

GASSER-MÜLLERŮV ODHAD

J. Poměnková

Došlo: 18. listopadu 2002

Abstract

POMĚNKOVÁ, J.: *The Gasser-Müller estimator*. Acta univ. agric. et silvic. Mendel. Brun., 2004, LII, No. 3, pp. 167-176

Kernel smoothing provides a simple way for finding structure in data. The idea of the kernel smoothing can be applied to a simple fixed design regression model and a random design regression model. This article is focused on kernel smoothing for fixed design regression model with using special type of estimator, the Gasser-Müller estimator, and on choice of the bandwidth. At the end of this article figures for illustration described methods on two data sets are shown. The first data set contains simulated values of function $\sin(2\pi x)$, the second contains January average temperatures measured in Basel 1755–1855.

kernel smoothing, Gasser-Müller estimator, bandwidth

1. ZÁKLADNÍ DEFINICE A VĚTY

Nechť $(x_i, Y_i)_{i=1}^n$, $n \in N$ je posloupnost pozorování dvojice (x, Y) , kde $x \in R$ a Y je reálná nebo simulovaná náhodná veličina. Jestliže hodnoty nezávislé proměnné x jsou pevně zvolené experimentátorem (tzn. nejsou náhodné veličiny), pak hovoříme o tzv. modelu s pevným plánem. Závislost hodnot Y na hodnotách x lze popsat tzv. regresní křivkou a příslušný funkční vztah mezi proměnnými (x, Y) , nazývaný regresní model, zapisujeme ve tvaru

$$Y_i = m(x_i) + \epsilon_i \quad i = 1, \dots, n, \quad (1)$$

kde m je neznámá funkce, kterou budeme odhadovat, x_i je bod plánu, Y_i je pozorování a platí

$$\begin{aligned} E(\epsilon_i) &= 0, & i &= 1, \dots, n \\ D(\epsilon_i) &= \sigma^2 > 0, & i &= 1, \dots, n, \end{aligned}$$

Hodnoty x_i , $i = 1, \dots, n$ můžeme zadat nezápornou integrovatelnou funkcí f

$$\int_{-\infty}^{+\infty} f(x) dx = 1; \quad \int_{-\infty}^{x_i} f(x) dx = \frac{i-1}{n-1}; \quad i = 1, \dots, n. \quad (2)$$

Označme $Lip[a, b]$ třídu spojitých funkcí splňujících Lipschitzovu podmínku na $[a, b]$:

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

Konstrukci jádrového odhadu funkce $\hat{m}(x)$ budeme provádět v bodech plánu x_i , $i = 1, \dots, n$, kde $x_i = \frac{i}{n}$, $0 \leq x_i \leq 1$. Zavedme si nyní definici jádra.

Definice 1.1.: Necht' v, k jsou nezáporná celá čísla, $0 \leq v < k$. Necht' $K \in Lip[-1, 1]$, nosič $(K) = [-1, 1]$. Necht' K splňuje následující momentové podmínky

$$\int_{-1}^1 x^j K(x) dx = \begin{cases} 0 & 0 \leq j < k, j \neq v \\ (-1)^v v! & j = v, \\ \beta_k & j = k. \end{cases} \quad (3)$$

Funkce s těmito vlastnostmi se nazývá jádro. Je-li

$\beta_k \neq 0$ říkáme, že jádro K je řádu (v, k) a píšeme $K \in \mathcal{M}_{v,k}$.

Definice 1.2.: Necht' $\mu \in \mathbb{N}$, $k, v + k$ je sudé číslo a necht' K je jádro splňující následující podmínky

1) $K \in C^\mu[-1, 1]$,

2) $K \in \mathcal{M}_{v,k}$,

3) $K^{(j)}(1) = K^{(j)}(-1) = 0, j = 0, 1, \dots, \mu - 1, \mu \geq 1, 0 \leq v \leq k - 2, v + k$ sudé.

Toto jádro nazýváme jádro hladkosti μ a označujeme $K \in \mathcal{M}_{v,k}^\mu$.

Věta 1.3.:

1) Necht' K splňuje požadavky z výše uvedené definice, pak $K^{(\mu)} \in \mathcal{M}_{v+\mu, k+\mu}, \mu \geq 1$

2) Pro libovolnou funkci $\psi \in \mathcal{M}_{v+\mu, k+\mu}$ existuje právě jedna funkce $q \in C^\mu[-1, 1]$ taková, že $q^{(\mu)} = \psi, q^{(j)}(-1) = 0, j = 0, 1, \dots, \mu - 1, q \in \mathcal{M}_{v,k}$.

Z obecného hlediska lze jádrový odhad zapsat jako $\hat{m}(x) = \sum_{i=1}^n W_i(x, h) Y_i$, kde $W_i(x, h)$ označuje váhovou funkci. Označme $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, $h = h_n$ je kladná konstanta.

Omezíme-li se na odhady v bodech plánu, pak můžeme jádrový odhad zapsat jako

$$\hat{\mathbf{m}} = \begin{pmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{pmatrix} = \mathbf{S}\mathbf{Y} = \begin{pmatrix} W(x_1, h) & \dots & W_1(x_n, h) \\ \vdots & \ddots & \vdots \\ W_n(x_1, h) & \dots & W_n(x_n, h) \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

kde $\hat{\mathbf{m}} = (\hat{m}(x_1), \dots, \hat{m}(x_n))^T$ je vektor odhadů funkce m , $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ je vektor pozorování, $\mathbf{x} = (x_1, \dots, x_n)^T$ je vektor bodů plánu a $\mathbf{S} = (s_{ij}), i, j = 1, \dots, n$ je vyhlazovací matice. Tato matice je dána tzv. váhovými funkcemi, které budeme v našem případě označovat $S = (s_{ij}), s_{ij} = W_i(x_j, h)$ a závisí na h, i, x a K ; h je šířka vyhlazovacího okna a řídí hladkost odhadu.

