-\frac{i2\pi kt}{T}}\right|^2 \\ &= E\left[\sum_{k=0}^{T-1}\left|\varepsilon_k e^{-\frac{i2\pi kt}{T}}\right|^2 + 2\mathrm{Re}\sum_{\substack{(k,l)\in\{0,\ldots,T-1\}^2 \\ k<l}} \varepsilon_k e^{-\frac{i2\pi kt}{T}}\varepsilon_l e^{\frac{i2\pi lt}{T}}\right]. \end{aligned}$$

Při výpočtu střední hodnoty využijeme vlastností $E(\varepsilon_k) = 0$ pro $k = 0, \ldots, T-1$ a také skutečnosti, že chyby jsou navzájem nezávislé, a proto $E(\varepsilon_k\varepsilon_l) = 0$ pro $k \neq l$

$$E|\varepsilon_t^-|^2 = \sum_{k=0}^{T-1} E(\varepsilon_k^2)\left|e^{-\frac{i2\pi kt}{T}}\right|^2 + 0 = \sum_{k=0}^{T-1} E(\varepsilon_k^2) = T\sigma^2.$$

□

Lemma 4.2.8. *Nechť $\boldsymbol{w}^-$ je diskrétní Fourierova transformace vektoru $\boldsymbol{w}$. Pak platí*

$$\sum_{t=0}^{T-1}\{1 - w_t^-\}^2 = T - 2Tw_0 + \sum_{t=0}^{T-1}(w_t^-)^2.$$

Důkaz. Nejprve umocníme závorku na druhou a dostáváme

$$\begin{aligned} \sum_{t=0}^{T-1}\{1 - w_t^-\}^2 &= \sum_{t=0}^{T-1}\{1 - 2w_t^- + (w_t^-)^2\} \\ &= T - 2\sum_{t=0}^{T-1} w_t^- + \sum_{t=0}^{T-1}(w_t^-)^2. \end{aligned}$$

Nyní stačí ukázat, že $\sum_{t=0}^{T-1} w_t^- = TW_0(0)$,

$$\sum_{t=0}^{T-1} w_t^- = \sum_{t=0}^{T-1}\sum_{k=-T+1}^{T-1} W_0(x_k)e^{-\frac{i2\pi kt}{T}} = \sum_{k=-T+1}^{T-1} W_0(x_k)\sum_{t=0}^{T-1} e^{-\frac{i2\pi kt}{T}}.$$

Spočtěme nyní sumu $\sum_{t=0}^{T-1} e^{-\frac{i2\pi kt}{T}}$ podle vzorce pro součet prvních T členů geometrické posloupnosti $\sum_{t=0}^{T-1} a_1 q^t = a_1 \frac{q^T-1}{q-1}$, $q \neq 1$, kde a_1 je první člen posloupnosti. V našem případě je $a_1 = e^{\frac{i2\pi k0}{T}} = 1$, $q = e^{\frac{i2\pi k}{T}}$ a tedy

$$\sum_{t=0}^{T-1} e^{-\frac{i2\pi kt}{T}} = \begin{cases} \frac{e^{-\frac{i2\pi kT}{T}}-1}{e^{-\frac{i2\pi k}{T}}-1} = 0, & \text{pro } k \neq 0 \\ T, & \text{pro } k = 0. \end{cases}$$

Odtud již plyne tvrzení, neboť zřejmě $W_0(0) = w_0$. □

Věta 4.2.9. *Nechť $N = \left[\frac{T-1}{2}\right]$, tj. N je celá část výrazu $\frac{T-1}{2}$. Pak*

$$R_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} (I_{S_t} + \frac{\sigma^2}{2\pi})\{1 - w_t^-\}^2 - \sigma^2 + 2\sigma^2 w_0. \tag{25}$$

Důkaz. Opět využijeme Parsevalovy identity

$$R_T(h) = \frac{1}{T} E \sum_{t=0}^{T-1} (m(x_t) - \widehat{m}(x_t; h))^2 = \frac{1}{T^2} E \sum_{t=0}^{T-1} |S_t^- - w_t^- Y_t^-|^2.$$

Dosazením za $Y_t^- = S_t^- + \varepsilon_t^-$ a roznásobením dostáváme

$$\begin{aligned} R_T(h) &= \frac{1}{T^2} E \sum_{t=0}^{T-1} |S_t^- - w_t^-(S_t^- + \varepsilon_t^-)|^2 \\ &= \frac{1}{T^2} E \sum_{t=0}^{T-1} |S_t^-(1 - w_t^-) - w_t^- \varepsilon_t^-|^2 \\ &= \frac{1}{T^2} E \sum_{t=0}^{T-1} \left[|S_t^-|^2 |1 - w_t^-|^2 + |w_t^-|^2 |\varepsilon_t^-|^2 - 2\mathrm{Re}\left\{ S_t^-(1 - w_t^-)\overline{w_t^- \varepsilon_t^-} \right\} \right] \\ &= \frac{2}{T^2} \sum_{t=1}^{N} |S_t^-|^2 \{1 - w_t^-\}^2 + \frac{1}{T^2} \sum_{t=0}^{T-1} (w_t^-)^2 E|\varepsilon_t^-|^2 \\ &\quad - \frac{2}{T^2} \sum_{t=0}^{T-1} \mathrm{Re}\left\{ S_t^-(1 - w_t^-) w_{-t}^- E\varepsilon_{-t}^- \right\}. \end{aligned}$$

U druhého členu tohoto součtu můžeme při výpočtu střední hodnoty užít Lemma 4.2.7. Třetí člen je nulový, neboť zřejmě $E\varepsilon_{-t}^- = 0$, a tedy

$$R_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} I_{S_t} \{1 - w_t^-\}^2 + \frac{\sigma^2}{T} \sum_{t=0}^{T-1} (w_t^-)^2.$$

Nyní využijeme Lemma 4.2.8 a upravíme poslední člen

$$R_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N} I_{S_t}\{1-w_t^-\}^2 + \frac{\sigma^2}{T}\sum_{t=0}^{T-1}\{1-w_t^-\}^2 - \sigma^2 + 2\sigma^2 w_0$$
$$= \frac{4\pi}{T}\sum_{t=1}^{N}(I_{S_t} + \frac{\sigma^2}{2\pi})\{1-w_t^-\}^2 - \sigma^2 + 2\sigma^2 w_0.$$

□

4.3 Metoda Fourierovy transformace

V tomto odstavci popíšeme metodu pro hledání optimální šířky okna, která využívá Fourierovu transformaci. Touto myšlenkou se zabývá Chiu ve svém článku [10], kde uvažuje podobný postup pro speciální třídu váhových funkcí. Zde bude tato metoda zobecněna na třídu $S_{0\kappa}$, κ sudé a následně i aplikována v dalším odstavci. Chiu se také zaměřil pouze na Pristleyho – Chaovy odhady, kdežto my budeme brát v úvahu všechny zmiňované typy.

Nyní podrobněji popíšeme tuto metodu a poté se pokusíme ukázat motivaci pro uvedený postup. Hlavní myšlenkou bude opět „upravit“ nějakým způsobem residuální součet čtverců $RSS_T(h)$ tak, aby jeho minimum bylo co nejlepším odhadem teoretické optimální hodnoty h_{opt}. Nebudeme však brát v úvahu původní chybovou funkci $RSS_T(h)$, nýbrž její ekvivalentní vyjádření pomocí Fourierovy transformace (24)

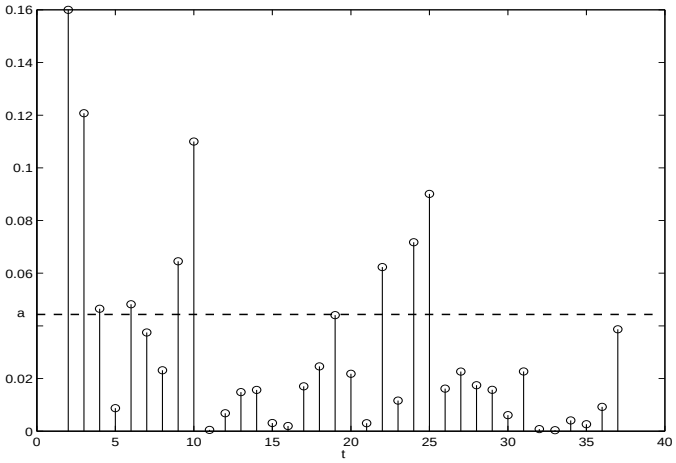
$$RSS_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N} I_{Y_t}\left\{1-w_t^-\right\}^2.$$

Dále budeme potřebovat odhad neznámého parametru σ^2. Použijeme odhadu, který navrhl Rice [17]

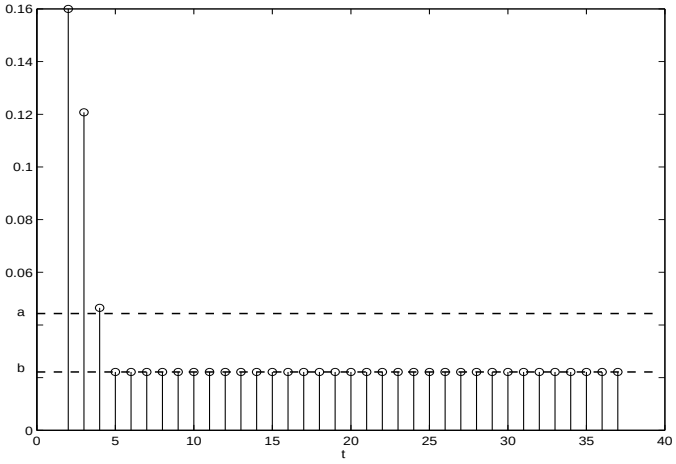
$$\hat{\sigma}^2 = \frac{1}{2T-2}\sum_{t=1}^{T-1}(Y_t - Y_{t-1})^2. \tag{26}$$

Hlavní idea navrhované metody spočívá ve vhodné úpravě periodogramu $\boldsymbol{I}_Y$. Najdeme nejmenší index $J_1 \in \{2,\dots N\}$ takový, že $I_{Y_{J_1}} < c\,\hat{\sigma}^2/2\pi$ pro vhodnou konstantu $c > 1$. Periodogram $\boldsymbol{I}_Y$ pozměníme tak, že všechny jeho složky I_{Y_t} pro $t \geq J_1$ položíme rovny $\hat{\sigma}^2/2\pi$. Situace je znázorněna na obr.17 a obr.18. Takto upravený periodogram označíme $\widetilde{\boldsymbol{I}}_Y$, tj.

$$\widetilde{I}_{Y_t} = \begin{cases} I_{Y_t}, & t < J_1 \\ \hat{\sigma}^2/2\pi, & t \geq J_1 \end{cases}.$$



Obrázek 17: *Složky periodogramu* $\boldsymbol{I}_Y$ *v závislosti na* t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$.



Obrázek 18: *Složky pozměněného periodogramu* $\widetilde{\boldsymbol{I}}_Y$ *v závislosti na* t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$, $b = \frac{\hat{\sigma}^2}{2\pi}$.

