JÁDROVÉ ODHADY DERIVACE REGRESNÍ FUNKCE

JIŘÍ ZELINKA, IVANA HOROVÁ

Abstrakt. It is possible to review the quality of kernel estimations of the regression function and its derivative from several points of view. According to them we can choose the parameters influencing the kernel estimations. The parameters are: the bandwidth, the shape and the order of the used kernel function. It can be shown that it is possible to introduce the criterion for the optimal shape as well as for optimal order of kernel function. This paper shortly aspires to summarize attained results from this branch and to demonstrate their application for simulated and real data sets.

Резюме: Качество оценок функции регрессии и её производной при помощи ядер возможно обсуждать из различных точек зрения и смотря по ним потом выбирать параметры которые влияют на эти оценки: прежде всего это чирина сглаживающего окна, но тоже форма ядра и его порядок. Возможно ввести критерий не только для выбора оптимальной формы ядра, но тоже для его порядка. Эта статья кратко подытоживает результаты в этой области включительно их прилозхений к имитирующим и реальным данным.

V tomto příspěvku budeme uvažovat pouze model s pevným plánem, tj. mějme pevně zadané body x_i, $i = 1, \dots, n$, v nich naměřené hodnoty Y_i a předpokládáme model v obecném tvaru

$$Y_i = m(x_i) + \sqrt{v(x_i)} \cdot \varepsilon_i. \quad i = 1, \dots, n, \tag{1}$$

kde m je neznámá regresní funkce, ε_i jsou nezávislé chyby jednotlivých měření, pro něž platí

$$E\varepsilon_i = 0, \qquad Var(\varepsilon_i) = 1, \quad i = 1, \dots, n,$$

a $v(x)$ je nezáporná varianční funkce. Bez újmy na obecnosti můžeme dále předpokládat, že všechny body x_i leží v intervalu $[0, 1]$ a jsou vzestupně uspořádány.

Pro odhad regresní funkce m, resp. její ν-té derivace $m^{(\nu)}$ použijeme tzv. jádrové odhady. Ty se konstruují užitím jádrových funkcí (zkráceně jader) označovaných zpravidla $K(x)$. Ukazuje se, že při výpočtu statistických vlastností jádrových odhadů regresní funkce je vhodné použít jádra splňující určité momentové podmínky. Proto zavádíme následující definici:

Definice: Nechť ν, k jsou nezáporná celá čísla stejné parity, $k > \nu$. Symbolem $\mathcal{M}_{\nu,k}$ označíme třídu funkcí K spojitých na $\mathbf{R}$ s nosičem $[-1, 1]$, které jsou na tomto intervalu lipschitzovsky spojité a které splňují následující podmínky:

$$\int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, 0 \le j < k, j \ne \nu \\ (-1)^\nu \nu!, j = \nu \\ \beta_k \ne 0, j = k. \end{cases} \tag{2}$$

2000 *Mathematics Subject Classification.* Primary 93E14; Secondary 30C40.
Klíčová slova. Jádrové odhady, Gasser-Müllerův odhad, křížové ověřování.
Příspěvek vznikl s podporou výzkumného záměru MŠMT, CEZ: J07/98:143100001.

Poznámka: Ze spojitosti funkcí K plyne $K(-1) = K(1) = 0$. Funkce třídy $\mathcal{M}_{\nu,k}$ nazýváme jádra řádu k.

Odhad ν-té derivace funkce m můžeme nyní zapsat pomocí Gasser-Müllerova tvaru odhadu jako

$$\hat{m}_{GM}^{(\nu)}(x) = \hat{m}^{(\nu)}(x) = \sum_{i=1}^{n} \frac{1}{h^{\nu+1}} \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du Y_i, \tag{3}$$

kde $K \in \mathcal{M}_{\nu,k}$ a hraniční meze jednotlivých integrálů jsou ve tvaru

$$s_0 = 0, \quad s_i = (x_{i+1} + x_i)/2 \text{ pro } i = 1, \dots n-1, \quad s_n = 1.$$

Místo $\hat{m}^{(0)}$ budeme psát $\hat{m}$. Parametr h se nazývá šířka vyhlazovacího okna a má podstatný vliv na kvalitu jádrového odhadu regresní funkce m nebo její derivace. Mezi další významné faktory ovlivňující odhad patří tvar jádra K a také jeho řád k.