2. GASSER-MÜLLERŮV ODHAD

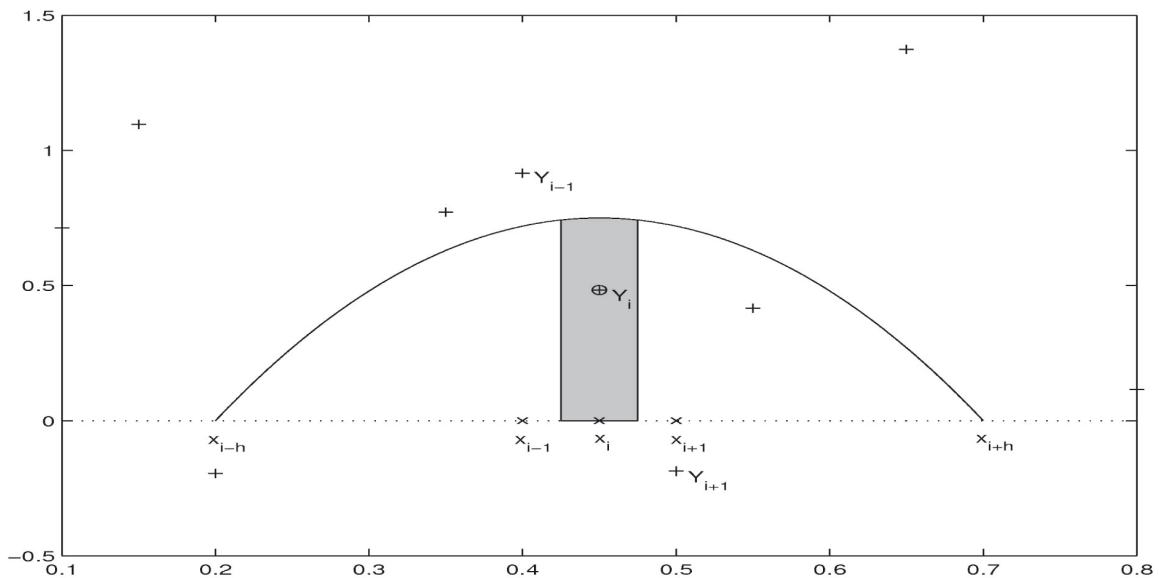
Mějme regresní model popsáný v 1. kapitole. Gasser-Müllerův odhad definujeme následovně

$$\hat{m}(x) = \sum_{i=1}^n Y_i \frac{1}{h} \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du = \sum_{i=1}^n Y_i W_i(x, h). \quad (4)$$

Body plánu $x_i, i = 1, \dots, n$ jsou vzestupně uspořádané a pro body $s_i, i = 1, \dots, n$ platí $s_0 = 0, s_i = \frac{x_{i-1} + x_i}{2}, i = 1, \dots, n-1$ a $s_n = 1$. Pro prvky vyhlazovací matice $s_{ij} = W_i(x_j, h), i, j = 1, \dots, n$ v bodě plánu x_j s šířkou vyhlazovacího okna h platí vztah

$$W_i(x_j, h) = \frac{1}{h} \int_{s_{i-1}}^{s_i} K\left(\frac{x_j - u}{h}\right) du. \quad (5)$$

V další části uvedeme vztah pro odhad derivace funkce m .



Graf ukazuje konstrukci váhy pro Gasser-Müllerův odhad na příkladě simulovaných dat funkce $f(x) = \sin(2\pi x)$. Měření jsou označena symboly $+$, Y_i je měření v bodě x_i a parabola je Epanečnikovo jádro $K(x) = \frac{3}{4}(1-x^2)$. Šedě vyznačená oblast je po přenásobení koeficientem $1/h_n$ váha odpovídající bodu plánu, měření a Epanečnikovu jádru v konkrétním případě.

Věta 2.1.: *Nechť je dán regresní model s pevným plánem a nechť dále platí*

1. $m \in C^k([0, 1])$, $k \in \mathbb{N}$

2. $K \in \mathcal{M}_{v,k}$

3. $\lim_{n \rightarrow \infty} h_n = 0$ a $\lim_{n \rightarrow \infty} nh_n^{2v+1} = \infty$

Pak pro odhad v -té derivace funkce m , a každé $x \in (0, 1)$ je

$$\begin{aligned} \hat{m}^{(v)}(x) &= \frac{1}{h^{v+1}} \sum_{i=1}^n Y_i W_i(x) = \\ &= \frac{1}{h^{v+1}} \sum_{i=1}^n Y_i \int_{s_i-1}^{s_i} K\left(\frac{x-u}{h}\right) du \end{aligned} \quad (6)$$

konzistentním odhadem $m^{(v)}(x)$. Pro rozptyl tohoto odhadu platí

$$\text{var}(\hat{m}^{(v)}(x)) = \frac{\sigma^2}{nh^{2v+1}} (C_K + o(1)), \quad (7)$$

kde

$$C_K = \int_{-1}^1 K^2(x) dx, \quad (8)$$

a pro jeho vychýlení platí

$$E\hat{m}^{(v)}(x) - m^{(v)}(x) = h^{k-v} m^{(k)}(x) B_K + O([nh]^{-1}) + o(h^{k-v}), \quad (9)$$

kde

$$B_K = (-1)^k \frac{\beta_k}{k!}, \quad \beta_k = \int_{-1}^1 x^k K(x) dx. \quad (10)$$

3. VOLBA OPTIMÁLNÍ ŠÍŘKY VYHLAZOVACÍHO OKNA

Vhodným nástrojem pro posouzení kvality odhadu je střední kvadratická chyba.