Při všech simulacích, které byly v souvislosti s touto metodou provedeny, poskytovala žádoucí výsledky volba $1 < c < 3$. Dostáváme tak modifikovaný residuální součet čtverců $MRSS_T(h)$

$$MRSS_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} \widetilde{I}_{Y_t} \{1 - w_t^-\}^2. \tag{27}$$

Takto upravenou funkci $RSS_T(h)$ můžeme použít v některé z klasických metod pro odhad optimální šířky okna. Protože známe odhad parametru $\hat{\sigma}^2$, nabízí se zde možnost použití chybové funkce $\widehat{R}_T(h)$, která je uvedena v kapitole 3.1 – vztah (18). Uveďme její tvar v cyklickém regresním modelu

$$\widehat{R}_T(h) = RSS_T(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0. \tag{28}$$

Nahrazením residuálního součtu čtverců $RSS_T(h)$ funkcí (27) vznikla nová chybová funkce, označme ji $\widetilde{R}_T(h)$

$$\widetilde{R}_T(h) = MRSS_T(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0. \tag{29}$$

Metoda Fourierovy transformace spočívá v minimalizaci této funkce, tj.

$$\hat{h}_{opt} = \arg\min_{h\in(0,1)} \widetilde{R}_T(h).$$

Pokusme se nyní vysvětlit důvod pro výše uvedené obměny funkce $RSS_T(h)$. U klasických metod pro odhad optimální šířky okna se často stává, že vedou k menším hodnotám než je skutečná optimální šířka okna. Tento jev je způsoben tím, že chybové funkce, které se minimalizují, jsou jen asymptoticky nevychýlenými odhady funkce $R_T(h)$. Pro lepší názornost budeme brát v úvahu nejprve funkci $\widehat{R}_T(h)$ jako odhad $R_T(h)$. Položme

$$D(h) = \widehat{R}_T(h) - R_T(h).$$

Dosazením podle (24) a (25) dostáváme

$$D(h) = \frac{4\pi}{T} \sum_{t=1}^{N} \left\{ I_{Y_t} - I_{S_t} - \frac{\sigma^2}{2\pi} \right\} \{1 - w_t^-\}^2. \tag{30}$$

Podobným způsobem budeme definovat funkci $\widetilde{D}(h)$ pro $\widetilde{R}_T(h)$

$$\widetilde{D}(h) = \widetilde{R}_T(h) - R_T(h)$$

a opět ji můžeme vyjádřit dle (29) a (25) pomocí periodogramů

$$\widetilde{D}(h) = \frac{4\pi}{T} \sum_{t=1}^{J_1-1} \left\{ I_{Y_t} - I_{S_t} - \frac{\sigma^2}{2\pi} \right\} \{1 - w_t^-\}^2 - \frac{4\pi}{T} \sum_{t=J_1}^{N} I_{S_t} \{1 - w_t^-\}^2. \tag{31}$$

Pro dostatečně hladkou regresní funkci, což je předpoklad našich úvah, periodogram I_{S_t} klesá rychle k nule s rostoucím t. Velikost indexu J_1 závisí do jisté míry na konstantě c, která určuje práh pro úpravu periodogramu $\boldsymbol{I}_Y$. V simulacích i praktických příkladech dávala dobré výsledky volba $c \in (1, 3)$. S takto nastavenou hodnotou c je druhý člen v (31) zanedbatelný. Porovnáním (30) a (31) je vidět, že $|\widetilde{D}(h)| \leq |D(h)|$, tj. navrhovaná funkce $\widetilde{R}_T(h)$ má menší odchylku od skutečné průměrné střední kvadratické chyby $R_T(h)$ než funkce $\widehat{R}_T(h)$.

Příklad. Na simulovaných datech v systému MATLAB jsme porovnávali odhady získané Riceho chybovou funkcí $\widehat{R}_T(h)$ a metodou Fourierovy transformace $\widetilde{R}_T(h)$, jak vzájemně, tak s teoretickou optimální šířkou okna. Pozorování Y_t, pro $t = 0, \ldots, T-1 = 99$, byla vygenerována s náhodnými normálně rozloženými chybami s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.2$. Regresní funkce byla v našem případě

$$m(x) = \sin(2\pi x).$$

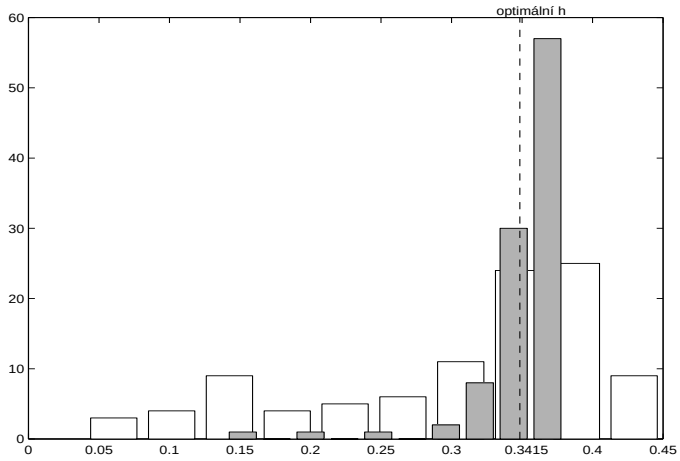
Při výpočtech jsme použili Nadarayovy – Watsonovy estimátory a jádro třídy S_{04} (viz tab.1). Teoretická optimální hodnota spočítaná dle vzorce (14) je v tomto případě $h_{opt} = 0.3415$. Bylo vygenerováno 200 řad. U každé řady byly získány odhady optimální šířky okna tak, že jsme spočítali hodnoty funkce $\widehat{R}_T(h)$, případně $\widetilde{R}_T(h)$, v 321 bodech ekvidistantně rozložených na intervalu $[0.01, 0.99]$ a pak z nich vybrali tu hodnotu, pro kterou tato funkce nabývala svého minima. V tabulce 4 jsou uvedeny střední hodnoty všech odhadů a také směrodatné odchylky. V prvním sloupci je označení chybových funkcí, které byly použity, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot, $std(\hat{h}_{opt})$ je jejich směrodatná odchylka.

Z hodnot uvedených v tabulce je zřejmé, že odhady získané naší navrhovanou metodou

Tabulka 4: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných minimalizací funkcí $\widehat{R}_T(h)$ a $\widetilde{R}_T(h)$. Teoretická optimální šířka okna $h_{opt} = 0.3415$.*

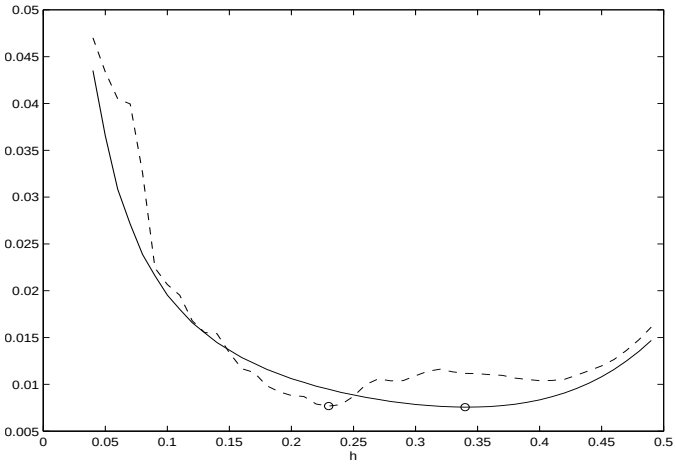
Metoda	$E(\hat{h})$	$std(\hat{h})$
$\widehat{R}_T$	0.3041	0.0903
$\widetilde{R}_T$	0.3475	0.0289

jsou v tomto případě blíže teoretické optimální šířce okna. Také směrodatná odchylka byla jen třetinová, což poukazuje na menší variabilitu odhadů. Rozložení všech výsledků znázorňují histogramy na obr.19.



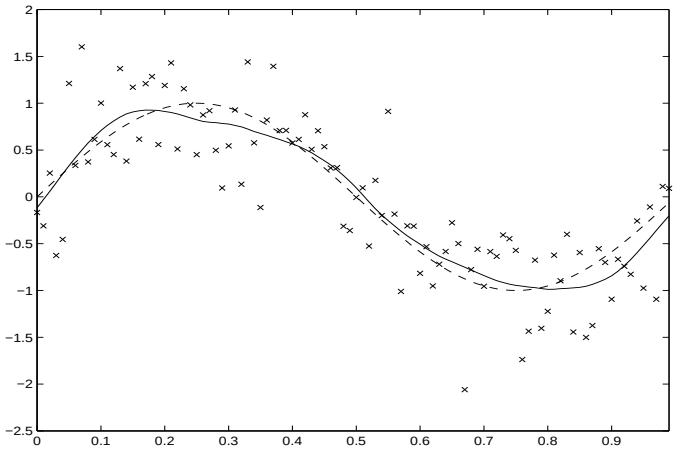
Obrázek 19: *Rozložení všech výsledků získaných minimalizací funkce $\widehat{R}_T$ (bílá) a funkce $\widetilde{R}_T$ (šedá).*

Zvolme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Na obr.20 jsou znázorněny obě chybové funkce $\widehat{R}_T(h)$, $\widetilde{R}_T(h)$ a jejich minima. Na obr.21

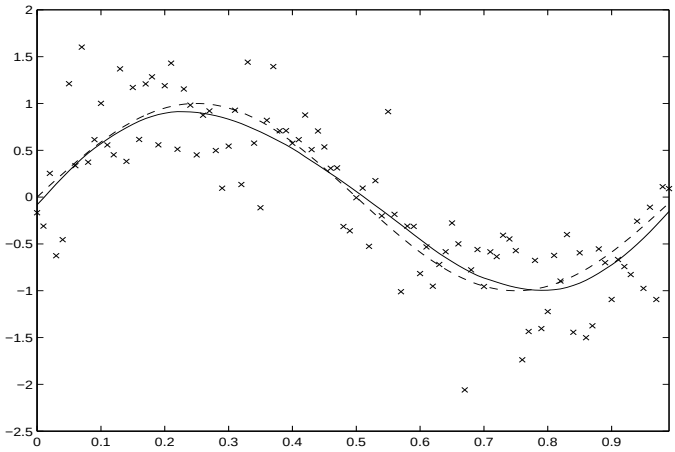


Obrázek 20: *Průběh obou chybových funkcí v závislosti na h a jejich minima. Čárkovaně je zobrazena funkce* $\widehat{R}_T(h)$. *Plná čára znázorňuje navrhovanou funkci* $\widetilde{R}_T(h)$.

jsou zobrazena simulovaná data, regresní funkce $m(x)$ a její odhad s parametrem $\hat{h}_{opt} = 0.2361$, který byl v tomto případě nalezen minimalizací funkce $\widehat{R}_T(h)$. Je zřejmé, že výsledný odhad je mírně podhlazený. Obr.22 znázorňuje jádrový odhad s šířkou okna $\hat{h}_{opt} = 0.3416$, která byla nalezena užitím funkce $\widetilde{R}_T(h)$.



Obrázek 21: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2361$.



Obrázek 22: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.3416$.

4.4 Plug-in metoda

Tato metoda je založena na minimalizaci průměrné střední kvadratické chyby $R_T(h)$, resp. jejího hlavního členu (13). Teoreticky je možné vyjádřit přímo minimum této funkce, tj. optimální hodnotu šířky vyhlazovacího okna (14). Ta ovšem závisí na neznámých parametrech σ^2 a $m^{(\kappa)}(x)$. Plug-in metoda se zabývá odhadem těchto parametrů. Tento postup je výhodný především v tom, že odpadá problém minimalizace nějaké funkce, protože hodnota minima je již teoreticky odvozena. Na druhé straně, odhad κ-té derivace neznámé regresní funkce $m(x)$ se zde zdá být problematický, neboť cílem jádrového vyhlazování je odhadnout právě tuto funkci. Využijeme k tomu poznatků získaných v minulých odstavcích, zejména budeme opět předpokládat cyklický model a aplikovat výsledky z teorie Fourierovy analýzy.

Nechť K je jádro třídy $S_{0\kappa}$. Nejprve připomeňme hlavní člen průměrné střední kvadratické chyby (13), str.25

$$\overline{R_T}(h) = \frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa,$$

kde

$$V(K) = \int_{-1}^{1} K^2(x)dx, \quad \beta_\kappa = \int_{-1}^{1} x^\kappa K(x)dx, \quad A_\kappa = \int_0^1 \left(m^{(\kappa)}(x)\right)^2 dx.$$

Teoretická hodnota minima funkce $\overline{R_T}(h)$ je optimální šířka okna (14)

$$h_{opt} = \left(\frac{\sigma^2 V(K)(\kappa!)^2}{2\kappa T \beta_\kappa^2 A_\kappa}\right)^{\frac{1}{2\kappa+1}}.$$

Naším cílem bude odhad neznámých parametrů σ^2 a A_κ.

V případě neznámého rozptylu σ^2 lze uvažovat odhad $\hat{\sigma}^2$ (26), který používá Rice v [17], a který jsme aplikovali v minulém odstavci

$$\hat{\sigma}^2 = \frac{1}{2T-2}\sum_{t=1}^{T-1}(Y_t - Y_{t-1})^2.$$

Chceme-li odhadnout parametr A_κ, situace je složitější a budeme potřebovat několik pomocných tvrzení, která uvedeme dále.

Nyní se vrátíme k odstavci 4.3. Připomeňme funkci $\widetilde{R}_T(h)$ danou vztahem (29), kterou jsme v daném případě uvedli jako další možnou metodu pro hledání optimální šířky okna. Při jejím odvozování jsme definovali index J_1, který částečně závisel na jisté konstantě c. V simulacích i praktických příkladech dávala dobré výsledky volba $c \in (1,3)$. Pro lepší přehlednost v dalším textu položme $c = 2$ pevné. Následující věta uvádí poněkud pozměněný tvar funkce $\widetilde{R}_T(h)$.

