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*.

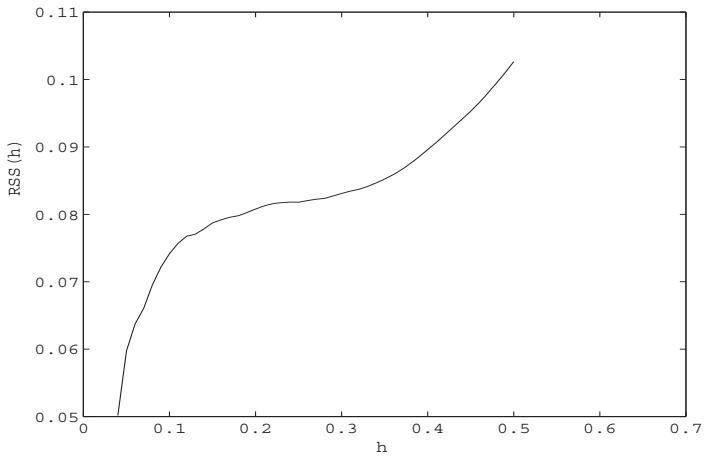
V praxi se také často minimalizuje tzv. *průměrná kvadratická chyba ASE* (Average Square Error), která je asymptoticky ekvivalentní s *AMSE* (viz např. [6])

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

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 $\hat{h}_{opt}$.

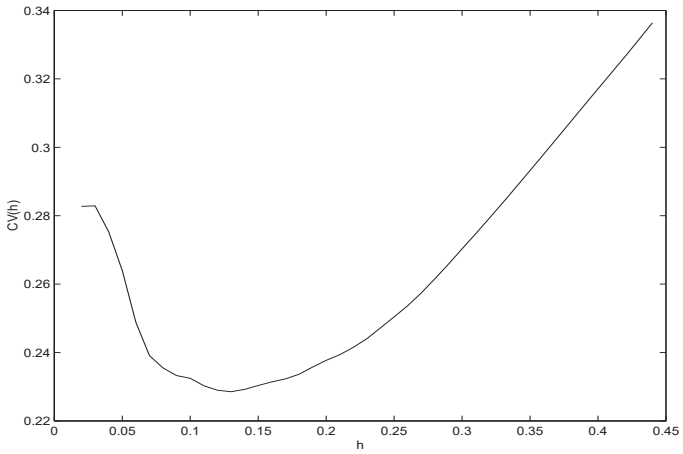
Označme $\widehat{m}_j(x_j;h)$ odhad hodnoty regresní funkce $\widehat{m}$ v bodě x_j bez použití tohoto bodu, tj.

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

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

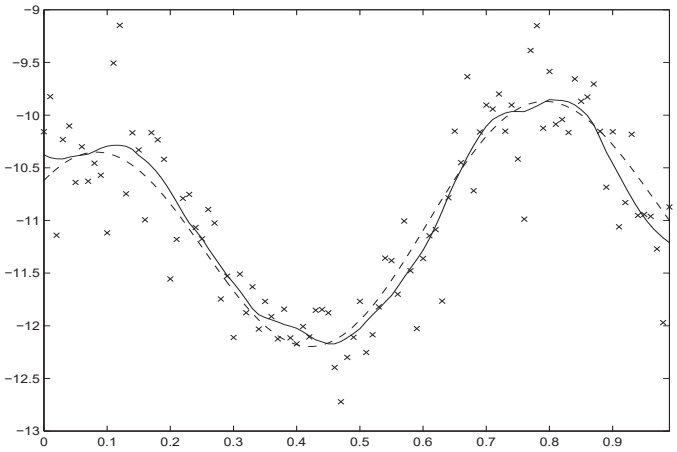
$$m(x) = \cos(9x - 7) - (3 + x^{12})/6 + 8^{x-1}.$$

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).

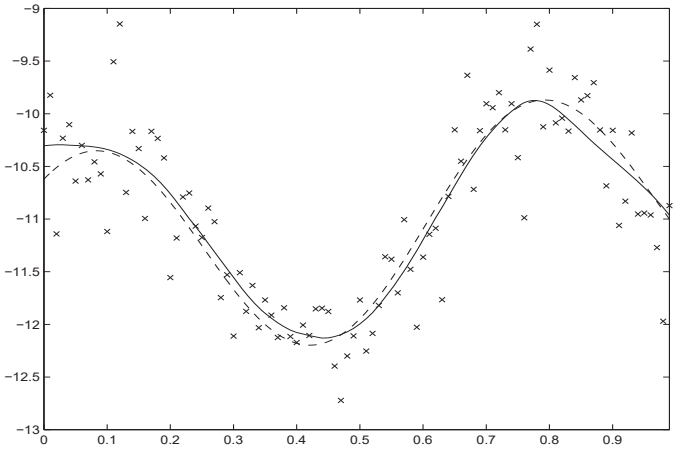
Tabulka 2: *Střední hodnoty a směrodatné odchylky odhadů parametru h_{opt} získaných metodou křížového ověřování.*