Vyjdeme proto z tvaru střední kvadratické chyby $MSE(\hat{m}^{(v)}(x))$, který je dán vztahem

$$MSE(\hat{m}^{(v)}(x)) = E(\hat{m}^{(v)}(x) - m^{(v)}(x))^2,$$

a který lze zapsat také ve tvaru

$$MSE(\hat{m}^{(v)}(x)) = \text{var}(\hat{m}^{(v)}(x)) + E(\hat{m}^{(v)}(x) - m^{(v)}(x))^2$$

Nechť platí předpoklady věty 2.1., potom střední

kvadratickou chybu $MSE(\hat{m}^{(v)}(x))$ můžeme vyjádřit ve tvaru

$$MSE(\hat{m}^{(v)}(x)) \approx \frac{\sigma^2 C_K}{nh^{2v+1}} + h^{2(k-v)} [m^{(k)}(x) B_K]^2, \quad (11)$$

kde

$$B_K = (-1)^k \frac{\beta_k}{k!}, \quad \beta_k = \int_{-1}^1 x^k K(x) dx, \quad C_K = \int_{-1}^1 K^2(x) dx.$$

Šířka vyhlazovacího okna v bodě x ($h_{MSE(\hat{m}^{(v)}(x))}$), která minimalizuje $MSE(\hat{m}^{(v)}(x))$ je dána vztahem

$$\frac{\partial MSE(\hat{m}^{(v)}(x))}{\partial h} = 0$$

a nazýváme ji optimální šířkou. Jde však o asymptotickou optimalitu mezi všemi možnými volbami hodnoty h , zvoleným typem a řádem jádra K . Obdržíme

$$h_{MSE(\hat{m}^{(v)}(x))} \approx \left[\frac{2v+1}{2(k-v)} \cdot \frac{\sigma^2 C_K}{B_K^2 (m^{(k)}(x))^2} \cdot \frac{1}{n} \right]^{\frac{1}{2k+1}} \quad (12)$$

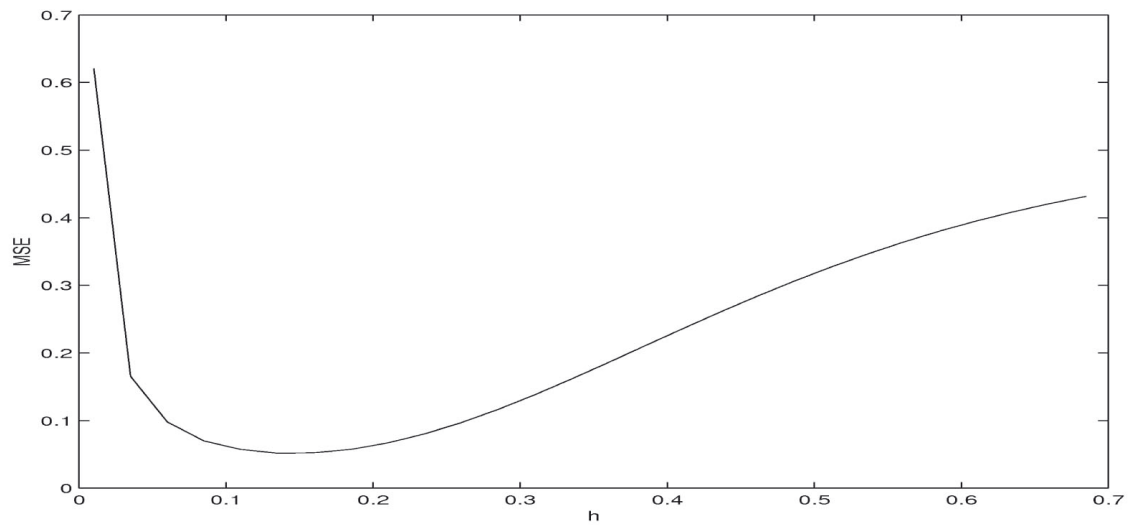
Ze vzorců (11) a (12) můžeme vidět, že s rostoucím h klesá rozptyl odhadu a roste kvadrát jeho vychýlení a naopak. Ze vzorce (12) je také patrné, že tento vztah lze použít pro nalezení vhodné šířky vyhlazovacího okna, když známe odhadovanou funkci. Analogicky lze získat šířku vyhlazovacího okna pro celý interval vyhlazované funkce z průměrné střední kvadratické chyby $\overline{AMSE}$. Jestliže odhadovanou funkci neznáme, nalezneme řešení vhodné šířky vyhlazovacího okna pomocí metody křížového ověřování. V případě této metody půjde o odhad optimální hodnoty.

Vyjděme opět z věty 2.1., z předpokladů a tvaru vychýlení a rozptylu pro odhad $\hat{m}^{(v)}(x)$. Označme $\hat{\mathbf{m}}^{(v)} = (\hat{m}^{(v)}(x_1), \dots, \hat{m}^{(v)}(x_n))^T$. Pak průměrnou střední kvadratickou chybu můžeme vyjádřit ve tvaru

$$\begin{aligned} \overline{AMSE}(\hat{\mathbf{m}}^{(v)}, h) &= \frac{1}{n} \sum_{j=1}^n E(\hat{m}^{(v)}(x_j) - m^{(v)}(x_j))^2 \approx \\ &\approx \frac{1}{n} \sum_{j=1}^n \left[\frac{C_K \cdot \sigma^2}{nh^{2v+1}} + h^{2(k-v)} B_K^2 (m^{(k)}(x_j))^2 \right] \end{aligned}$$

a C_K, B_K jsou definovány ve větě 2.1. Optimalizací získáme vyjádření pro optimální šířku vyhlazovacího okna ve tvaru

$$h_{opt, \overline{AMSE}} \approx \left[\frac{(2v+1)C_K}{2n(k-v)B_K^2 (\bar{m}^{(k)})^2} \right]^{\frac{1}{2k+v+1}} \quad (13)$$

2: Závislost $\overline{AMSE}$ na h **Metoda křížového ověřování**

Chceme-li tedy odhadnout šířku vyhlazovacího okna h , vypočteme hodnotu funkce křížového ověřování

$$CV(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_{-i}(x_i))^2, \quad (14)$$

kde $\hat{m}_{-i}(x_i)$ je označení odhadu v bodě x_i , při kterém jsme vynechali bod x_i a odpovídající hodnotu Y_i , přičemž hledáme minimum této funkce na množině $H_n = [a_k n^{-\frac{1}{2k+1}}, b_k n^{-\frac{1}{2k+1}}]$, pro nějaké $0 < a_k < b_k < \infty$, tj.