Věta 4.4.1. *Nechť J_1 je nejmenší index takový, že $I_{Y_{J_1}} < \hat{\sigma}^2/\pi T$. Pak*

$$\widetilde{R}_T(h) = \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2.$$

Důkaz. Napíšeme funkci $\widetilde{R}_T(h)$ dle vztahu (29) a upravíme

$$\begin{aligned}
\widetilde{R}_T(h) &= \frac{4\pi}{T}\sum_{t=1}^{N}\tilde{I}_{Y_t}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1} I_{Y_t}\{1-w_t^-\}^2 + \frac{4\pi}{T}\sum_{t=J_1}^{N}\frac{\hat{\sigma}^2}{2\pi}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0.
\end{aligned}$$

Pro druhý člen tohoto součtu můžeme aplikovat Lemma 4.2.8, roznásobením obdržíme požadovaný tvar

$$\begin{aligned}
\widetilde{R}_T(h) &= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\left(T - 2Tw_0 + \sum_{t=0}^{T-1}(w_t^-)^2\right) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2.
\end{aligned}$$

□

Lemma 4.4.2. *Nechť $t \in \{0,\dots,T-1\}$. Pak pro všechny typy odhadů platí*

$$W_0(x_t) = \frac{1}{T}K_h(x_t) + O(T^{-2}).$$

Důkaz. Pro Pristleyho – Chaovy váhy platí dokonce rovnost, neboť

$$W_0(x_t) = \frac{1}{T}K_h(0 - x_t) = \frac{1}{T}K_h(x_t).$$

V případě Nadarayových – Watsonových a lokálně lineárních odhadů jsou hodnoty vah v bodech plánu stejné (Důsledek 4.1.3). Můžeme tedy psát

$$W_0(x_t) = \frac{1}{TC_T}K_h(x_t),$$

kde

$$C_T = \frac{1}{T}\sum_{k=-T+1}^{T-1} K_h(x_k).$$

Podobným postupem jako v důkazu Lemmatu 2.2.2 by se dalo ukázat, že

$$C_T = \int_{-1}^{1} K(x)dx + O(T^{-1}) = 1 + O(T^{-1}),$$

po dosazení dostáváme

$$W_0(x_t) = \frac{1}{T(1 + O(T^{-1}))}K_h(x_t) = \frac{1}{T}K_h(x_t) + O(T^{-2}).$$

Nakonec pak zbývá dokázat tvrzení pro Gasserovy – Müllerovy estimátory. Podle definice je

$$W_0(x_t) = \int_{s_{-1}}^{s_0} K_h(t - x_t)dt = \int_{s_{-1}-x_t}^{s_0-x_t} K_h(u)du,$$

kde $s_k = \frac{x_k + x_{k+1}}{2}$. Uvedený integrál vypočteme obdélníkovým pravidlem

$$\begin{aligned}
\int_{s_{-1}-x_t}^{s_0-x_t} K_h(u)du &= (s_0 - x_t - (s_{-1} - x_t))K_h\left(\frac{s_{-1} - x_t + s_0 - x_t}{2}\right) + O(T^{-2}) \\
&= \left(\frac{x_0 + x_1 - (x_{-1} + x_0)}{2}\right)K_h\left(\frac{x_{-1} + x_0 + x_0 + x_1 - 2x_t}{2}\right) + O(T^{-2}) \\
&= \frac{1}{T}K_h(x_t) + O(T^{-2}).
\end{aligned}$$

□

Věta 4.4.3. *Nechť $\boldsymbol{w}^-$ je diskrétní Fourierova transformace vektoru $\boldsymbol{w}$. Pak platí*

$$\sum_{t=0}^{T-1}(w_t^-)^2 = \frac{1}{h}V(K) + O(T^{-1}). \tag{32}$$

Důkaz. Dosazením dle definice w_t^- a postupnými úpravami dostáváme

$$\begin{aligned}\sum_{t=0}^{T-1}(w_t^-)^2 &= \sum_{t=0}^{T-1}|w_t^-|^2 = \sum_{t=0}^{T-1} w_t^- \overline{w_t^-} \\ &= \sum_{t=0}^{T-1}\sum_{j=-T+1}^{T-1} W_0(x_j)e^{-\frac{i2\pi jt}{T}} \sum_{k=-T+1}^{T-1} W_0(x_k)e^{\frac{i2\pi kt}{T}} \\ &= \sum_{t=0}^{T-1}\sum_{j=-T+1}^{T-1}\sum_{k=-T+1}^{T-1} W_0(x_j)W_0(x_k)e^{\frac{i2\pi(k-j)t}{T}} \\ &= \sum_{j=-T+1}^{T-1}\sum_{k=-T+1}^{T-1} W_0(x_j)W_0(x_k)\sum_{t=0}^{T-1} e^{\frac{i2\pi(k-j)t}{T}}.\end{aligned}$$

Spočtěme nyní sumu $\sum_{t=0}^{T-1} e^{\frac{i2\pi(k-j)t}{T}}$ podle vzorce pro součet prvních T členů geometrické posloupnosti. Analogicky jako v důkazu Lemmatu 4.2.8 dostáváme

$$\sum_{t=0}^{T-1} e^{\frac{i2\pi(k-j)t}{T}} = \begin{cases} \frac{e^{\frac{i2\pi(k-j)T}{T}}-1}{e^{\frac{i2\pi(k-j)}{T}}-1} = 0, & \text{pro } k \neq j \\ T, & \text{pro } k = j. \end{cases}$$

Odtud

$$\begin{aligned}\sum_{t=0}^{T-1}(w_t^-)^2 &= T\sum_{k=-T+1}^{T-1} W_0^2(x_k) = T\sum_{k=-T+1}^{T-1}\left[\frac{1}{T}K_h(x_k)\right]^2 + O(T^{-2}) \\ &= \sum_{k=-T+1}^{T-1}\frac{1}{T}K_h^2(x_k) + O(T^{-2}).\end{aligned}$$

Stejným způsobem jako v důkazu Lemmatu 2.2.2 můžeme nahradit sumu integrálem, tj.

$$\sum_{t=0}^{T-1}(w_t^-)^2 = \int_{-1}^{1} K_h^2(u)du + O(T^{-1}) = \frac{1}{h}\int_{-1}^{1} K^2(x)dx + O(T^{-1}).$$

□

Poznámka. Připomeňme nyní Taylorův rozvoj funkce $\cos x$ na intervalu $(-h, h)$ se středem v bodě $x_0 = 0$, který budeme potřebovat v dalších úvahách

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n\frac{x^{2n}}{(2n)!} + R(x),$$

kde

$$R(x) = (-1)^{n+1}\frac{x^{2n+1}}{(2n+1)!}\sin\xi, \qquad \xi \in (-h, h)$$

je chyba aproximace. Nechť $\kappa \in \mathbb{N}$ je sudé, pak

$$1 - \cos x = \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots + (-1)^{\frac{\kappa}{2}+1} \frac{x^\kappa}{(\kappa)!} + R(x),$$

kde

$$R(x) = (-1)^{\frac{\kappa}{2}} \frac{x^{\kappa+1}}{(\kappa+1)!} \sin \xi, \qquad \xi \in (-h, h).$$

Odtud

$$1 - \cos(2\pi jx) = \frac{(2\pi jx)^2}{2!} - \frac{(2\pi jx)^4}{4!} + \frac{(2\pi jx)^6}{6!} + \cdots + (-1)^{\frac{\kappa}{2}+1} \frac{(2\pi jx)^\kappa}{\kappa!} + R(2\pi jx),$$

odhad chyby

$$|R(2\pi jx)| \leq \frac{(2\pi jx)^{\kappa+1}}{(\kappa+1)!}.$$

Otázkou je, jak volit index j, aby aproximace byla dobrá, tj. aby $|R(2\pi jx)| \leq \varepsilon$ pro pevně zvolené $\varepsilon > 0$. Odhad se provádí se středem v bodě $x_0 = 0$ pro všechna x v intervalu $(-h, h)$, takže

$$|R(2\pi jx)| \leq \frac{(2\pi jh)^{\kappa+1}}{(\kappa+1)!} \leq \varepsilon \Rightarrow j \leq \frac{\sqrt[\kappa+1]{\varepsilon(\kappa+1)!}}{2\pi h}.$$

Vzniká zde však další problém. Jak zvolit ε a také jak aproximovat h? Při různých simulacích se ukázala jako dobrá volba $\varepsilon = 10^{-3}$ a pro parametr h jeho hrubý dolní odhad $\frac{\kappa}{T}$. Označme J_2 největší takový index, pro který platí výše uvedená nerovnost, tj.

$$J_2 \leq \frac{\sqrt[\kappa+1]{\varepsilon(\kappa+1)!}}{2\pi h}.$$

Připomeňme také index J_1, který je definován ve Větě 4.4.1

„J_1 je nejmenší index takový, že $I_{Y_{J_1}} < \hat{\sigma}^2/\pi T$“.

V dalších úvahách budeme požadovat, aby byly splněny obě podmínky pro indexy současně. Proto budeme definovat index J

$$J = \min\{J_1, J_2 + 1\}. \tag{33}$$

Věta 4.4.4. *Nechť $\kappa \in \mathbb{N}$ je sudé a J je index definovaný vztahem (33). Pak pro všechna $j \in \mathbb{N}$, $1 \leq j \leq J - 1$, platí*

$$\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) \approx (-1)^{\frac{\kappa}{2}+1} \frac{h^\kappa}{\kappa!} \beta_\kappa. \tag{34}$$

Důkaz. Nejprve vyjádříme w_t^- podle Důsledku 4.2.5. Poté můžeme podle Lemmatu 4.4.2 nahradit výraz $W_0(x_t)$ výrazem $\frac{1}{T}K_h(x_t) + O(T^{-1})$

$$\begin{aligned}\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) &= \frac{1}{(2\pi j)^\kappa}\left(1 - 2\sum_{t=0}^{T-1} W_0(x_t)\cos(\frac{2\pi tj}{T})\right) \\ &= \frac{1}{(2\pi j)^\kappa}\left(1 - 2\sum_{t=0}^{T-1} \frac{1}{T}K_h(x_t)\cos(\frac{2\pi tj}{T})\right) + O(T^{-1}).\end{aligned}$$

Stejným postupem jako v důkazu Lemmatu 2.2.2 se dá ukázat, že $\sum_{t=0}^{T-1} \frac{1}{T}K_h(x_t)\cos(\frac{2\pi tj}{T}) = \int_0^1 K_h(u)\cos(2\pi ju)du + O(T^{-1})$. Dále také využijeme toho, že pro jádra třídy $S_{0\kappa}$ platí $\int_{-1}^1 K_h(u)du = 1$

$$\begin{aligned}\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) &= \frac{1}{(2\pi j)^\kappa}\left(1 - 2\int_0^1 K_h(u)\cos(2\pi ju)du\right) + O(T^{-1}) \\ &= \frac{1}{(2\pi j)^\kappa}\left(\int_{-1}^1 K_h(u)du - \int_{-1}^1 K_h(u)\cos(2\pi ju)du\right) + O(T^{-1}) \\ &= \frac{1}{(2\pi j)^\kappa}\int_{-1}^1 [1 - \cos(2\pi ju)]K_h(u)du + O(T^{-1}).\end{aligned}$$

Funkci $1 - \cos(2\pi ju)$ nahradíme Taylorovým rozvojem řádu κ. Označme R_κ chybu této aproximace

$$\begin{aligned}\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) &= \frac{1}{(2\pi j)^\kappa}\int_{-1}^1 \left[\frac{(2\pi ju)^2}{2} - \frac{(2\pi ju)^4}{24} + \cdots + \frac{(-1)^{\frac{\kappa}{2}+1}(2\pi ju)^\kappa}{\kappa!}\right] K_h(u)du \\ &\quad + \frac{R_\kappa}{(2\pi j)^\kappa} + O(T^{-1}) \\ &= \frac{1}{2}\frac{(2\pi j)^2}{(2\pi j)^\kappa}\int_{-1}^1 u^2 K_h(u)du - \frac{1}{24}\frac{(2\pi j)^4}{(2\pi j)^\kappa}\int_{-1}^1 u^4 K_h(u)du + \ldots \\ &\quad \cdots + \frac{(-1)^{\frac{\kappa}{2}+1}(2\pi j)^\kappa}{\kappa!(2\pi j)^\kappa}\int_{-1}^1 u^\kappa K_h(u)du + \frac{R_\kappa}{(2\pi j)^\kappa} + O(T^{-1}).\end{aligned}$$