Pro posouzení kvality jádrového odhadu v bodě x zpravidla používáme střední kvadratickou chybu

$$MSE(\hat{m}^{(\nu)}(x), h) = E\left(\hat{m}^{(\nu)}(x) - m^{(\nu)}(x)\right)^2 = \tag{4}$$

$$= Var\, \hat{m}^{(\nu)}(x) + (E\hat{m}^{(\nu)}(x) - m(x))^2.$$

Za předpokladu dostatečné hladkosti funkce m lze tuto střední kvadratickou chybu v bodě x zapsat ve tvaru (viz [5])

$$MSE(\hat{m}^{(\nu)}(x), h) = \frac{v(x)[V(K) + o(1)]}{nh^{2\nu+1}} + h^{2(k-\nu)} \left[\frac{\beta_k^2}{(k!)^2}\left(m^{(k)}(x)\right)^2 + o(1)\right], \tag{5}$$

kde

$$V(K) = \int_{-1}^{1} K^2(t)dt, \quad \beta_k = \int_{-1}^{1} t^k K(t)dt.$$

Vidíme, že pokud n poroste do nekonečna a současně se bude měnit šířka vyhlazovacího okna $h = h_n$ tak, aby platilo $h_n \to 0$, $n \cdot h_n^{2\nu+1} \to \infty$, bude střední kvadratická chyba $MSE(\hat{m}^{(\nu)}(x), h_n)$ konvergovat k nule, tj. jedná se o konzistentní odhad $m^{(\nu)}$ v bodě x.

V dalším textu budeme používat tzv. hlavní člen střední kvadratické chyby, který označíme $\overline{MSE}$ a který dostaneme zanedbáním členů $o(1)$ konvergujících k nule pro $n \to \infty$. Platí tedy

$$\overline{MSE}(\hat{m}^{(\nu)}(x), h) = \frac{v(x)V(K)}{nh^{2\nu+1}} + h^{2(k-\nu)} \frac{\beta_k^2}{(k!)^2}\left(m^{(k)}(x)\right)^2 \tag{6}$$

Jako globální kritérium kvality odhadu použijeme průměrnou střední kvadratickou chybu

$$AMSE(\hat{m}^{(\nu)}, h) =$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E\left(\hat{m}^{(\nu)}(x_i) - m^{(\nu)}(x_i)\right)^2 \approx \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{v(x_i)V(K)}{nh^{2\nu+1}} + h^{2(k-\nu)}\frac{\beta_k^2}{(k!)^2}\left(m^{(k)}(x_i)\right)^2\right) = \\
&= \frac{V(K)}{nh^{2\nu+1}}\frac{1}{n}\sum_{i=1}^{n} v(x_i) + h^{2(k-\nu)}\frac{\beta_k^2}{(k!)^2}\frac{1}{n}\sum_{i=1}^{n}\left(m^{(k)}(x_i)\right)^2.
\end{aligned}$$

Pokud označíme

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v(x_i), \quad (\bar{m}^{(k)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(m^{(k)}(x_i)\right)^2,$$

dostaneme pro hlavní člen průměrné střední kvadratické chyby vztah

$$\overline{AMSE}(\hat{m}^{(\nu)}, h) = \frac{V(K)\bar{v}}{nh^{2\nu+1}} + h^{2(k-\nu)}\beta_k^2\frac{(\bar{m}^{(k)})^2}{(k!)^2}. \tag{7}$$

Známe-li varianční funkci v i regresní funkci m (např. pro simulovaná data), můžeme určit hodnotu šířky okna $h_{opt,\nu,k}$, která minimalizuje $\overline{AMSE}$. Výpočtem dostaneme

$$h_{opt,\nu,k}^{2k+1} = \frac{(2\nu+1)V(K)\bar{v}\,(k!)^2}{2n(k-\nu)\beta_k^2(\bar{m}^{(k)})^2}. \tag{8}$$

Dalším cílem je taková úprava vztahu (7), která by umožnila zbavit se neznámého výrazu $\bar{m}^{(k)}$. K tomu použijeme následující postup použitý v [4]:
Zavedeme označení $K_\delta(\cdot) = \frac{1}{\delta^{\nu+1}}K(\frac{\cdot}{\delta})$ pro $\delta > 0$. Použitím jádra K_δ dostaneme stejný efekt, jako kdybychom šířku vyhlazovacího okna h nahradili šířkou $h^* = h/\delta$. Platí totiž, že „množství vyhlazení“ pomocí jádra K s vyhlazovacím parametrem h je stejné jako „množství vyhlazení“ jádrem K_δ s parametrem $h^* = h/\delta$, neboť

$$\begin{aligned}
\hat{m}^{(\nu)}(x) &= \frac{1}{h^{\nu+1}}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right)duY_i = \\
&= \frac{1}{(h^*\delta)^{\nu+1}}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h^*\delta}\right)duY_i = \\
&= \frac{1}{h^{*\nu+1}}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} \frac{1}{\delta^{\nu+1}}K\left((\frac{x-u}{h^*})/\delta\right)duY_i = \\
&= \frac{1}{h^{*\nu+1}}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K_\delta\left(\frac{x-u}{h^*}\right)duY_i.
\end{aligned}$$

Průměrná střední kvadratická chyba vyjádřená pomocí jádra K_δ pak má tvar:

$$\overline{AMSE}(\hat{m}^{(\nu)}, h^*) = \frac{\bar{v}}{nh^{*2\nu+1}}\int_{-\delta}^{\delta} K_\delta^2(t)dt + h^{*2(k-\nu)}\frac{(\bar{m}^{(k)})^2}{(k!)^2}\left(\int_{-\delta}^{\delta} t^k K_\delta(t)dt\right)^2. \tag{9}$$