$$\hat{h}_{optCV} = \arg \min_{h \in H_n} CV(h). \quad (15)$$

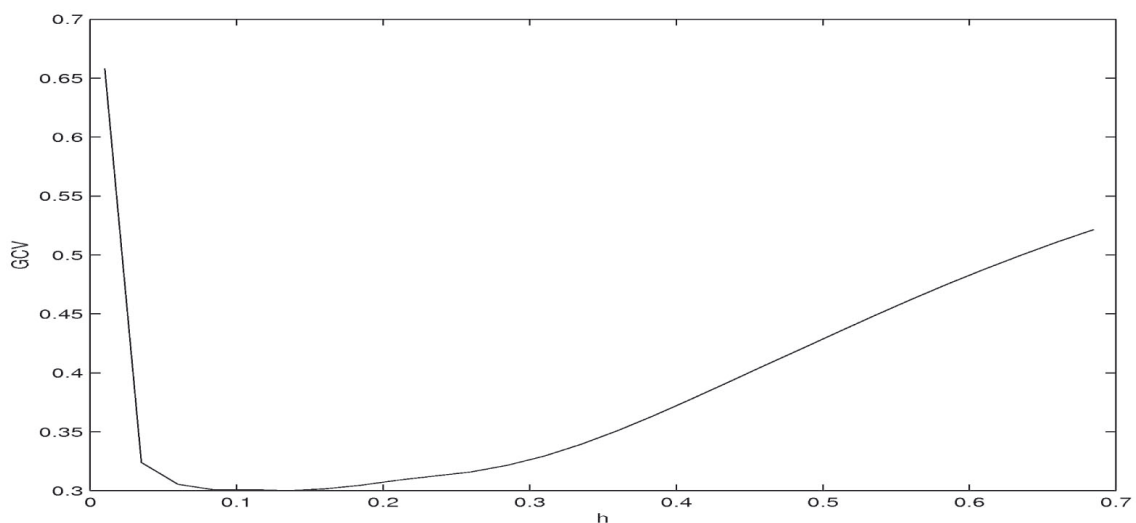
Vztah (14) lze zjednodušit a zobecnit do vhodnějšího tvaru

$$GCV(h) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{Y_i - \hat{m}(x_i)}{1 - \text{tr}(W)/n} \right\}^2, \quad (16)$$

kde

$\text{tr}(W) = \text{trace}(W) = \sum_{i=1}^n W_i(x, h)$, kde W je definováno v (5).

Upravená metoda, která využívá vztahu (16) se nazývá zobecněná metoda křížového ověřování.

3: Závislost GCV na h

Ukázka typů jader

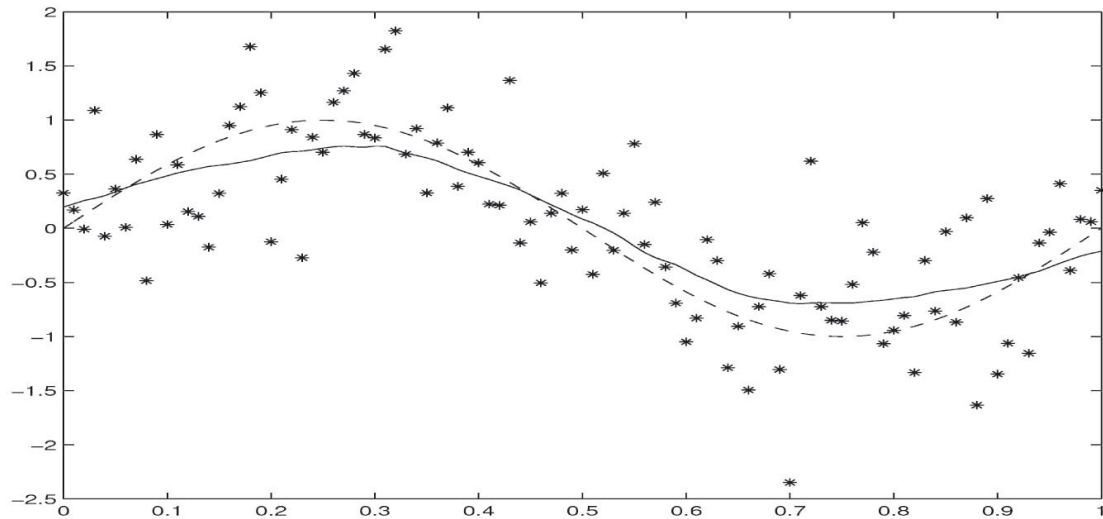
ν	k	μ	jádro
0	2	1	$\frac{3}{4}(1-x^2)$
0	2	2	$\frac{15}{16}(1-2x^2+x^4)$
0	4	0	$\frac{3}{8}(3-5x^2)$
0	4	1	$\frac{15}{32}(3-10x^2+7x^4)$
0	4	2	$\frac{105}{64}(1-5x^2+7x^4-3x^6)$
1	3	0	$\frac{3}{2}x$
1	3	1	$\frac{15}{4}(x-x^3)$
1	3	2	$\frac{105}{16}(-x+2x^3-x^5)$
1	5	0	$\frac{15}{8}(-5x+7x^3)$
1	5	1	$\frac{105}{32}(-5x+14x^3-9x^5)$
1	5	2	$\frac{315}{64}(-5x+21x^3-27x^5+11x^7)$