Nyní využijeme další vlastnosti jader třídy $S_{0\kappa}$, a sice že $\int_{-1}^{1} u^j K_h(u)du = 0$ pro $0 < j < \kappa$ a $\int_{-1}^{1} u^\kappa K_h(u)du \neq 0$. Nakonec provedeme substituci $x = \frac{u}{h}$ a dostáváme výsledek.

$$\begin{aligned}\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) &= \frac{(-1)^{\frac{\kappa}{2}+1}}{\kappa!}\int_{-1}^{1} u^\kappa K_h(u)du + \frac{R_\kappa}{(2\pi j)^\kappa} + O(T^{-1}) \\ &= \frac{(-1)^{\frac{\kappa}{2}+1}}{\kappa!}\int_{-h}^{h} u^\kappa \frac{1}{h}K(\frac{u}{h})du + \frac{R_\kappa}{(2\pi j)^\kappa} + O(T^{-1}) \\ &= (-1)^{\frac{\kappa}{2}+1}\frac{h^\kappa}{\kappa!}\int_{-1}^{1} x^\kappa K(x)dx + \frac{R_\kappa}{(2\pi j)^\kappa} + O(T^{-1}).\end{aligned}$$

Poslední dva členy můžeme zanedbat, neboť $O(T^{-1})$ s rostoucím T konverguje k nule a podle předpokladu pro index j platí $\left|\frac{R_\kappa}{(2\pi j)^\kappa}\right| \leq \frac{\varepsilon}{(2\pi)^\kappa}$ pro libovolné dostatečně malé $\varepsilon > 0$. Můžeme tedy určit přibližné vyjádření pro odhadovaný výraz

$$\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) \approx (-1)^{\frac{\kappa}{2}+1}\frac{h^\kappa}{\kappa!}\beta_\kappa.$$

□

Věta 4.4.5. *Nechť $\kappa \in \mathbb{N}$ je sudé a J je index definovaný vztahem (33). Pak platí*

$$\frac{4\pi}{T}\sum_{j=1}^{J-1}(2\pi j)^{2\kappa}\{I_{Y_j} - \frac{\hat{\sigma}^2}{2\pi}\} \approx A_\kappa.$$

Důkaz. Při odhadu parametru A_κ využijeme poznatků získaných v předchozím odstavci. Definovali jsme tam chybovou funkci $\widetilde{R}_T(h)$ jako odhad teoretické průměrné střední kvadratické chyby $R_T(h)$. Bereme-li v úvahu pouze hlavní člen $\overline{R}_T(h)$ funkce $R_T(h)$, můžeme psát

$$\widetilde{R}_T(h) \approx \frac{\sigma^2}{Th}V(K) + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa.$$

Dále podle Věty 4.4.1 máme

$$\widetilde{R}_T(h) = \frac{\hat{\sigma}^2}{T}\sum_{j=0}^{T-1}\{w_j^-\}^2 + \frac{4\pi}{T}\sum_{j=1}^{J-1}\{I_{Y_j} - \frac{\hat{\sigma}^2}{2\pi}\}\{1 - w_j^-\}^2.$$

Nyní můžeme použít vztahu (32) z Věty 4.4.3

$$\frac{\hat{\sigma}^2}{T}\sum_{j=0}^{T-1}\{w_j^-\}^2 \approx \frac{\sigma^2}{Th}V(K),$$

a tedy

$$\frac{4\pi}{T}\sum_{j=1}^{J-1}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\}\{1-w_j^-\}^2 \approx \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa.$$

Nechť J je index definovaný vztahem (33), pak dle Věty 4.4.4 pro všechna j, $1 \leq j \leq J-1$ platí

$$\frac{1}{(2\pi j)^{2\kappa}}(1-w_j^-)^2 \approx \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2.$$

Odtud

$$\begin{aligned}
\frac{4\pi}{T}\sum_{j=1}^{J-1}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\}\{1-w_j^-\}^2 &= \frac{4\pi}{T}\sum_{j=1}^{J-1}\frac{1}{(2\pi j)^{2\kappa}}\{1-w_j^-\}^2(2\pi j)^{2\kappa}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\} \\
&\approx \frac{4\pi}{T}\sum_{j=1}^{J-1}\frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2(2\pi j)^{2\kappa}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\} \\
&= \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2\frac{4\pi}{T}\sum_{j=1}^{J-1}(2\pi j)^{2\kappa}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\} \\
&\approx \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa.
\end{aligned}$$

Porovnáním posledních dvou výrazů dostáváme odhad A_κ

$$\frac{4\pi}{T}\sum_{j=1}^{J-1}(2\pi j)^{2\kappa}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\} \approx A_\kappa.$$

□

Podle předchozích úvah tedy můžeme definovat odhad hlavního členu (13) průměrné střední kvadratické chyby získaný plug-in metodou

$$Q(h) = \frac{\hat{\sigma}^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2\widehat{A}_\kappa, \tag{35}$$

kde

$$\hat{\sigma}^2 = \frac{1}{2T-2}\sum_{t=1}^{T-1}(Y_t - Y_{t-1})^2, \qquad \widehat{A}_\kappa = \frac{4\pi}{T}\sum_{j=1}^{J-1}(2\pi j)^{2\kappa}\{I_{Y_j}-\frac{\hat{\sigma}^2}{2\pi}\}$$

a index J je definován vztahem (33).
Řešením rovnice $\frac{\partial Q}{\partial h} = 0$ dostáváme minimum této funkce, tj. odhad optimální šířky okna plug-in metodou

$$\hat{h}_{opt} = \left(\frac{\hat{\sigma}^2 V(K)(\kappa!)^2}{2\kappa T\beta_\kappa^2\widehat{A}_\kappa}\right)^{\frac{1}{2\kappa+1}}. \tag{36}$$

Příklad.
Na simulovaných datech v systému MATLAB jsme porovnávali odhady získané metodou křížového ověřování s teoretickou optimální šířkou okna. Pozorování Y_t, pro $t = 0, \ldots, T-1 = 99$, byla vygenerována s náhodnými normálně rozloženými chybami s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.003^2$. Regresní funkce byla v našem případě

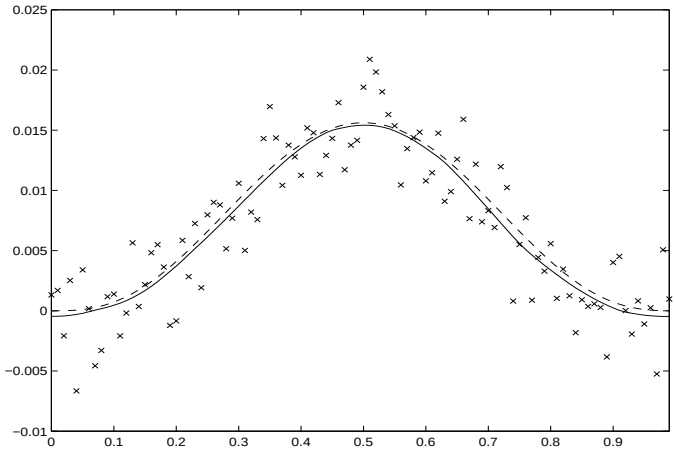
$$m(x) = x^3(1-x)^3.$$

Při výpočtech jsme použili jádra třídy $S_{0\kappa}$ (viz tab.1) pro $\kappa = 2,\ 4,\ 6$ (pro větší κ je $m^{(\kappa)}(x)$ nulová). Bylo vygenerováno 200 řad. U každé řady byly získány odhady optimální šířky plug-in metodou. V tabulce 5 jsou uvedeny střední hodnoty a směrodatné odchylky všech odhadů, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot a $std(\hat{h}_{opt})$ je jejich směrodatná odchylka, h_{opt} označuje teoretickou optimální hodnotu spočtenou dle vzorce (14).

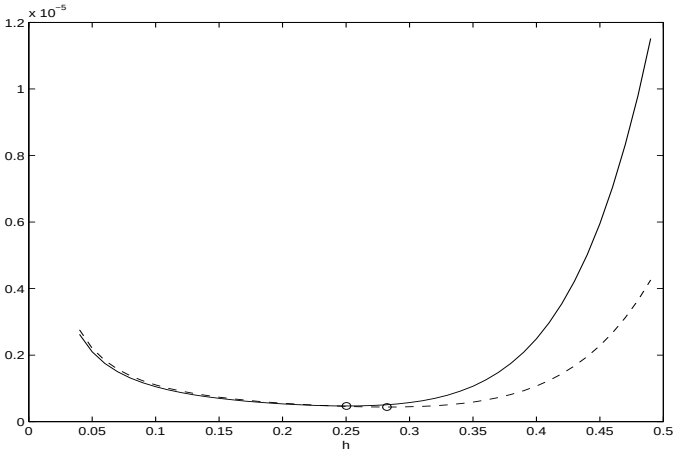
Tabulka 5: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných plug-in metodou.*

κ	2	4	6
h_{opt}	0.1184	0.2521	0.4939
$E(\hat{h}_{opt})$	0.1219	0.2937	0.4884
$std(\hat{h}_{opt})$	0.0049	0.0397	0.0882

Z hodnot uvedených v tabulce je zřejmé, že odhady jsou velmi blízké optimální šířce okna h_{opt}. Zvolíme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Pro odhad regresní funkce jsme použili jádro třídy S_{04}. V tomto případě byla plug-in metodou vybrána optimální šířka okna $\hat{h}_{opt} = 0.2804$. Na obr.23 jsou zobrazena simulovaná data, regresní funkce $m(x)$ a její odhad s tímto parametrem. Obr.24 znázorňuje průběh chybové funkce $R_T(h)$ a jejího odhadu $Q(h)$ získaného plug-in metodou.



Obrázek 23: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.003^2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2804$.



Obrázek 24: *Průběh skutečné chybové funkce* $R_T(h)$ *(plná čára) a jejího odhadu* $Q(h)$ *získaného plug-in metodou (čárkovaně).*

5 Simulace

Cílem této kapitoly je vzájemně porovnat uvedené metody pro odhad optimální šířky okna. Srovnání jsme prováděli na simulovaných datech v systému MATLAB následujícím způsobem:
Nejprve jsme vygenerovali měření s příslušnou regresní funkcí $m(x)$ a rozptylem σ^2 a poté jsme vypočítali odhady optimální šířky okna pomocí jednotlivých metod. Tento postup jsme opakovali 200-krát. Ve všech případech byl rozsah dat $T = 74$ a při výpočtech bylo použito Nadarayových – Watsonových odhadů. Pro každou metodu jsme tedy získali 200 odhadů optimální šířky okna. Z těchto hodnot jsme vyjádřili střední hodnotu a určili směrodatnou odchylku. Protože v tomto případě známe regresní funkci $m(x)$, můžeme zde udat skutečnou teoretickou hodnotu optimální šířky okna a porovnat se středními hodnotami odhadů u jednotlivých metod. Směrodatná odchylka pak popisuje variabilitu srovnávaných metod.
Při simulacích bylo porovnáváno těchto 6 metod:

- Metoda křížového ověřování – minimalizujeme funkci křížového ověřování (19) $CV(h)$, na obrázcích označena symbolem „CV“.
- Metoda penalizačních funkcí – minimalizujeme funkci (20) $G(h)$, jako penalizační funkce $\Xi(u)$ byly vybrány
  - *Rice's bandwidth selector* $\Xi_R(u) = \frac{1}{1-2u}$, na obrázcích označena „Rice-pen“.
  - *ET bandwidth selector* $\Xi_{ET}(u) = e^{\frac{4}{\pi}\tan\frac{\pi}{2}u}$, označena „ET-pen“.
- Riceho chybová funkce (18) $\widehat{R}_T(h)$, pro lepší přehlednost označena „Mallows“ podle původního autora.
- Metoda Fourierovy transformace – minimalizujeme funkci (29) $\widetilde{R}_T(h)$, na obrázcích je tato metoda označena jako „Fourier“.
- Plug-in metoda – minimalizujeme funkci (35) $Q(h)$, její minimum je dáno vztahem (36). Tato metoda je označena symbolem „plug-in“.

