O neparametrickém odhadu polohy maxima

Zdeněk Hlávka

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky

Robust 2010

Neparametrická regrese

Pozorujeme (X_i, Y_i), $i = 1, \ldots, n$, předpokládáme, že

$$Y_i = m(X_i) + \varepsilon_i,$$

kde $\operatorname{Var} \varepsilon_i = \sigma^2$ a chceme odhadnout neznámou funkci $m(.)$.

Nadaraya-Watsonův odhad

$$m_h^{NW}(x) = \frac{\sum_{i=1}^n K_h(x - X_i) Y_i}{\sum_{i=1}^n K_h(x - X_i)} = \sum_{i=1}^n \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)} Y_i,$$

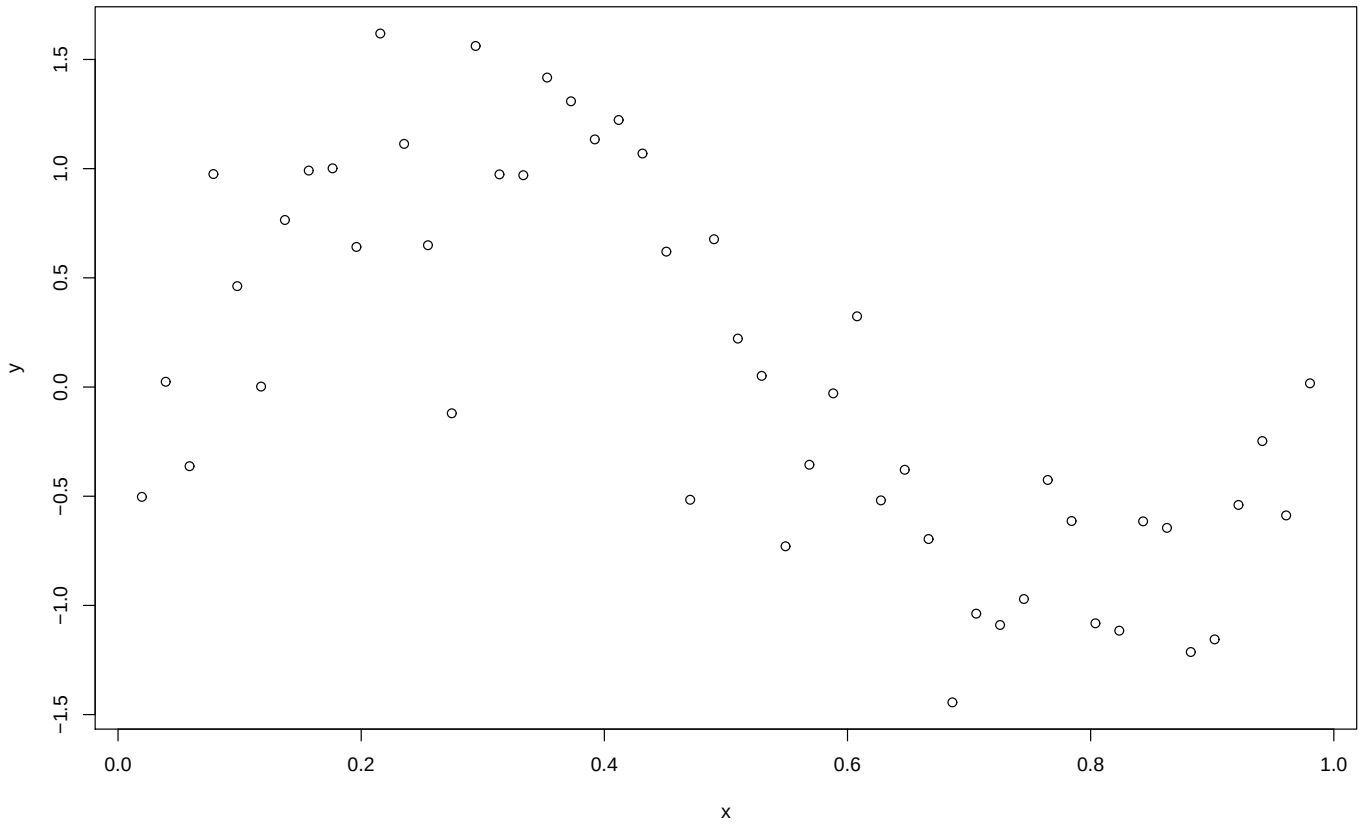
kde $K_h(x) = K(x/h)/h$, $K(.)$ je vhodná *jádrová funkce* a h je *bandwidth*.