5. APLIKACE GASSER-MÜLLEROVA ODHADU

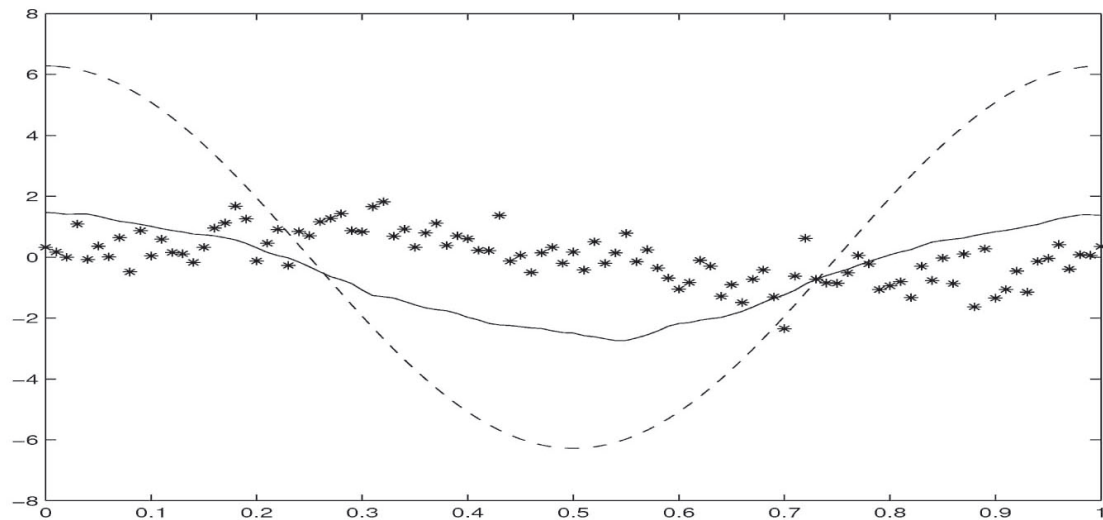
V této části je provedena aplikace poznatků jádrového vyhlazování s užitím Gasser-Müllerova estimátoru na konkrétní data. V první části jde o aplikaci na simulované hodnoty, ve druhé na reálná data. Pro simulaci byla vybrána známá funkce $\sin(2\pi x)$ o rozsahu $n = 101$, $\sigma^2 = 0,3$. Reálná data tvořil soubor průměrných lednových teplot naměřených v Basileji mezi lety 1755-1855.

5.1. Simulovaná data

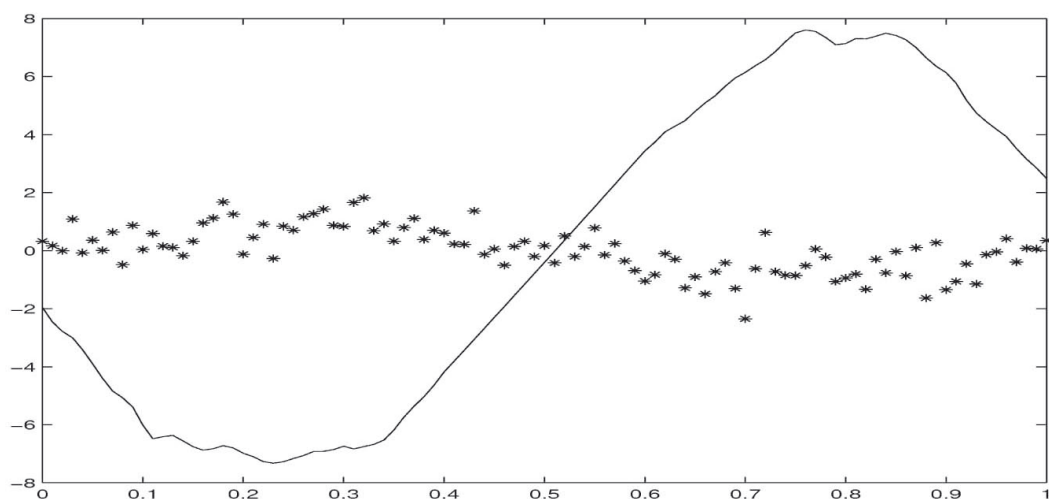
Pro ukázkou jádrového vyhlazování byla simulována data funkce $m = \sin(2\pi x)$.