Snažíme se určit δ tak, aby „příspěvek“ jader k oběma částem chyby byl stejný, tedy aby platilo

$$\int_{-\delta}^{\delta} K_\delta^2(t)dt = \left(\int_{-\delta}^{\delta} t^k\, K_\delta(t)dt\right)^2. \tag{10}$$

Takové δ se dá snadno určit:

$$\delta^{2k+1} = \frac{\int_{-1}^{1} K^2(x)dx}{\left(\int_{-1}^{1} x^k K(x)dx\right)^2},$$

tj.

$$\delta^{2k+1} = \frac{V(K)}{\beta_k^2}.$$

Toto číslo se nazývá kanonický faktor a značíme ho δ_0, tedy

$$\delta_0 = \left(\frac{V(K)}{\beta_k^2}\right)^{\frac{1}{2k+1}}. \tag{11}$$

Jádro K_{δ_0} příslušné tomuto kanonickému faktoru se nazývá kanonické jádro.

Položme nyní $h = h^*\delta_0$ a dosadíme tuto hodnotu do vztahu (7) pro $\overline{AMSE}$. Po úpravách dostaneme

$$\overline{AMSE}(\hat{m}^{(\nu)}, h^*) = T(K)\left\{\frac{\bar{v}}{nh^{*2\nu+1}} + h^{*2(k-\nu)}\frac{(\bar{m}^{(k)})^2}{(k!)^2}\right\}, \tag{12}$$

kde $T(K)$ je funkcionál závislý jenom na jádrové funkci K a dá se zapsat ve tvaru

$$T(K) = \left(\left|\int_{-1}^{1} x^k K(x)dx\right|^{2\nu+1}\left(\int_{-1}^{1} K^2(x)dx\right)^{k-\nu}\right)^{\frac{2}{2k+1}}. \tag{13}$$

Toto vyjádření $\overline{AMSE}$ pomocí funkcionálu $T(K)$ umožňuje posoudit vliv tvaru jádra na $\overline{AMSE}$.

Najít funkci, která jej minimalizuje, je úloha variačního počtu. Její řešení lze nalézt např. v článku [4]. Jedná se o polynom s nosičem $[-1, 1]$, který se dá vyjádřit pomocí Legendrových polynomů. Následující tabulka udává explicitní tvar těchto jádrových funkcí pro některé hodnoty ν a k a současně také příslušný kanonický faktor δ_0.

$\nu = 0$		
k	δ_0	K_{opt}
2	1.7188	$-\frac{3}{4}(x^2-1)$
4	2.0165	$\frac{15}{32}(x^2-1)(7x^2-3)$
6	2.0834	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$
8	2.1021	$\frac{315}{4096}(x^2-1)(715x^6-1001x^4+385x^2-35)$
10	2.1062	$-\frac{3465}{65536}(x^2-1)(4199x^8-7956x^6+4914x^4-1092x^2+63)$
12	2.1051	$\frac{9009}{524288}(x^2-1)(52003x^{10}-124355x^8+106590x^6-39270x^4+$ $+5775x^2-231)$

$\nu = 1$		
k	δ_0	K_{opt}
3	1.4204	$\frac{15}{4}x(x^2-1)$
5	1.7656	$-\frac{105}{32}x(x^2-1)(9x^2-5)$
7	1.8931	$\frac{315}{32}x(x^2-1)(143x^4-154x^2+35)$
9	1.9526	$-\frac{3465}{4096}x(x^2-1)(1105x^6-1755x^4+819x^2-105)$
11	1.9843	$\frac{45045}{65536}x(x^2-1)(6783x^8-14212x^6+10098x^4-2772x^2+231)$
13	2.0027	$-\frac{45045}{524288}x(x^2-1)(260015x^{10}-676039x^8+646646x^6-$ $-277134x^4+51051x^2-3003)$

$\nu = 2$		
k	δ_0	K_{opt}
4	1.3925	$-\frac{105}{16}(x^2-1)(5x^2-1)$
6	1.6964	$\frac{315}{64}(x^2-1)(77x^4-58x^2+5)$
8	1.8269	$-\frac{3465}{2048}(x^2-1)(1755x^6-2249x^4+721x^2-35)$
10	1.8946	$\frac{45045}{8192}(x^2-1)(3553x^8-6392x^6+3618x^4-672x^2+21)$
12	1.9341	$-\frac{45045}{262144}(x^2-1)(676039x^{10}-1562351x^8+1271974x^6-$ $-429726x^4+52899x^2-1155)$
14	1.9590	$\frac{765765}{1048576}(x^2-1)(884925x^{12}-2495270x^{10}+2653027x^8-$ $-1315028x^6+301587x^4-26598x^2+429)$