Pokud mají X_i rovnoměrné rozdělení, pak "za jistých předpokladů":

$$\mathrm{MSE}\{m_h^{NW}(x)\} \approx \frac{1}{nh}\sigma^2 \int K^2(s)ds + \frac{1}{4}h^4\{m''(x)\}^2.$$

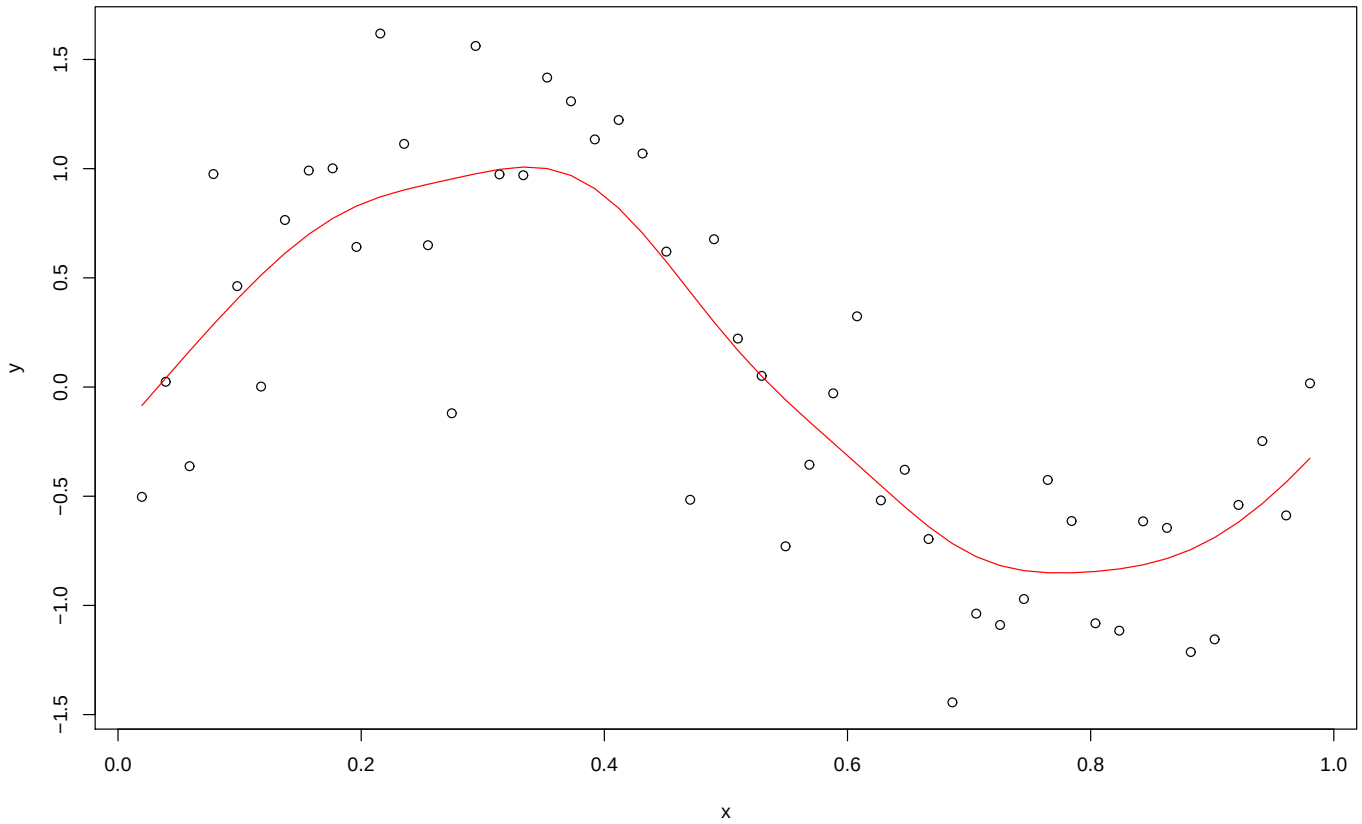
Flexibilita × horší vlastnosti než parametrické odhady

Obvyklé použití: průzkumová analýza dat, grafické znázornění, zkoumání závislosti na čase nebo geografické poloze.



Flexibilita × horší vlastnosti než parametrické odhady

Obvyklé použití: průzkumová analýza dat, grafické znázornění, zkoumání závislosti na čase nebo geografické poloze.



Neparametrická regrese

Nadaraya-Watsonův odhad:

$$m_h^{NW}(x) = \sum_{i=1}^{n} \frac{n^{-1}K_h(x - X_i)}{n^{-1}\sum_{j=1}^{n} K_h(x - X_j)} Y_i = \frac{\sum_{i=1}^{n} n^{-1}K_h(x - X_i)Y_i}{\widehat{f}_X(x)}$$

Pokud známe $f_X(x)$ (*pro jednoduchost předpokládejme rovnoměrné rozdělení* $R(0,1)$), můžeme použít **Gasser-Müllerův odhad**

$$m_h^{GM}(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(x - u)du Y_i,$$

kde $s_i = (x_{(i)} + x_{(i+1)})/2$ a $\int_{s_{i-1}}^{s_i} K_h(x - u) \approx n^{-1}K_h(x - x_{(i)})$.