4: Odhad $\hat{m}$, $K \in \mathcal{M}_{0,4}$, $h = 0,4$

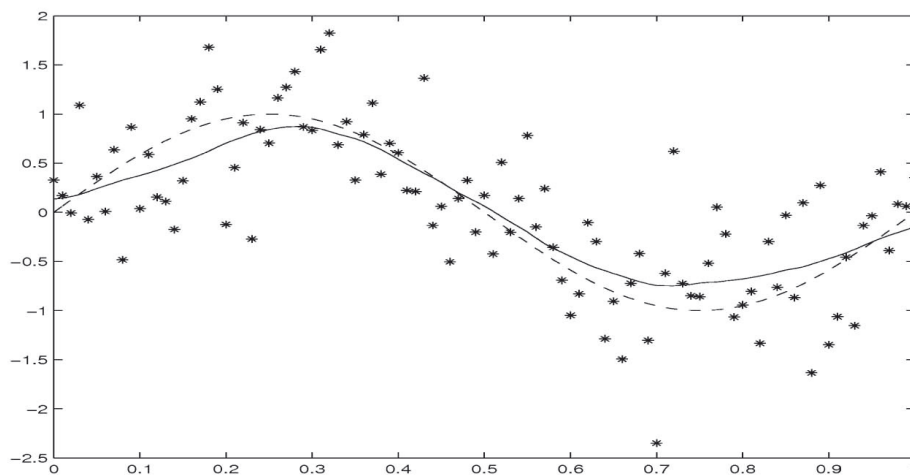


5: Odhad $\hat{m}^{(1)}$, $K \in \mathcal{M}_{1,3}$, $h = 0,5$

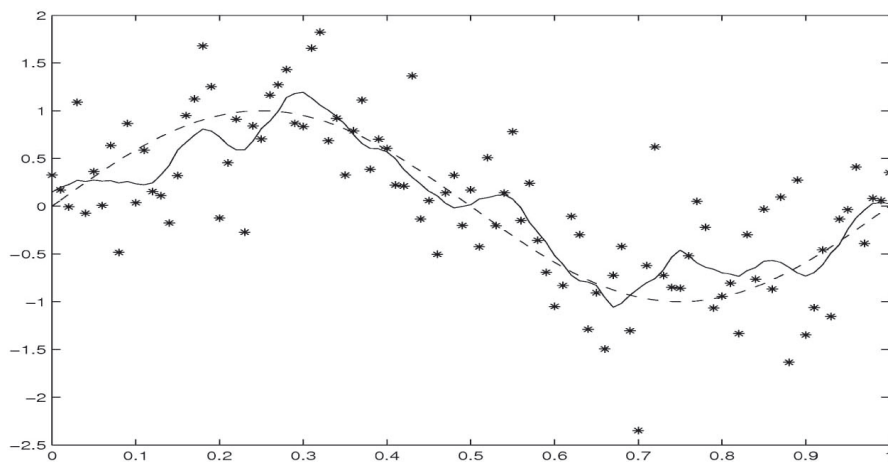


6: Odhad $\hat{m}^{(2)}$, $K \in \mathcal{M}_{2,4}$, $h = 0.4$

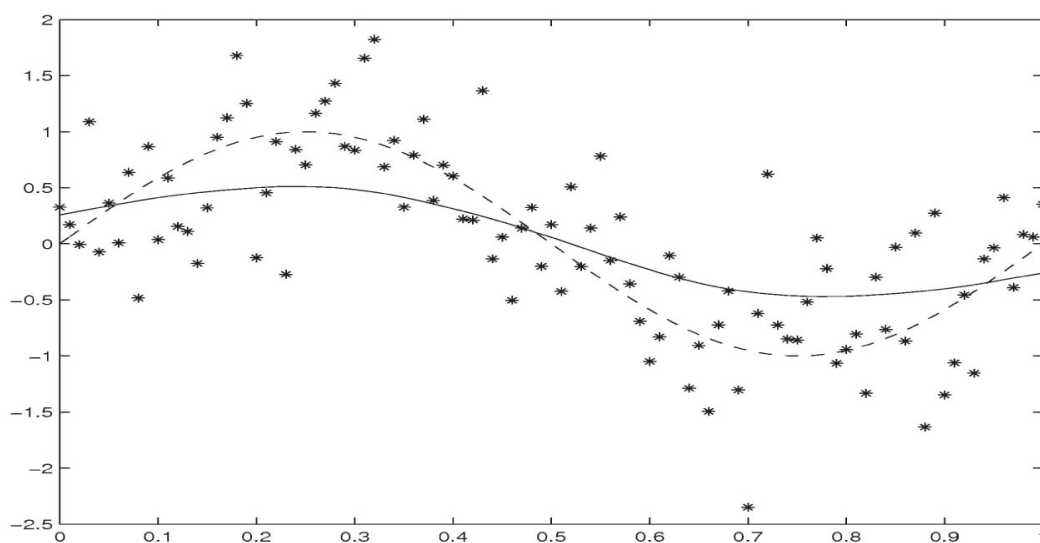
5.2. Vliv výběru šířky vyhlazovacího okna na simulovaná data



7: Odhad funkce s optimální hodnotou šířky vyhlazovacího okna



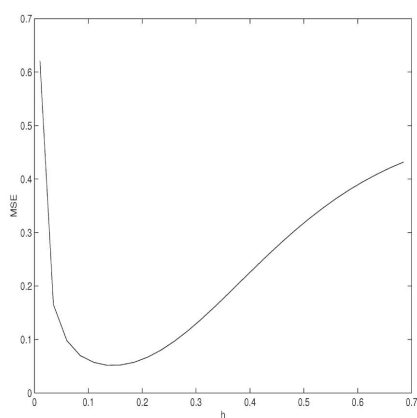
8: Odhad funkce s malou hodnotou šířky vyhlazovacího okna



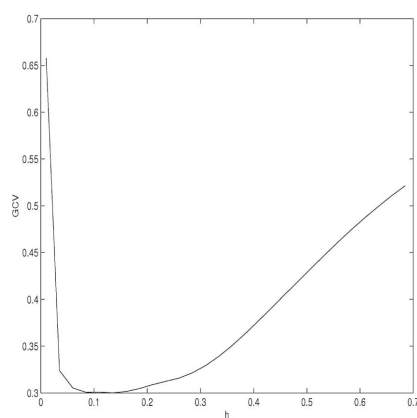
9: Odhad funkce s velkou hodnotou šířky vyhlazovacího okna

5.3. Výběr optimální hodnoty h

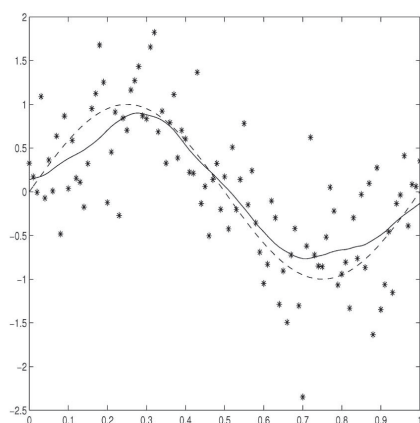
Užitím $\overline{AMSE}$ a metody *Cross-validation*