$\nu = 3$		
k	δ_0	K_{opt}
5	1.4086	$\frac{945}{16}x(x^2-1)(7x^2-3)$
7	1.6725	$-\frac{10395}{64}x(x^2-1)(39x^4-38x^2+7)$
9	1.7958	$\frac{135135}{2048}x(x^2-1)(935x^6-1405x^4+597x^2-63)$
11	1.8641	$-\frac{135135}{8192}x(x^2-1)(29393x^8-59432x^6+40018x^4-$ $-10032x^2+693)$
13	1.9060	$\frac{2297295}{262144}x(x^2-1)(382375x^{10}-969703x^8+895622x^6-$ $-364078x^4+61347x^2-3003)$
15	1.9334	$-\frac{43648605}{1048576}x(x^2-1)(510255x^{12}-1554570x^{10}+1825625x^8-$ $-1034540x^6+288145x^4-35178x^2+1287)$

Vidíme, že pro parametry $\nu = 0$ a $k = 2$ dostáváme známé Epanečnikovo jádro.

Nyní můžeme pomocí standardního postupu minimalizovat funkci ve vztahu (12) vzhledem k šířce okna h^*. Optimální hodnotu získáme jako stacionární bod, výpočtem tedy dostaneme

$$h^*_{opt,\nu,k}{}^{2k+1} = \frac{(2\nu+1)\bar{v}(k!)^2}{2n(k-\nu)(\bar{m}^{(k)})^2} \tag{14}$$

a odtud

$$\frac{(\bar{m}^{(k)})^2}{(k!)^2} = \frac{(2\nu+1)\bar{v}}{2n(k-\nu){h^*_{opt,\nu,k}}^{2k+1}} \tag{15}$$

Dosazením do rovnice (12) vyjde

$$\overline{AMSE}(\hat{m}^{(\nu)}, h^*_{opt,\nu,k}) = T(K)\frac{\bar{v}(2k+1)}{2n(k-\nu){h^*}^{2\nu+1}_{opt,\nu,k}} \tag{16}$$

a ze vztahu (15) navíc plyne

$$h^*_{opt,\nu,k} = \left(\frac{(2\nu+1)\,k}{k-\nu}\right)^{\frac{1}{2k+1}} \cdot h^*_{opt,0,k}, \tag{17}$$

pokud k a ν je sudé, respektive

$$h^*_{opt,\nu,k} = \left(\frac{(2\nu+1)\,(k-1)}{3(k-\nu)}\right)^{\frac{1}{2k+1}} \cdot h^*_{opt,1,k}, \tag{18}$$

pokud k a ν je liché.

Optimální šířku okna $h^*_{opt,\nu,k}$ můžeme odhadnout například pomocí metody křížového ověřování (anglicky cross-validation method, viz [1], [2]), přičemž z (17) a (18) vidíme, že nemusíme určovat hodnoty $h^*_{opt,\nu,k}$ pro všechna ν, což v obecném případě je výpočetně složité, ale stačí zjistit tyto veličiny pro $\nu = 0$ nebo $\nu = 1$ podle toho, zda počítáme odhad sudé nebo liché derivace regresní funkce m.

Základní myšlenku metody křížového ověřování lze formulovat takto:

Odhadneme hodnotu $\hat{m}$ postupně v každém bodě x_i, $i = 1, \ldots, n$ bez použití vlastního bodu x_i, tedy pouze pomocí zbývajících $n-1$ bodů. Pak vybereme takovou hodnotu h, pro kterou jsou hodnoty v chybějících bodech nejlépe odhadnuty pomocí zbývajících bodů. To znamená, že minimalizujeme funkci křížového ověřování $CV(h)$, kterou lze pro mnoho typů odhadů vyjádřit ve tvaru součtu reziduí, tj.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_{-i}(x_i)\right)^2,$$

kde $\hat{m}_{-i}(x_i)$ je výše zmíněný odhad funkce m v bodě x_i při vynechání i-tého pozorování.

Pro Gasser-Müllerův tvar odhadu je možné tuto funkci zapsat též jako

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}(x_i)}{1 - s_{ii}}\right)^2 \tag{19}$$

pro

$$s_{ii} = \frac{1}{h}\int_{s_{i-1}}^{s_i} K\left(\frac{x_i - u}{h}\right) du, \qquad i = 1, \ldots, n.$$

Můžeme také použít zobecněné funkce křížového ověřování (podrobněji viz [5], [6])

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}(x_i)}{1 - s/n}\right)^2, \qquad s = \sum_{i=1}^{n} s_{ii}. \tag{20}$$

Funkce křížového ověřování pro odhad $h_{opt,1,k}$ má tvar (viz [5])

$$CV^{(1)}(h) = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{Y_{i+1} - Y_i}{x_{i+1} - x_i} - \hat{m}^{(1)}_{-(i,i+1)}\left(\frac{x_{i+1} + x_i}{2}\right)\right)^2, \tag{21}$$

$\hat{m}^{(1)}_{-(i,i+1)}((x_{i+1}+x_i)/2)$ je odhad první derivace funkce m v bodě $((x_{i+1}+x_i)/2)$, při vynechání pozorování v bodech x_i a x_{i+1}.