Střední hodnoty a směrodatné odchylky pro výše uvedené metody jsou zaznamenány v tabulkách. Na obrázcích jsou pomocí histogramů znázorněna rozdělení všech 200 odhadů optimální šířky okna získaných jednotlivými metodami. Svislá čára představuje teoretickou hodnotu optimální šířky okna.
Nakonec jsme vybrali jednu z vygenerovaných řad a zobrazili jádrové odhady regresní funkce s použitím vyhlazovacích parametrů nalezených jednotlivými metodami. Na obrázcích jsou znázorněny také bodové intervaly spolehlivosti dané vztahem

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

kde $u_{1-\frac{\alpha}{2}}$ je α-kvantil standardizovaného normálního rozložení a pro rozptyl $\sigma^2(x)$ v bodě x platí

$$\sigma^2(x) = \sum_{i=0}^{T-1} W_i(x)(Y_i - m(x;h))^2.$$

Konstrukce intervalů spolehlivosti je podrobněji popsána např. v [6].

5.1 Simulace 1

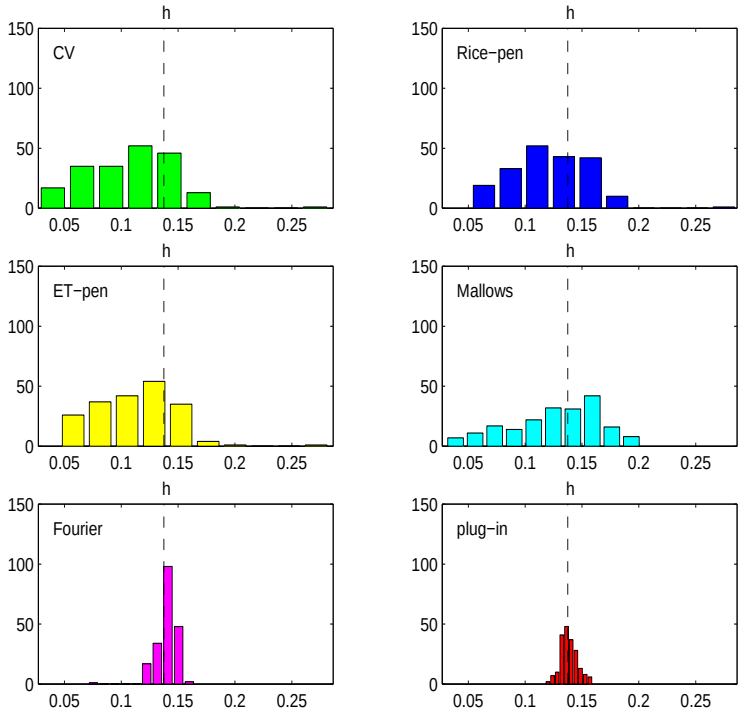
V tomto případě byla generována data s regresní funkcí

$$m(x) = \sin(2\pi x).$$

Chyby měření měly normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.2$. V tabulce 6 jsou zapsány střední hodnoty všech odhadů a také směrodatné odchylky. V prvním sloupci je označení chybových funkcí, které byly použity, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot, $std(\hat{h}_{opt})$ je jejich směrodatná odchylka. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$. Na obr.25 je znázorněno rozložení všech 200 výsledků získaných výše uvedenými metodami pro případ $\kappa = 2$. Pro $\kappa = 4, 6$ je rozložení podobné.
Z tabulky i z obrázku je zřejmé, že nejlepších výsledků bylo dosaženo posledními dvěma metodami, tj. metodou Fourierovy transformace a plug-in metodou. U ostatních metod byly často získány menší hodnoty vyhlazovacího parametru než je jeho teoretická optimální hodnota. Také rozptyl všech výsledků je několikanásobně větší, což do jisté míry ukazuje na menší stabilitu těchto metod.

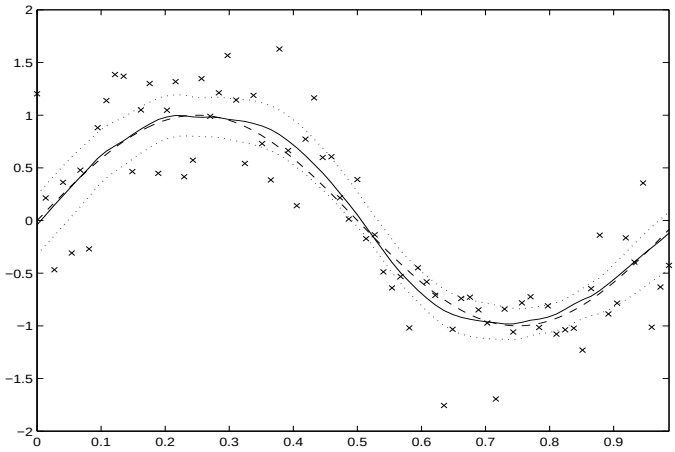
Tabulka 6: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných pomocí výše uvedených metod, $m(x) = \sin(2\pi x)$.*

	$\kappa = 2$ $h_{opt} = 0.1374$		$\kappa = 4$ $h_{opt} = 0.3521$		$\kappa = 6$ $h_{opt} = 0.5783$	
	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$
CV	0.1063	0.0391	0.2232	0.0712	0.3273	0.1056
Rice-pen	0.1222	0.0329	0.2493	0.0585	0.3691	0.0877
ET-pen	0.1114	0.0342	0.2312	0.0625	0.3397	0.0915
Mallows	0.1269	0.0402	0.3354	0.0938	0.4432	0.1078
Fourier	0.1409	0.0095	0.3625	0.0306	0.4967	0.0172
plug-in	0.1383	0.0074	0.3422	0.0348	0.5604	0.0623

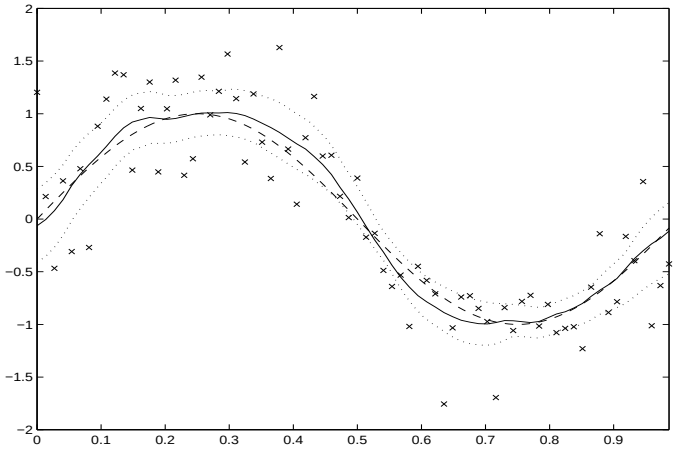


Obrázek 25: *Rozložení všech 200 výsledků získaných pomocí výše uvedených metod pro případ* $\kappa = 2$. *Svislá čára znázorňuje teoretickou hodnotu optimální šířky okna.*

Zvolme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Při výpočtech jsme použili jádra třídy S_{02} – viz str.8. Na obr.26 jsou zobrazena simulovaná data, regresní funkce $m(x)$ a její odhad s vyhlazovacím parametrem $\hat{h}_{opt} = 0.1323$, který byl v tomto případě nalezen plug-in metodou. Hodnota tohoto parametru byla nejblíže teoretické optimální šířce okna $h_{opt} = 0.1374$ a i z obrázku je patrné, že odhad regresní funkce je uspokojivý. Obr.27 znázorňuje jádrový odhad s šířkou okna $\hat{h}_{opt} = 0.0976$, která byla nalezena užitím metody křížového ověřování. Je zřejmé, že výsledný odhad je mírně podhlazený.



Obrázek 26: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1323$ *pro jádro* $K \in S_{02}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.



Obrázek 27: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.0976$ *pro jádro* $K \in S_{02}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.

5.2 Simulace 2

Ve druhém případě byla generována data s regresní funkcí

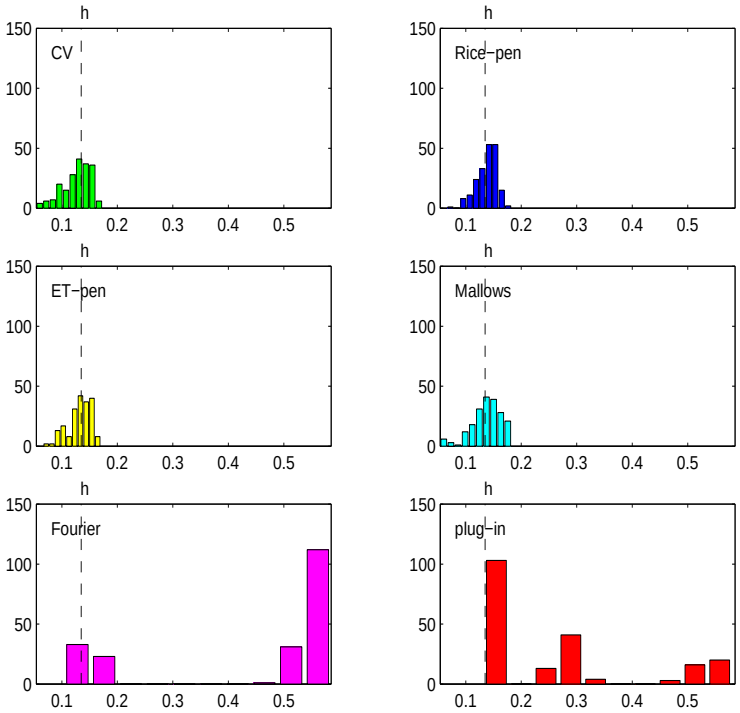
$$m(x) = -2\,\sin(-4 + 1/6\,x) + 5 + \cos(20\,x).$$

Chyby měření měly normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.3$. V tabulce 7 jsou zapsány střední hodnoty všech odhadů a také směrodatné odchylky. V prvním sloupci je označení chybových funkcí, které byly použity, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot, $std(\hat{h}_{opt})$ je jejich směrodatná odchylka. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$. Na obr.28 je znázorněno rozložení všech 200 výsledků získaných výše uvedenými metodami pro případ $\kappa = 4$. Pro $\kappa = 2, 6$ je rozložení podobné.
Z obrázku i z hodnot uvedených v tabulce je patrné, že situace je zde oproti minulé simulaci opačná. Příčinu neúspěchu posledních dvou metod lze hledat především v tom, že daná regresní funkce nesplňuje předpoklady pro cyklické rozšíření modelu, na němž jsou založeny právě tyto metody. Na druhé straně, tento fakt nemusí nutně znamenat nezdar těchto metod, jak bude ukázáno v další simulaci.