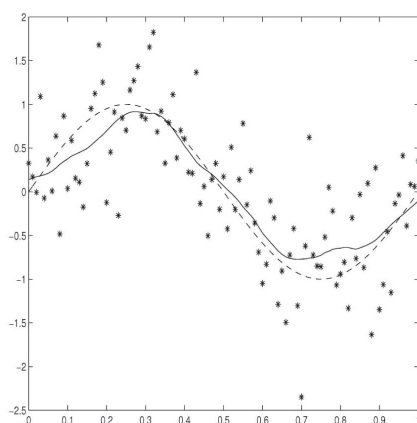
10.1: Závislost $\overline{AMSE}$ na h



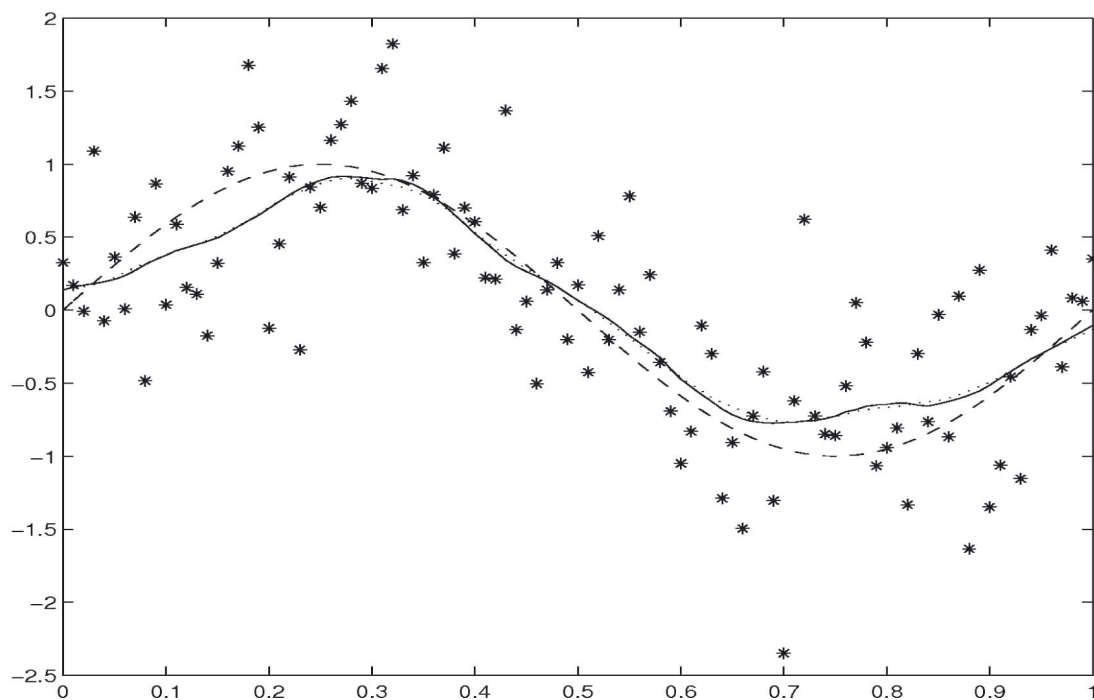
11.1: Závislost GCV na h



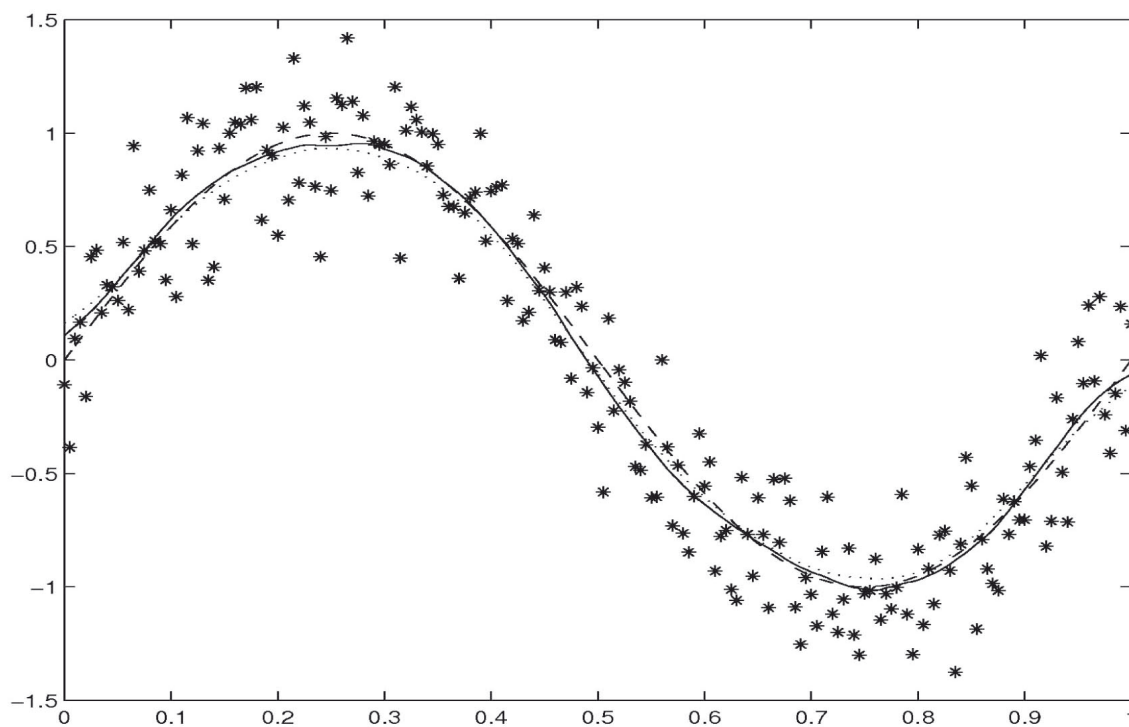
10.2: Odhad funkce m s $h_{opt, \overline{AMSE}}$



11.2: Odhad funkce m s $\hat{h}_{optGCV}$



12: Odhad funkce $\sin(2\pi x)$ (čárkovaně) s h_{AMSE} (tečkovaně) a $\hat{h}_{opt,GCV}$ (plnou čarou), $n = 101$

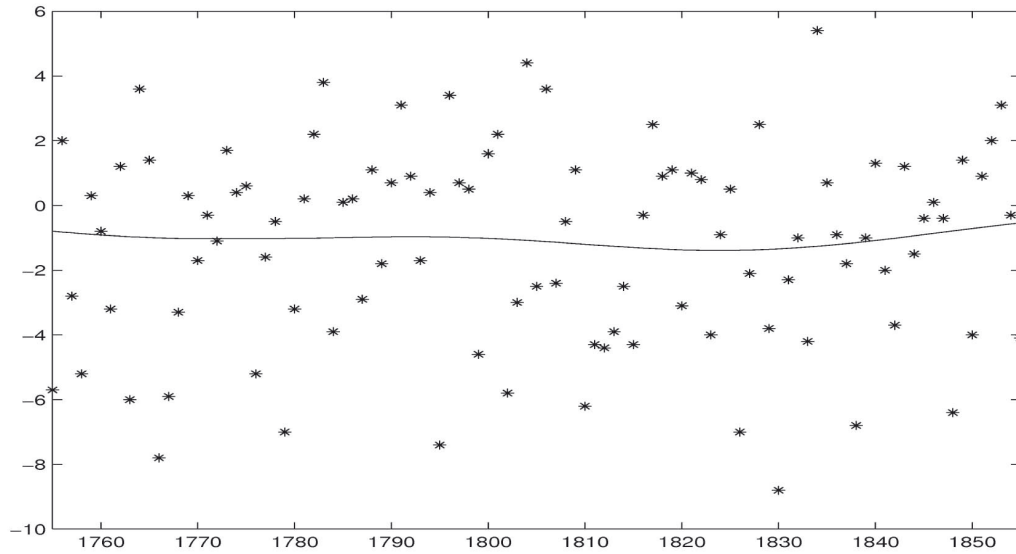


13: Odhad funkce $\sin(2\pi x)$ (čárkovaně) s h_{AMSE} (tečkovaně) a $\hat{h}_{opt,GCV}$ (plnou čarou), $n = 201$