Odhad $\hat{h}^*_{opt,0,k}$ nebo $\hat{h}^*_{opt,1,k}$ optimální šířky okna $h^*_{opt,0,k}$ nebo $h^*_{opt,1,k}$ tedy určíme jako tu hodnotu, která minimalizuje funkci $CV(h^*)$, resp. $GCV(h^*)$ nebo $CV^{(1)}(h^*)$ s jádrem K_{δ_0}. Hodnoty h^* přitom bereme z vhodné množiny H_n vyhlazovacích parametrů, což je zpravidla uzavřený interval. Na základě praktických zkušeností můžeme vzít jako dolní mez tohoto intervalu vzdálenost mezi body x_i a jako horní mez dvojnásobek délky nejmenšího intervalu obsahujícího všechny body x_i. Pro ekvidistatní body na intervalu $[0,1]$ je pak $H_n = [1/n, 2]$. Tedy

$$\hat{h}^*_{opt,0,k} = \arg\min_{h^*\in H_n} CV(h^*), \quad \text{resp.}$$

$$\hat{h}^*_{opt,1,k} = \arg\min_{h^*\in H_n} CV^{(1)}(h^*),$$

přičemž při výpočtu funkce křížového ověřování používáme optimální jádra z třídy $\mathcal{M}_{0,k}$, resp. $\mathcal{M}_{1,k}$. Pomocí (17) či (18) pak určíme hodnotu $\hat{h}^*_{opt,\nu,k}$.

Pro určení závislosti kvality odhadu $\hat{m}^{(\nu)}$ na řádu jádra k zavedeme optimalizační kritérium

$$L(k) = T(K_{opt})\frac{2k+1}{2(k-\nu)n\hat{h}^{*\,2\nu+1}_{opt,\nu,k}}, \tag{22}$$

které dostaneme z (16) vynecháním členu $\bar{v}$, který je pro daná data konstantní a nemá tedy vliv na porovnání kvality odhadu.

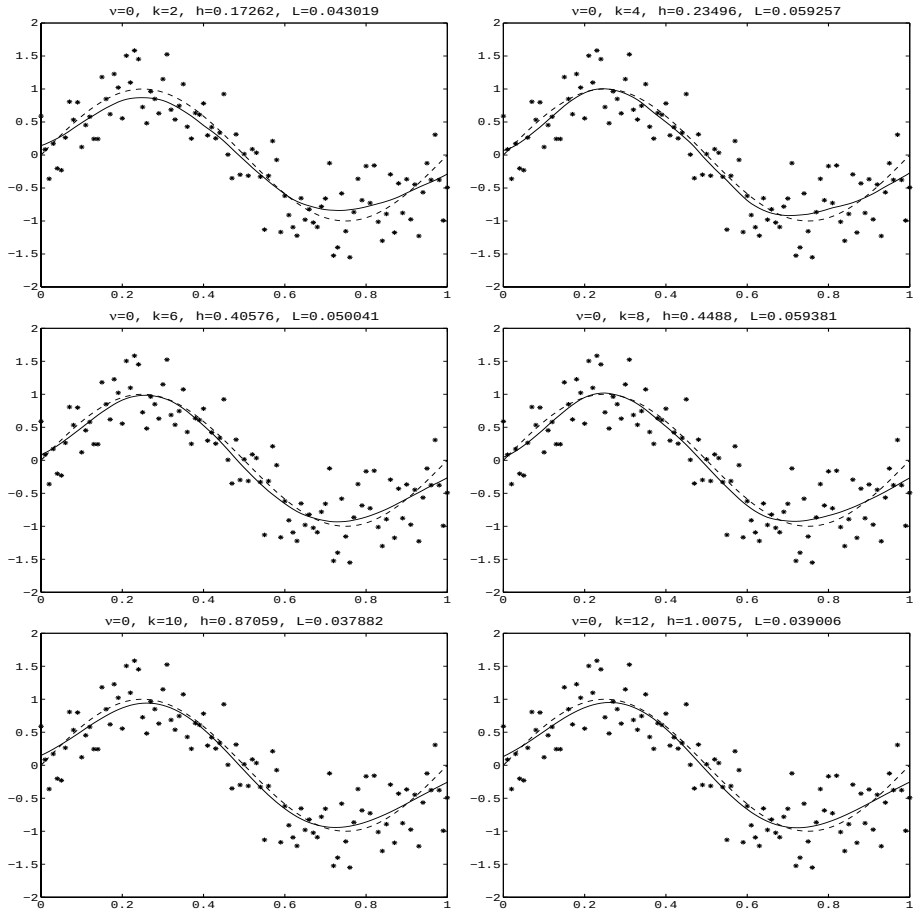
Nyní můžeme navrhnout postup pro stanovení optimálního řádu jádra k pro odhad ν-té derivace regresní funkce $m^{(\nu)}$. Nejprve určíme množinu I_ν těch hodnot k, pro něž chceme odhady konstruovat. Zpravidla klademe $I_\nu = \{\nu+2, \nu+4, \ldots, k_{max}\}$. Pak pro každé $k \in I_\nu$ najdeme optimální jádro K_{opt} (např. z uvedených tabulek) a vypočítáme hodnotu funkcionálu $T(K_{opt})$. Potom pomocí metody křížového ověřování určíme $\hat{h}^*_{opt,0,k}$ nebo $\hat{h}^*_{opt,1,k}$ a odtud $\hat{h}^*_{opt,\nu,k}$. Nakonec najdeme takové $\hat{k}$, pro které $L(k)$ nabývá minimální hodnoty:

$$\hat{k} = \arg\min_{k\in I_\nu} L(k).$$

Dá se ukázat (viz [4]), že odhad, který získáme tímto způsobem je asymptoticky optimální ve smyslu $\overline{AMSE}$ mezi všemi možnými volbami vyhlazovacího parametru h, tvaru a řádu jádra.

Následující obrázky ukazují aplikaci navrženého postupu pro stanovení optimálního odhadu derivace regresní funkce na simulovaná data. Jedná se o funkci $m(x) = \sin 2\pi x$, $x_i = i/100$ pro $i = 0, \ldots, 100$, měli jsme tedy celkem 101 hodnot. Chyby ε_i měly normální rozdělení s nulovou střední hodnotou a rozptylem 0.16. Obrázky jsou doplněny tabulkami spočítaných hodnot $L(k)$, $\hat{h}_{opt,\nu,k} = \delta_0\,\hat{h}^*_{opt,\nu,k}$ a optimálních hodnot $h_{opt,\nu,k}$ určených podle (8).

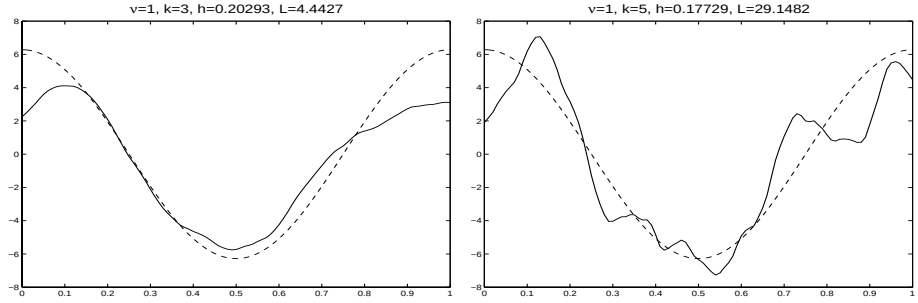
Odhad funkce $\sin 2\pi x$:

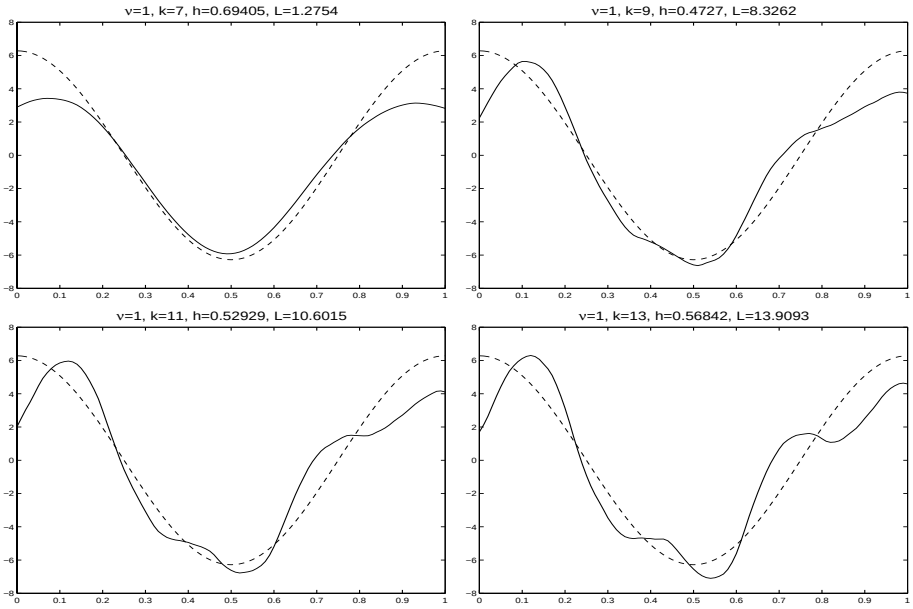


k	2	4	6	8	10	12
$L(k)$	0.0430	0.0593	0.0500	0.0594	0.0379	0.0390
$\hat{h}_{opt,\nu,k}$	0.1726	0.2350	0.4058	0.4488	0.8706	1.0075
$h_{opt,\nu,k}$	0.1252	0.3344	0.5580	0.7864	1.0170	1.2486