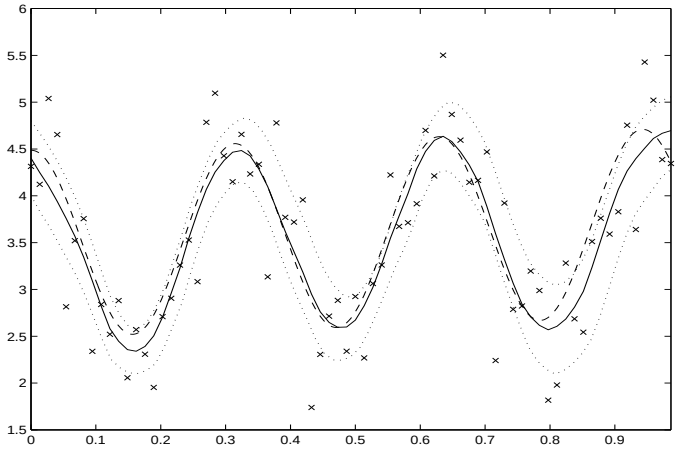
Tabulka 7: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných pomocí výše uvedených metod, $m(x) = -2\,\sin(-4 + 1/6\,x) + 5 + \cos(20\,x)$.*

	$\kappa = 2$ $h_{opt} = 0.0609$		$\kappa = 4$ $h_{opt} = 0.1349$		$\kappa = 6$ $h_{opt} = 0.2075$	
	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$
CV	0.0589	0.0144	0.1264	0.0259	0.2016	0.0379
Rice-pen	0.0683	0.0094	0.1392	0.0189	0.2183	0.0267
ET-pen	0.0647	0.0095	0.1321	0.0204	0.2112	0.0281
Mallows	0.0623	0.0137	0.1362	0.0261	0.2169	0.0394
Fourier	0.3251	0.1653	0.4485	0.1920	0.5482	0.2057
plug-in	0.1533	0.0665	0.2594	0.1495	0.3899	0.2025

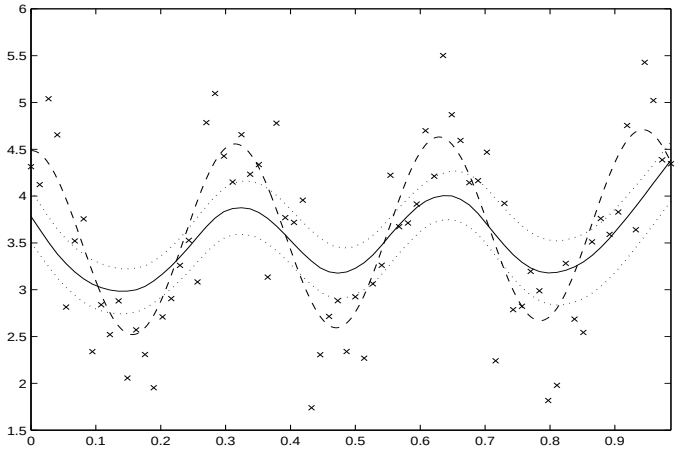
Zvolme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Při výpočtech jsme použili jádra třídy S_{04} – viz str.8. Na obr.29 jsou zobrazena simulovaná data, regresní funkce $m(x)$ a její odhad s vyhlazovacím parametrem $\hat{h}_{opt} = 0.1353$, který byl v tomto případě nalezen metodou křížového ověřování. Hodnota tohoto parametru byla nejblíže teoretické optimální šířce okna $h_{opt} = 0.1374$ a i z obrázku je patrné, že odhad regresní funkce je uspokojivý. Při aplikaci metody penalizačních funkcí a Mallowsovy metody byly hodnoty parametru $\hat{h}_{opt}$ velmi podobné. Naopak, při použití metod založených na cyklickém rozšíření modelu byly získány příliš vysoké hodnoty tohoto parametru. Obr.30 znázorňuje jádrový odhad s šířkou okna $\hat{h}_{opt} = 0.2681$, která byla nalezena plug-in metodou. Je zřejmé, že výsledný odhad je přehlazený.



Obrázek 28: *Rozložení všech 200 výsledků získaných pomocí výše uvedených metod pro případ* $\kappa = 4$.



Obrázek 29: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.3$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1353$ *pro jádro* $K \in S_{04}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.



Obrázek 30: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.3$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2681$ *pro jádro* $K \in S_{04}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.

5.3 Simulace 3

V tomto případě byla generována data s regresní funkcí

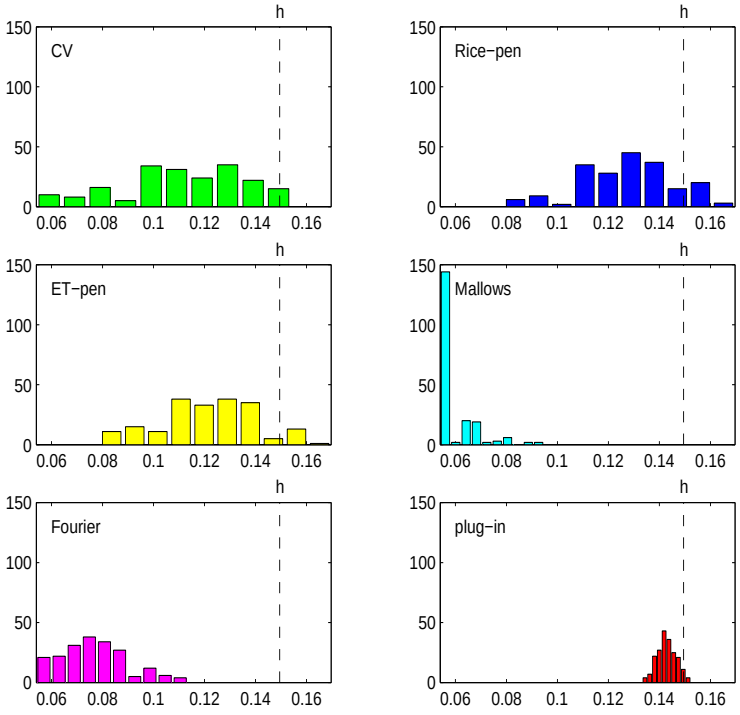
$$m(x) = -6\,\frac{\sin(11\,x + 5)}{\cot(x - 7)}.$$

Chyby měření měly normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2 = 0.2$. V tabulce 8 jsou zapsány střední hodnoty všech odhadů a také směrodatné odchylky. V prvním sloupci je označení chybových funkcí, které byly použity, $E(\hat{h}_{opt})$ je průměr všech 200 hodnot, $std(\hat{h}_{opt})$ je jejich směrodatná odchylka. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$. Na obr.31 je znázorněno rozložení všech 200 výsledků získaných výše uvedenými metodami pro případ $\kappa = 4$. Pro $\kappa = 2, 6$ je rozložení podobné.
Z obrázku i z hodnot uvedených v tabulce je zřejmé, že jednoznačně nejlepší výsledky byly dosaženy při použití plug-in metody. Tato metoda byla úspěšná i přesto, že daná regresní funkce nesplňuje předpoklad pro cyklické rozšíření modelu. Ostatní metody vedly často k příliš malým hodnotám vyhlazovacího parametru. Také rozptyl všech výsledků byl značný, největší jsme zaznamenali u metody křížového ověřování. Nejhorší odhady byly získány Mallowsovou metodou a metodou Fourierovy transformace. Obě závisí na odhadu rozptylu, který byl v tomto případě poněkud menší.

Zvolme jednu z vygenerovaných řad ke grafickému znázornění dosažených výsledků. Při výpočtech jsme použili jádra třídy S_{04} – viz str.8. Na obr.32 jsou zobrazena simulovaná data, regresní funkce $m(x)$ a její odhad s vyhlazovacím parametrem $\hat{h}_{opt} = 0.1466$, který byl v tomto případě nalezen plug-in metodou. Hodnota tohoto parametru byla nejblíže

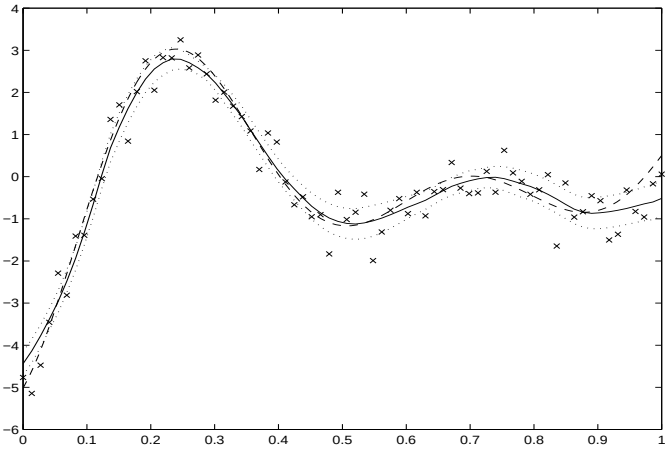
Tabulka 8: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných pomocí výše uvedených metod, $m(x) = -6\,\frac{\sin(11\,x+5)}{\cot(x-7)}$.*

	$\kappa = 2$ $h_{opt} = 0.0631$		$\kappa = 4$ $h_{opt} = 0.1495$		$\kappa = 6$ $h_{opt} = 0.2312$	
	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$	$E(\hat{h}_{opt})$	$std(\hat{h}_{opt})$
CV	0.0483	0.0104	0.1111	0.0252	0.1652	0.0378
Rice-pen	0.0581	0.0064	0.1273	0.0186	0.1844	0.0309
ET-pen	0.0558	0.0062	0.1211	0.0191	0.1794	0.0311
Mallows	0.0292	0.0033	0.0584	0.0080	0.0882	0.0106
Fourier	0.0372	0.0052	0.0766	0.0135	0.1149	0.0221
plug-in	0.0673	0.0025	0.1429	0.0037	0.2233	0.0042

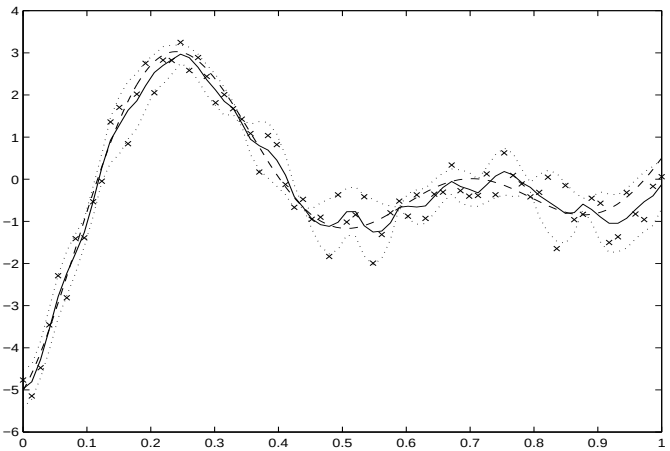


Obrázek 31: *Rozložení všech 200 výsledků získaných pomocí výše uvedených metod pro případ $\kappa = 4$.*

teoretické optimální šířce okna $h_{opt} = 0.1495$ a i z obrázku je patrné, že odhad regresní funkce je uspokojivý. Obr.33 znázorňuje jádrový odhad s šířkou okna $\hat{h}_{opt} = 0.0541$, která byla nalezena užitím metody Fourierovy transformace a Mallowsovy metody. Je zřejmé, že výsledný odhad je podhlazený.



Obrázek 32: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1466$ *pro jádro* $K \in S_{04}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.



Obrázek 33: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x)$. *Plná čára znázorňuje odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.0541$ *pro jádro* $K \in S_{04}$. *Tečkovaně je zobrazen interval spolehlivosti pro* $\alpha = 0.05$.

6 Příklady

Záměrem této kapitoly je porovnat jednotlivé metody pro odhad optimální šířky okna aplikací na reálných datech. V prvních dvou příkladech jsou použita data většího rozsahu, a přestože všechny srovnávané metody jsou asymptoticky ekvivalentní, výsledné odhady jsou rozdílné. V posledním příkladě jsou naopak analyzována data poměrně malého rozsahu. I zde jsou podle očekávání výsledky zcela rozdílné.
Při odhadech optimální šířky okna bylo použito týchž šesti metod jako v předchozí kapitole při simulacích. Také jejich označení v tabulkách je totožné s předchozím. Při jádrových odhadech regresní funkce bylo použito Nadarayových – Watsonových estimátorů. Pro ostatní typy jsou výsledky analogické.

6.1 Průměrné jarní teploty

V prvním příkladě budeme analyzovat průměrné jarní teploty naměřené v pražském Klementinu v letech 1771 – 2000[1]. Rozsah dat je tedy $T = 230$. V tabulce 9 jsou zapsány hodnoty odhadů optimální šířky okna získané jednotlivými metodami. V prvním sloupci je označení chybových funkcí, které byly použity. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$ - viz str.8.
Z tabulky je patrné, že se ve všech případech hodnoty odhadů dají rozdělit do tří skupin. V první skupině je Mallowsova metoda, neboť odhady získané touto metodou byly nejmenší. Do druhé skupiny můžeme zahrnout metodu křížového ověřování a obě penalizační funkce, které vedly ve všech případech k totožným výsledkům. Třetí skupinu reprezentují metody založené na cyklickém rozšíření regresního modelu, tj. metoda Fourierovy transformace a plug-in metoda. Aplikací těchto metod byly získány největší hodnoty vyhlazovacího parametru.