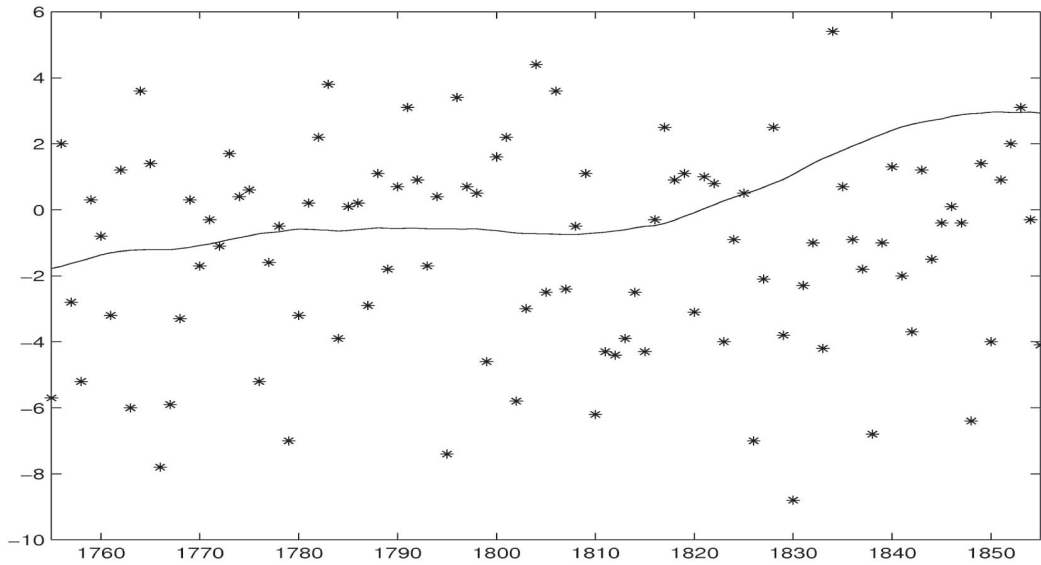
5.4. Aplikace na reálná data

Pro aplikace jádrového vyhlazování na reálných datech byly vybrány průměrné lednové teploty na-

měřené v Basileji mezi lety 1755–1855. Tyto hodnoty byly rovněž otestovány Kendalovým τ testem náhodnosti a jsou nezávislé.



14: Odhad průměrných lednových teplot, $K \in \mathcal{M}_{0,2}$, $h_{optGCV} = 0.45$



15: Odhad první derivace funkce popisující průměrné lednové teploty, $K \in \mathcal{M}_{0,2}$, $h_{optGCV} = 0.4$

6. ZÁVĚR

Předkládaný článek se věnuje problematice popsání struktury dat, a to prostřednictvím jádrového vyhlazování s užitím Gasser-Müllerova odhadu. Dále se zaměřuje na problematiku volby šířky vyhlazovacího okna, která významnou měrou ovlivňuje kvalitu výsledného odhadu. V programovém prostředí MATLAB byl rovněž zpracován výpočtový program a s jeho pomocí byly následně popsány dvě struktury dat. První byla tvořena simulovanými hodnotami

funkce $\sin(2\pi x)$, druhá průměrnými lednovými teplotami naměřenými v Basileji mezi lety 1755–1855. Získané výsledky jsou graficky znázorněny. Graf číslo 15 ukazuje nalezenou křivku, která popisuje odhad vývoje průměrných lednových teplot naměřených v Basileji mezi lety 1755–1855. Graf číslo 16 ukazuje nalezenou křivku, která popisuje odhad první derivace vývoje lednových teplot naměřených v Basileji mezi lety 1755–1855.

SOUHRN

Jádrové vyhlazování nabízí jednoduchý způsob jak najít a popsat strukturu dat. Myšlenka jádrového vyhlazování může být aplikována na regresní model s pevným i náhodným plánem. Předložený článek se zaměřil na jádrové vyhlazování s užitím speciálního typu odhadu, a to Gasser-Müllerova odhadu, a na výber šířky vyhlazovacího okna. Na konci článku jsou připojeny ilustrativní grafy uvedených metod. Pro grafické zpracování bylo využito jak simulovaných hodnot funkce $\sin(2\pi x)$, tak reálných hodnot, kterými byly průměrné lednové teploty naměřené v Basileji mezi lety 1755–1855.

jádrové vyhlazování, Gasser-Müllerův odhad, šířka vyhlazovacího okna

LITERATURA

- | | |
|--|--|
| GASSER, T., EGEL, J.: The choice of weights in kernel regression estimation. <i>Biometrika</i> 1990. | statistické metody, Lázně Bohdaneč 2.11.1 - 24.11.2000, 141 - 166 str. |
| HOROVÁ, I.: Some Remarks on Kernels, <i>Journal of Computational Analysis</i> , Vol.2., No.2.2000. | HÄRDLE, W.: <i>Applied Nonparametric Regression</i> . Cambridge University Press, 1990. |
| HOROVÁ, I., ZELINKA, J.: Základy a aplikace jádrových odhadů. <i>Analýza dat 2000/II</i> . Moderní | HÄRDLE, W., SCHIMEK, M. G.: <i>Statistical Theory and Computational Aspects of Smoothing</i> 1994. |
| | WAND, M. P., JONES, M. S.: <i>Kernel Smoothing</i> . Chapman & Hall, London, 1995. |

Adresa

Mgr. Jitka Poměnková, Ústav statistiky a operačního výzkumu, Mendelova zemědělská a lesnická univerzita v Brně, Zemědělská 5, 613 00 Brno, Česká republika