Optimální hodnota: $\hat{k} = 10$

Odhad první derivace funkce $\sin 2\pi x$:

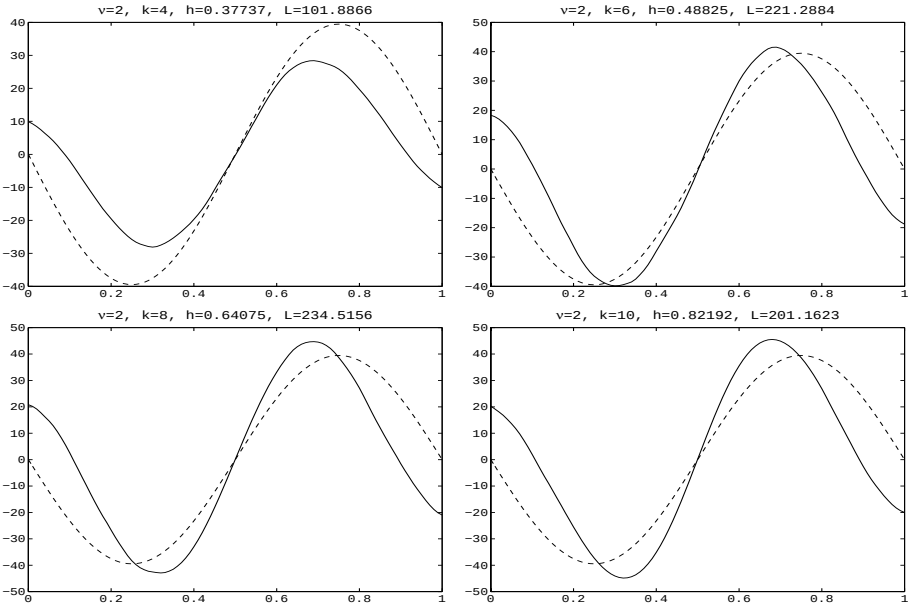


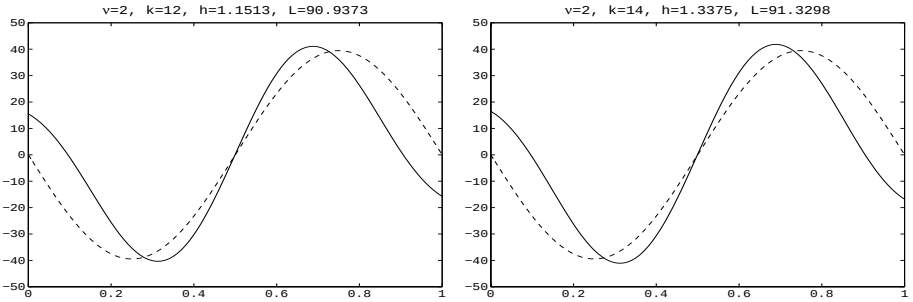


k	3	5	7	9	11	13
$L(k)$	4.4427	29.1482	1.2754	8.3262	10.6015	13.9093
$\hat{h}_{opt,\nu,k}$	0.2029	0.1773	0.6940	0.4727	0.5293	0.5684
$h_{opt,\nu,k}$	0.2066	0.4295	0.6589	0.8905	1.1231	1.3563

Optimální hodnota: $\hat{k} = 7$

Odhad druhé derivace funkce $\sin 2\pi x$:





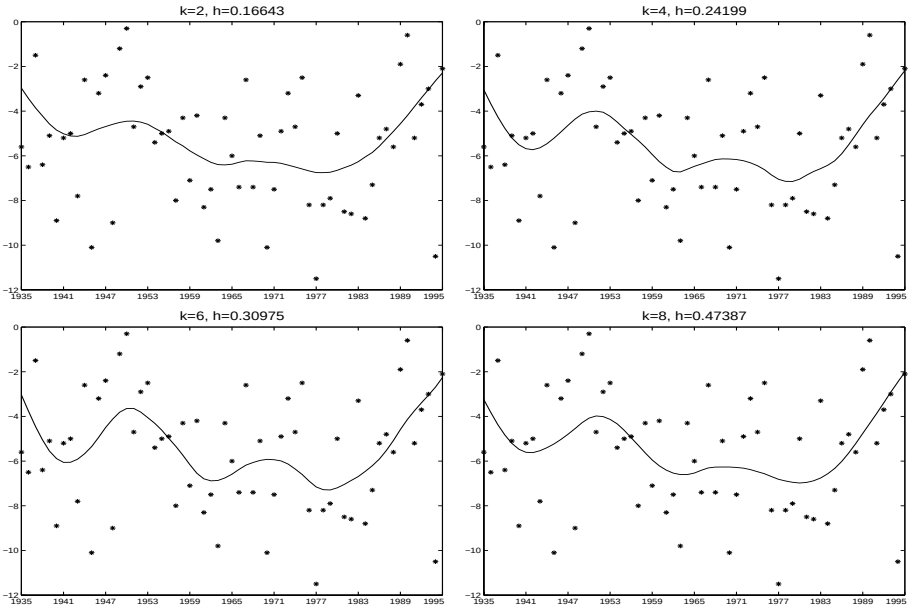
k	4	6	8	10	12	14
$L(k)$	101.8866	221.2884	234.5156	201.1623	90.9373	91.3298
$\hat{h}_{opt,\nu,k}$	0.3774	0.4883	0.6407	0.8219	1.1513	1.3375
$h_{opt,\nu,k}$	0.2983	0.5305	0.7642	0.9982	1.2325	1.4668