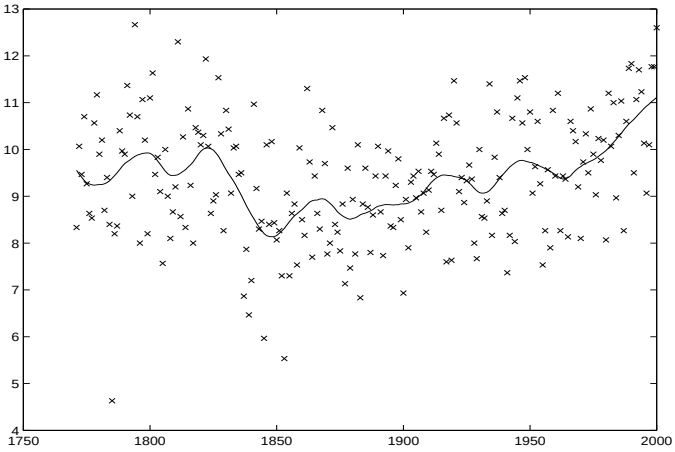
Tabulka 9: *Odhady optimální šířky okna h_{opt} získané pomocí výše uvedených metod.*

	$\kappa = 2$	$\kappa = 4$	$\kappa = 6$
CV	0.0743	0.2955	0.4011
Rice-pen	0.0743	0.2955	0.4042
ET-pen	0.0743	0.2955	0.4011
Mallows	0.0649	0.0923	0.1448
Fourier	0.1993	0.4549	0.5011
plug-in	0.1914	0.4233	0.6569

Pro jádrové odhady regresní funkce použijeme jádro řádu $\kappa = 4$. V ostatních případech

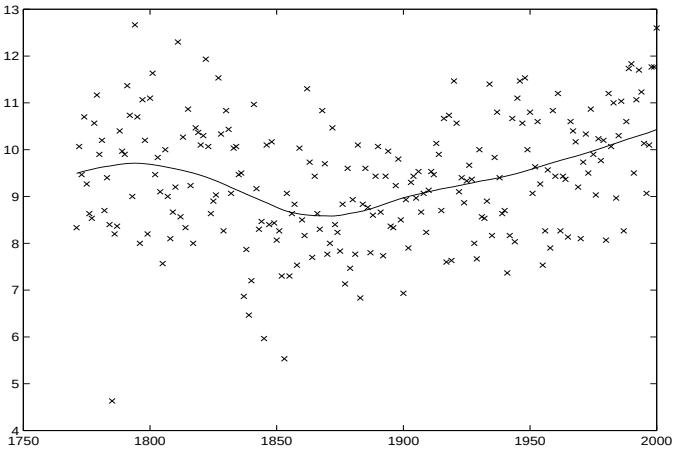
[1]Tato data poskytla katedra geografie Přírodovědecké fakulty Masarykovy univerzity v Brně.

jsou výsledky podobné. Na obr.34 je znázorněn odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.0923$ získaným Mallowsovou metodou.



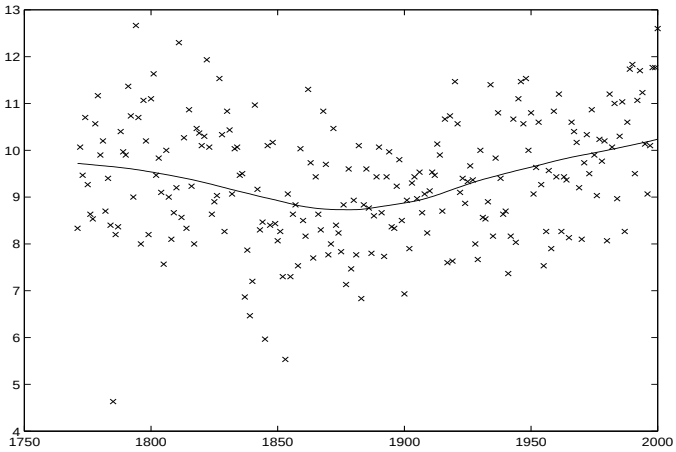
Obrázek 34: *Symboly* $\times$ *označují průměrné jarní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.0923$.

Další obrázek představuje jádrový odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.2955$, který byl nalezen metodou křížového ověřování a oběma penalizačními funkcemi.



Obrázek 35: *Symboly* $\times$ *označují průměrné jarní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2955$.

Obr.36 ilustruje kvalitu jádrového odhadu s parametrem $\hat{h}_{opt} = 0.4233$ získaným posledními dvěma metodami, tj. metodou Fourierovy transformace a plug-in metodou.
Protože neznáme skutečnou regresní funkci $m(x)$, je těžké objektivně posoudit, který z jádrových odhadů je nejlepší. Je třeba si tedy uvědomit, že konečné rozhodnutí o odhadované křivce je částečně subjektivní, neboť odhady optimální šířky okna jsou pouze



Obrázek 36: *Symboly* $\times$ *označují průměrné jarní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.4233$.

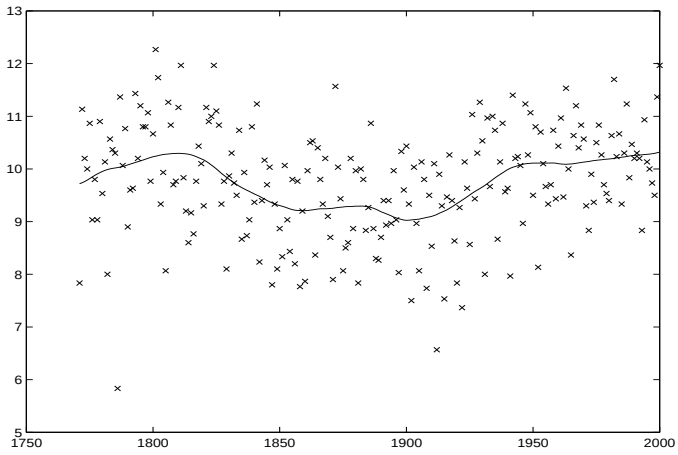
asymptoticky optimální. Z obrázků i z hodnot uvedených v tabulce je patrné, že odhady vyhlazovacího parametru Mallowsovou metodou nabývají malých hodnot a výsledný odhad regresní křivky je tedy podhlazený. Na druhé straně, metoda Fourierovy transformace a plug-in metoda vedou k již dost vysokým hodnotám a dochází tak k přehlazení dat. Dle mého názoru je v tomto případě vhodné použít buď metodu křížového ověřování nebo některou z výše uvedených penalizačních funkcí (u ostatních penalizačních funkcí docházelo k podhlazení, tj. k příliš malým hodnotám vyhlazovacího parametru). I když se výsledný odhad regresní funkce jeví jako mírně přehlazený, z výše uvažovaných křivek je nejuspokojivější.

6.2 Průměrné podzimní teploty

V dalším příkladě budeme analyzovat průměrné podzimní teploty naměřené opět v pražském Klementinu v letech 1771 – 2000. V tabulce 10 jsou zapsány hodnoty odhadů optimální šířky okna získané jednotlivými metodami. V prvním sloupci je označení chybových funkcí, které byly použity. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$ - viz str.8.
Z tabulky je patrné, že se ve všech případech hodnoty odhadů dají opět rozdělit do tří skupin. V první skupině je plug-in metoda, neboť odhady získané touto metodou byly nejmenší. Do druhé skupiny můžeme zahrnout metodu křížového ověřování, obě penalizační funkce a také Mallowsovu metodu. Tyto metody vedly ve všech případech k totožným výsledkům. Třetí skupinu reprezentuje metoda Fourierovy transformace. Aplikací této metody byly získány největší hodnoty vyhlazovacího parametru.
Stejně jako v předchozím příkladě, pro jádrové odhady regresní funkce použijeme jádro řádu $\kappa = 4$. V ostatních případech jsou výsledky podobné.
Na obr.37 je znázorněn odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.1775$ získaným plug-in metodou.

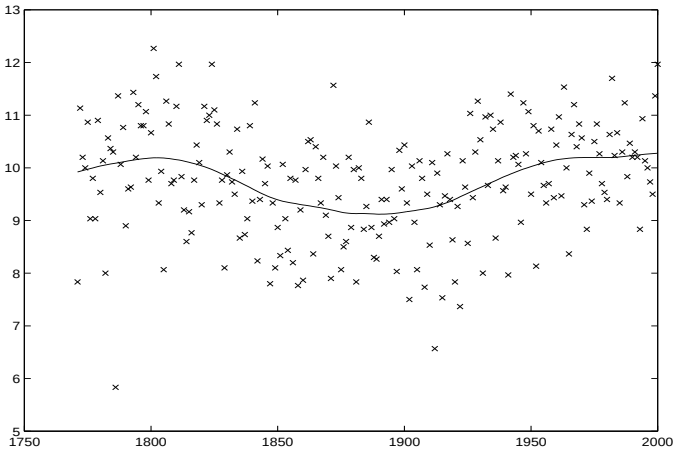
Tabulka 10: *Odhady optimální šířky okna h_{opt} získané pomocí výše uvedených metod.*

	$\kappa = 2$	$\kappa = 4$	$\kappa = 6$
CV	0.1556	0.2643	0.3261
Rice-pen	0.1556	0.2643	0.3261
ET-pen	0.1556	0.2643	0.3230
Mallows	0.1556	0.2674	0.3355
Fourier	0.2056	0.4643	0.5011
plug-in	0.1142	0.1775	0.2360

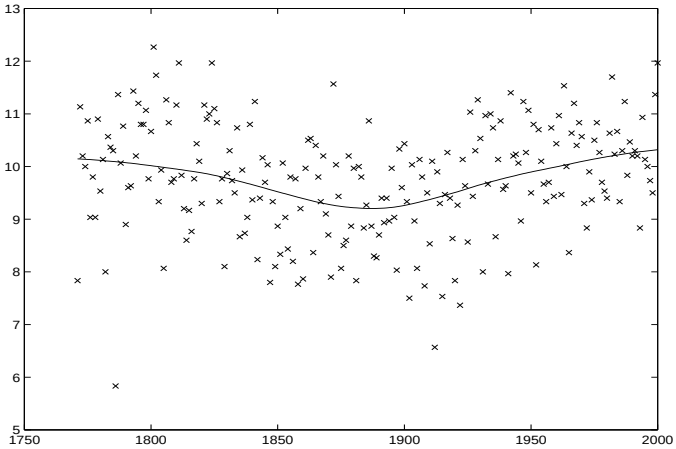


Obrázek 37: *Symboly $\times$ označují průměrné podzimní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna $\hat{h}_{opt} = 0.1775$.*

Další obrázek představuje jádrový odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.2643$, který byl nalezen metodou křížového ověřování, oběma penalizačními funkcemi a Mallowsovou metodou.
Obr.39 ilustruje kvalitu jádrového odhadu s parametrem $\hat{h}_{opt} = 0.4643$ získaným metodou Fourierovy transformace.
Protože neznáme skutečnou regresní funkci $m(x)$, je těžké objektivně posoudit, který z jádrových odhadů je nejlepší. Nezbývá nám tedy než opět subjektivně vybrat jeden z nabízených odhadů. Z obrázků i z hodnot uvedených v tabulce je patrné, že odhady vyhlazovacího parametru získané metodou Fourierových hodnot nabývají vysokých hodnot a výsledný odhad regresní křivky je přehlazený. Tuto metodu můžeme vyloučit, je třeba se tedy rozhodnout, zda použít vyhlazovací parametr získaný plug-in metodou nebo parametr získaný některou ze zbývajících metod. Domnívám se, že v tomto pří-



Obrázek 38: *Symboly* $\times$ *označují průměrné podzimní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2643$.



Obrázek 39: *Symboly* $\times$ *označují průměrné podzimní teploty. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.4643$.

padě je vhodnější použít plug-in metodu, neboť výsledný odhad s optimální šířkou okna $\hat{h}_{opt} = 0.1775$ vystihuje dle mého názoru lépe regresní funkci $m(x)$. Jádrový odhad s parametrem $\hat{h}_{opt} = 0.2643$ by mohl být také dostačující, zdá se se však, že zde již dochází k mírnému přehlazení dat.