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)

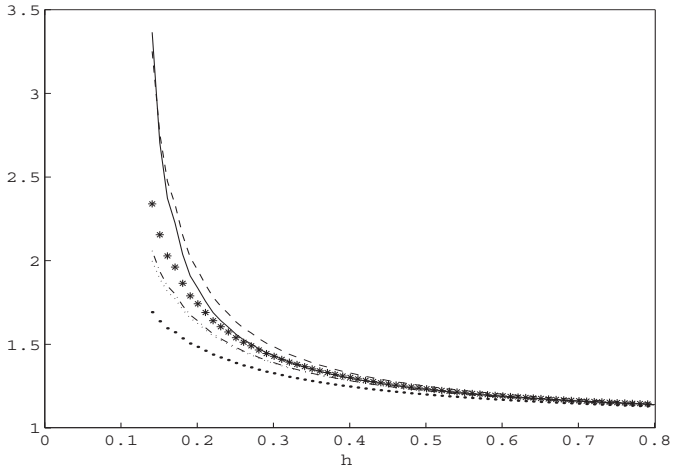
$$\Xi_S(u) = 1 + 2u$$

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

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

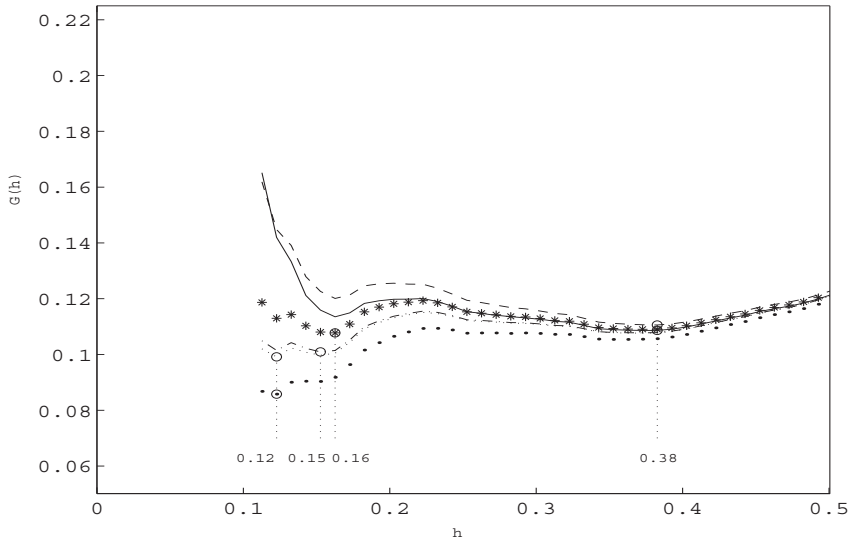
Při výpočtech jsme použili Nadarayovy – Watsonovy estimátory a jádra třídy $S_{0\kappa}$ (viz tab.1) pro $\kappa = 2,\ 4,\ 6,\ 8$. Bylo vygenerováno 200 řad. U každé řady byly získány odhady optimální šířky okna tak, že jsme nejprve spočítali hodnoty funkce $G(h)$ v 321 bodech ekvidistantně rozložených 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

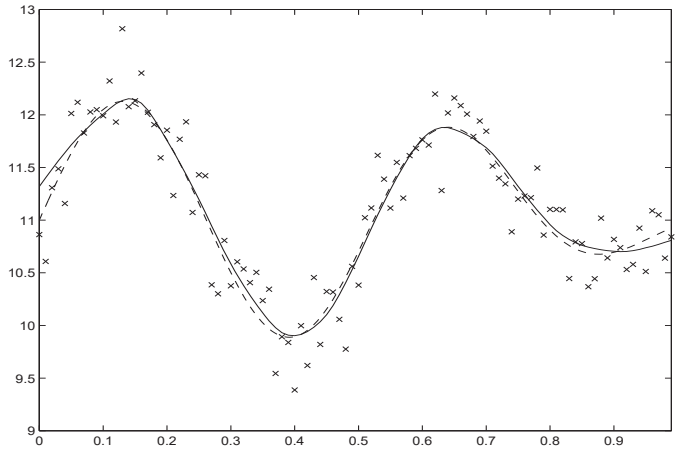
Porovnáme-li v tabulce hodnoty $\hat{h}_{opt}$ pro různé penalizační funkce s teoretickou optimální šířkou okna, je zřejmé, že jsou pro všechna κ výsledné odhady menší. S rostoucím κ jsou větší rozdíly mezi výsledky získanými jednotlivými metodami. To je způsobeno především tím, že odhady regresní funkce 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 $h_{opt} = 0.3742$.*