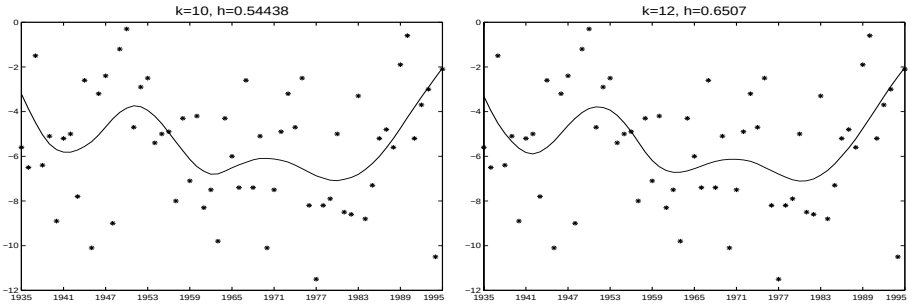
Optimální hodnota: $\hat{k} = 12$

Z obrázků je vidět, že odhad první derivace dopadl hůř než odhad regresní funkce nebo její druhé derivace. Je to pravděpodobně způsobeno tím, že optimalizace šířky okna pomocí metody křížového ověřování dává pro první derivaci regresní funkce horší výsledky než pro regresní funkci samotnou.

Obrázky také ukazují typický efekt pro jádrové vyhlazování: horší kvalitu odhadu na krajích intervalu – tzv. hraniční nebo okrajový efekt. Ten je způsoben tím, že blízko krajních bodů daného intervalu nosič jádrové funkce zasahuje do oblasti, kde nejsou žádná data, což zhoršuje odhad. Tento problém je možno řešit např. použitím hraničních jader (viz [3]).

Následující obrázky se týkají reálných dat. Jedná se o průměrné lednové teploty v Delhi v Kanadě v letech 1935–1995. Připojená tabulka opět ukazuje hodnoty $L(k)$ a $\hat{h}_{opt,0,k}$.





k	2	4	6	8	10	12
$L(k)$	0.0739	0.0953	0.1085	0.0931	0.1003	0.1000
$\hat{h}_{opt,0,k}$	0.1664	0.2420	0.3097	0.4739	0.5444	0.6507

Optimální hodnota: $\hat{k} = 2$

Gasser-Müllerův odhad použitý v tomto příspěvku je vhodný jen pro model s pevným plánem, a proto může být pro značnou část reálných dat nepoužitelný. To však není omezení v případě, že potřebujeme jenom odhad regresní funkce m a nikoliv některé její derivace. V takovém případě totiž lze v podstatě stejným způsobem použít tvar odhadu vhodnější pro data s náhodným plánem, např. Nadaraya-Watsonův nebo lokálně lineární odhad.

Poděkování: Autoři děkují recenzentovi za podnětné připomínky, které přispěly ke zlepšení kvality příspěvku. Dále děkují panu Mgr. J. Koláčkovi a panu Mgr. M. Řezáčovi za pečlivé přečtení rukopisu.

Literatura

[1] Hastie T.S,Tibshirami,R.J.: Generalized Additive Models, Chapman& Hall, London, 1990
[2] Härdle,W.: Applied Nonparametric Regression. Cambridge University Press, 1990
[3] Horová I., Optimization Problems Connected with Kernel Smoothig, Signal Processing, Communications and Computer Science, World Scientific and Engeneering Society Press, 2000
[4] Horová I.,Vieu P., Zelinka J.: Optimal Choice of Nonparametric Estimates of a Density and of its Derivatives. Preprint
[5] Müller,H.-G.: Nonparametric Regression Analysis of Longitudinal data. Lecture Notes in Statistics 46. Springer-Verlag, 1988
[6] Wand,M.P.,Jones,M.C.: Kernel Smoothing. Chapman& Hall, London, 1995

MU PřF, KAM, Janáčkovo nám. 2a, 662 95 Brno
E-mail: zelinka@math.muni.cz; horova@math.muni.cz