6.3 Rozvodovost v České republice

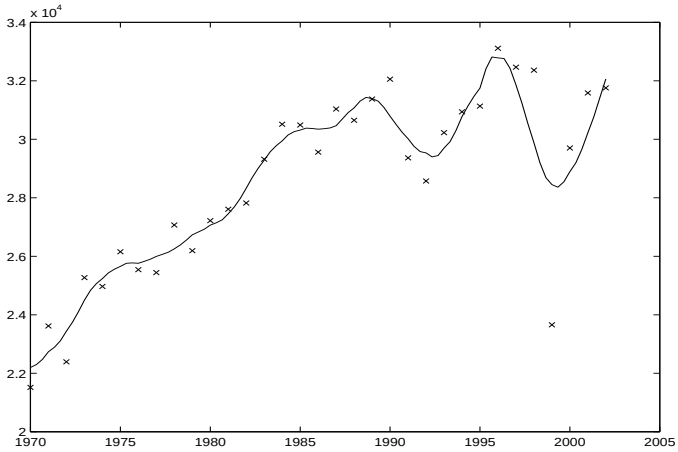
V tomto příkladě budeme sledovat počet rozvodů v České republice v letech 1970 – 2002. Data byla získána ze Statistické ročenky České republiky [2]. Na rozdíl od předchozích příkladů máme k dispozici daleko menší rozsah dat $T = 33$. Protože srovnávané metody jsou pouze asymptoticky ekvivalentní, dá se očekávat pro tak malý počet dat velká rozmanitost výsledků. V tabulce 11 jsou zapsány hodnoty odhadů optimální šířky okna získané jednotlivými metodami. V prvním sloupci je označení chybových funkcí, které byly použity. Odhady byly získány postupně pro jádra řádu $S_{0\kappa}$, kde $\kappa = 2, 4, 6$ - viz str.8.
Z tabulky je patrné, že se ve všech případech hodnoty odhadů dají rozdělit do čtyř skupin. Do první skupiny můžeme zahrnou metodu křížového ověřování a Mallowsovu metodu, neboť odhady získané těmito metodami byly nejmenší. Druhých nejnižších hodnot nabývaly odhady získané metodou Fourierovy transformace. Třetí skupinu reprezentuje plug-in metoda a do poslední skupiny můžeme zahrnout obě penalizační funkce, které vedly ve všech případech k největším hodnotám optimální šířky okna.

Tabulka 11: *Odhady optimální šířky okna h_{opt} získané pomocí výše uvedených metod.*

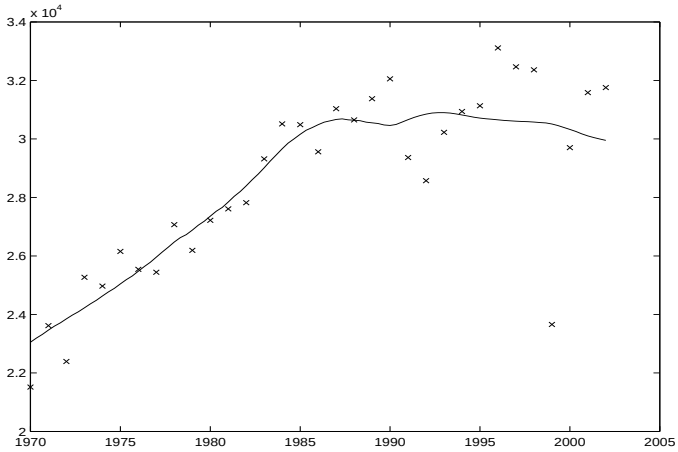
	$\kappa = 2$	$\kappa = 4$	$\kappa = 6$
CV	0.0731	0.1243	0.7193
Rice-pen	0.2825	0.5431	0.7193
ET-pen	0.2825	0.5212	0.7193
Mallows	0.0637	0.1212	0.1818
Fourier	0.1356	0.2743	0.3849
plug-in	0.1715	0.3983	0.6298

Stejně jako v předchozích příkladech použijeme pro jádrové odhady regresní funkce jádro řádu $\kappa = 4$. V ostatních případech jsou výsledky analogické.
Na obr.40 je znázorněn odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.1212$ získaným Mallowsovou metodou. Další obrázek představuje jádrový odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.2743$, který byl nalezen metodou Fourierovy transformace. Obr.42 ilustruje jádrový odhad s parametrem $\hat{h}_{opt} = 0.3983$ získaným plug-in metodou. Na posledním obrázku je znázorněn odhad regresní funkce s parametrem $\hat{h}_{opt} = 0.5212$ získaným metodami penalizačních funkcí.

Protože neznáme skutečnou regresní funkci $m(x)$, je těžké objektivně posoudit, který z jádrových odhadů je nejlepší. Nezbývá nám tedy než opět subjektivně vybrat jeden

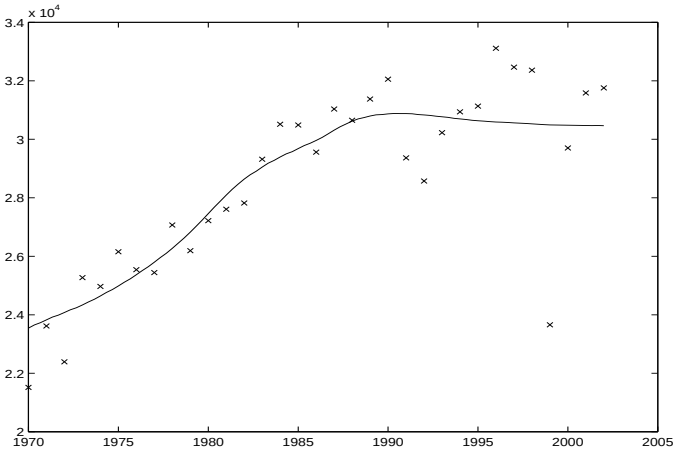


Obrázek 40: *Symboly* $\times$ *označují celkový počet rozvodů v každém roce. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.1212$.

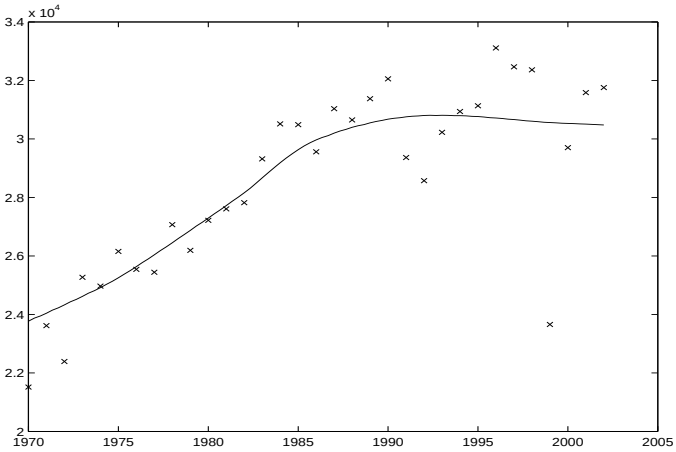


Obrázek 41: *Symboly* $\times$ *označují celkový počet rozvodů v každém roce. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.2743$.

z nabízených odhadů. Z obrázků i z hodnot uvedených v tabulce je patrné, že odhady vyhlazovacího parametru získané metodou křížového ověřování a Mallowsovou metodou nabývají malých hodnot a výsledný odhad regresní křivky je podhlazený. Tyto metody můžeme tedy vyloučit. Na druhé straně, obě penalizační funkce vedou k již dost vysokým hodnotám a dochází tak k přehlazení dat. Je třeba se tedy rozhodnout, zda použít vyhlazovací parametr získaný plug-in metodou nebo parametr získaný metodou Fourierovy transformace. Domnívám se, že v tomto případě je vhodnější použít metodu Fourierovy transformace, neboť výsledný odhad s optimální šířkou okna $\hat{h}_{opt} = 0.2743$ lépe vystihuje průběh rozvodovosti v devadesátých letech.



Obrázek 42: *Symboly* $\times$ *označují celkový počet rozvodů v každém roce. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.3983$.



Obrázek 43: *Symboly* $\times$ *označují celkový počet rozvodů v každém roce. Plná čára znázorňuje jádrový odhad regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.5212$.

Závěr

Cílem této práce bylo vytvořit ucelený souhrn poznatků z teorie hledání optimální šířky okna. Tento parametr nejvíce ovlivňuje výsledný odhad a jeho volba je zásadním problémem ve vyhlazovacích metodách.
Mezi původní výsledky můžeme zařadit např. novou penalizační funkci *ET bandwidth selector*. Při testování na simulovaných příkladech i při aplikacích na reálných datech patřila tato funkce k nejlepším ze všech uvedených penalizačních funkcí.
Také myšlenka využití Fourierovy transformace k odhadu střední kvadratické chyby patří k novějším trendům v problematice hledání optimální šířky okna. V souvislosti s odhady regresní funkce byla poprvé uvedena v [10]. V této práci se podařilo zobecnit tento postup pro všechny uvažované typy jádrových estimátorů a také pro libovolná jádra třídy $S_{0\kappa},\ \kappa$ sudé. Částečně se tak vyřešil problém podhlazování.
Dalším původním výsledkem je odhad parametru A_κ, kde se vyskytuje κ - tá derivace neznámé regresní funkce $m(x)$. Tohoto odhadu lze využít v plug-in metodě, jejíž výhoda je především v tom, že odpadá problém minimalizace chybové funkce, protože hodnota minima je již teoreticky odvozena. Plug-in metoda je v literatuře často uváděna (např. [6], [9], [19]), avšak mnohdy jen v souvislosti s faktem, že odhad neznámých parametrů je obtížný. Explicitní formule pro vyjádření těchto parametrů se často vyskytují jen při odhadech hustoty.
Jelikož se jedná o velmi rozsáhlou problematiku, je celá práce zaměřena pouze na model s ekvidistantními body plánu a na odhad globálního vyhlazovacího parametru. Konečný odhad regresní funkce je ovlivněn ještě dalšími skutečnostmi. Například na okrajích intervalu se mohou objevit tzv. hraniční efekty a odhadům v těchto bodech je třeba věnovat zvláštní pozornost. Tento problém je možno řešit např. použitím hraničních jader (viz [7]). Zajímavé jsou také otázky optimální volby řádu jádra nebo hledání optimální šířky okna při jádrových odhadech regresní funkce ve více dimenzích. Tyto oblasti by mohly být podnětem pro další studium.

Všechny typy jádrových odhadů i všechny metody pro odhad optimální šířky okna uvedené v této práci jsem naprogramoval v systému MATLAB. Vznikla tak knihovna pro jádrové vyhlazování, která umožňuje srovnání uvedených metod na simulovaných nebo reálných datech. Tato knihovna, včetně podrobného návodu k použití, je na přiloženém CD.

Reference

[1] Craven P., Wahba G., *Smoothing Noisy Data with Spline Function*, Numer. Math. 31, 377-403, 1979

[2] Český statistický úřad, *Statistická ročenka České republiky 2003*, Scientia, Praha, 2003

[3] Čížek V., *Diskrétní Fourierova transformace a její použití*, SNTL, Praha, 1981

[4] Droge B., *Some Comments on Cross-Validation*, Statistical Theory and Computational Aspects of Smoothing, 178-199, Physica-Verlag, 1996

[5] Härdle W., Hall P., Marron J.S., *How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum?*, Journal of the American Statistical Association 83, 86-95, 1988

[6] Härdle W., *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990

[7] Horová I., *Optimization Problems Connected with Kernel Smoothing, Signal Processing, Communications, and Computer Science*, World Scientific and Engeneering Society Press, 2000

[8] Horová I., Zelinka J., *Základy a aplikace jádrových odhadů*, Sborník Analýza dat 2000, 141-167, Trilobyte, Pardubice, 2000

[9] Chiu S.T., *Some Stabilized Bandwidth Selectors for Nonparametric Regression*, Annals of Statistics, 19, 1528-1546, 1991

[10] Chiu S.T., *Why Bandwidth Selectors Tend to Choose Smaller Bandwidths, and a Remedy*, Biometrika, 77, 222-6, 1990

[11] Koláček J., *Kernel Estimation of the Regression Function - Bandwidth Selection*, Summer School DATASTAT'01 Proceedings FOLIA, 129 - 138, 2002

[12] Koláček J., *Problems of Automatic Data-Driven Bandwidth Selectors for Nonparametric Regression*, Journal of Electrical Engineering, vol. 53, No. 12, SCAM'02 Bratislava, 48 - 52, 2002

[13] Koláček J., *Some Stabilized Bandwidth Selectors for Nonparametric Regression*, Journal of Electrical Engineering, vol. 54, No. 12, ISCAM'03 Bratislava, 65 - 68, 2003

[14] Koláček J., *Use of Fourier Transformation for Bandwidth Selection*, Summer school DATASTAT'03, Proceedings FOLIA, 118 - 127, 2004

[15] Koláček J., *Use of Fourier Transformation for Kernel Smoothing*, Proceedings in Computational Statistics COMPSTAT'04, 1329 - 1336, 2004

[16] Mallows C., *Some Comments on C_p*, Technometrics 15, 661-675, 1973

[17] Rice J., *Bandwidth Choice for Nonparametric Regression*, The Annals of Statistics 12, 1215-1230, 1984

[18] Stone M., *Cross-validatory Choice and Assessment of Statistical Predictions*, Journal of the Royal Statistical Society B 36, 111-147, 1974

[19] Wand M.P., Jones M.C., *Kernel Smoothing*, Chapman & Hall, London, 1995