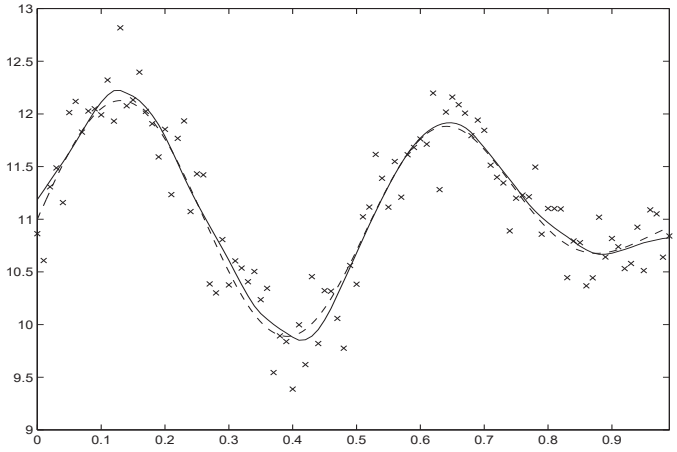
byly získány při použití AIC a Shibatovy penalizační funkce.
Pro odhad regresní funkce jsme použili jádro třídy S_{04}. Obr.12 znázorňuje jádrový odhad s použitím teoretické optimální šířky okna $h_{opt} = 0.1739$.



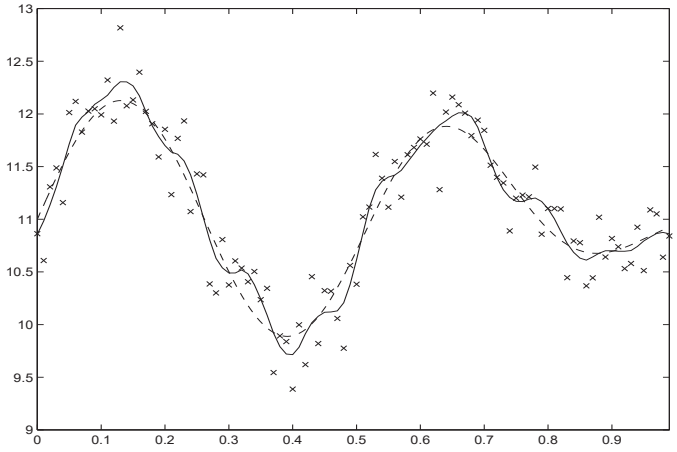
Obrázek 12: *Symboly × označují naměřené hodnoty Y s chybou $\sigma^2 = 0.2$. Čárkovaně je zobrazena skutečná regresní funkce $m(x) = 11 - 1/3 \tan(5 + x^6) \sin(12\,x)$. Plná čára znázorňuje jádrový odhad této regresní funkce s optimální šířkou okna $h_{opt} = 0.1739$. pro jádro $K \in S_{04}$*

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* $K \in S_{04}$

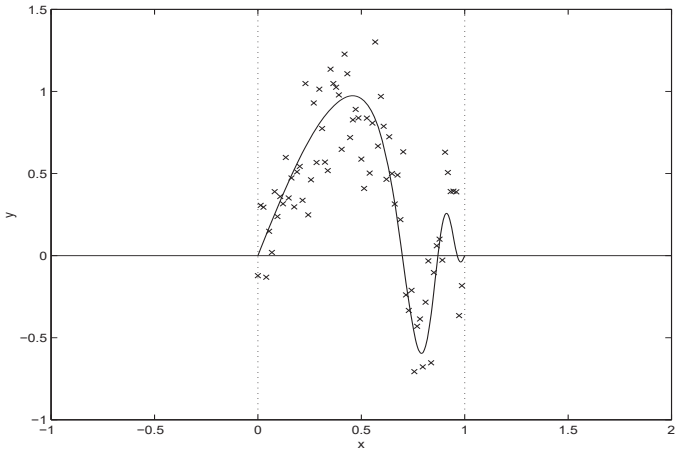


Obrázek 14: *Symboly* $\times$ *označují naměřené hodnoty* Y *s chybou* $\sigma^2 = 0.2$. *Čárkovaně je zobrazena skutečná regresní funkce* $m(x) = 11 - 1/3\,\tan(5 + x^6)\sin(12\,x)$. *Plná čára znázorňuje jádrový odhad této regresní funkce s šířkou okna* $\hat{h}_{opt} = 0.0672$. *pro jádro* $K \in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

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.