Odhad "funguje" díky tomu, že $\sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(x - u)du = \int K(x)dx = 1$.

MSE

Nadaraya-Watsonův odhad

Pokud mají X_i rovnoměrné rozdělení, pak "za jistých předpokladů":

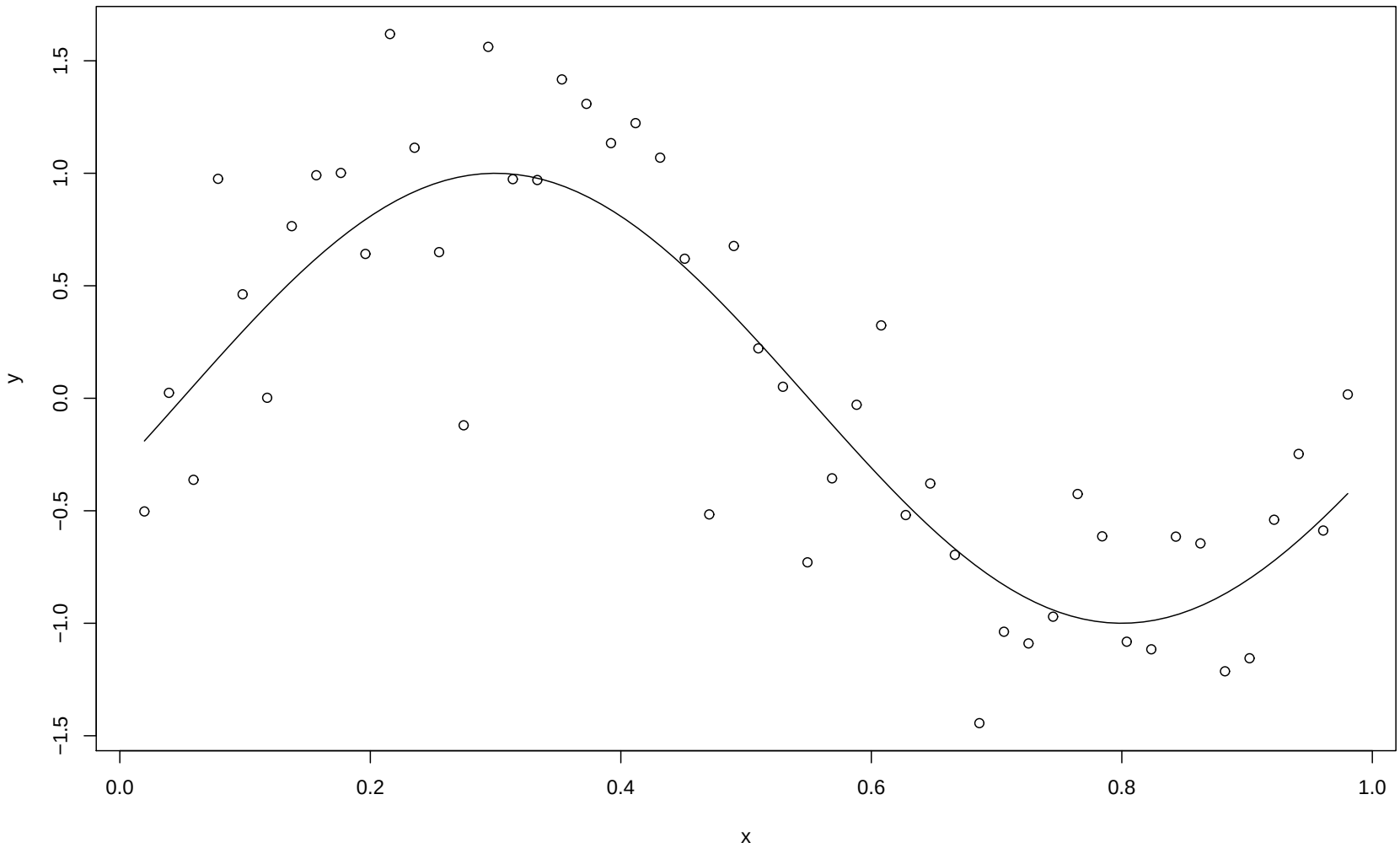
$$\mathrm{MSE}\{m_h^{NW}(x)\} \approx \frac{1}{nh}\sigma^2 \int K^2(s)ds + \frac{1}{4}h^4\{m''(x)\}^2.$$

Gasser-Müllerův odhad

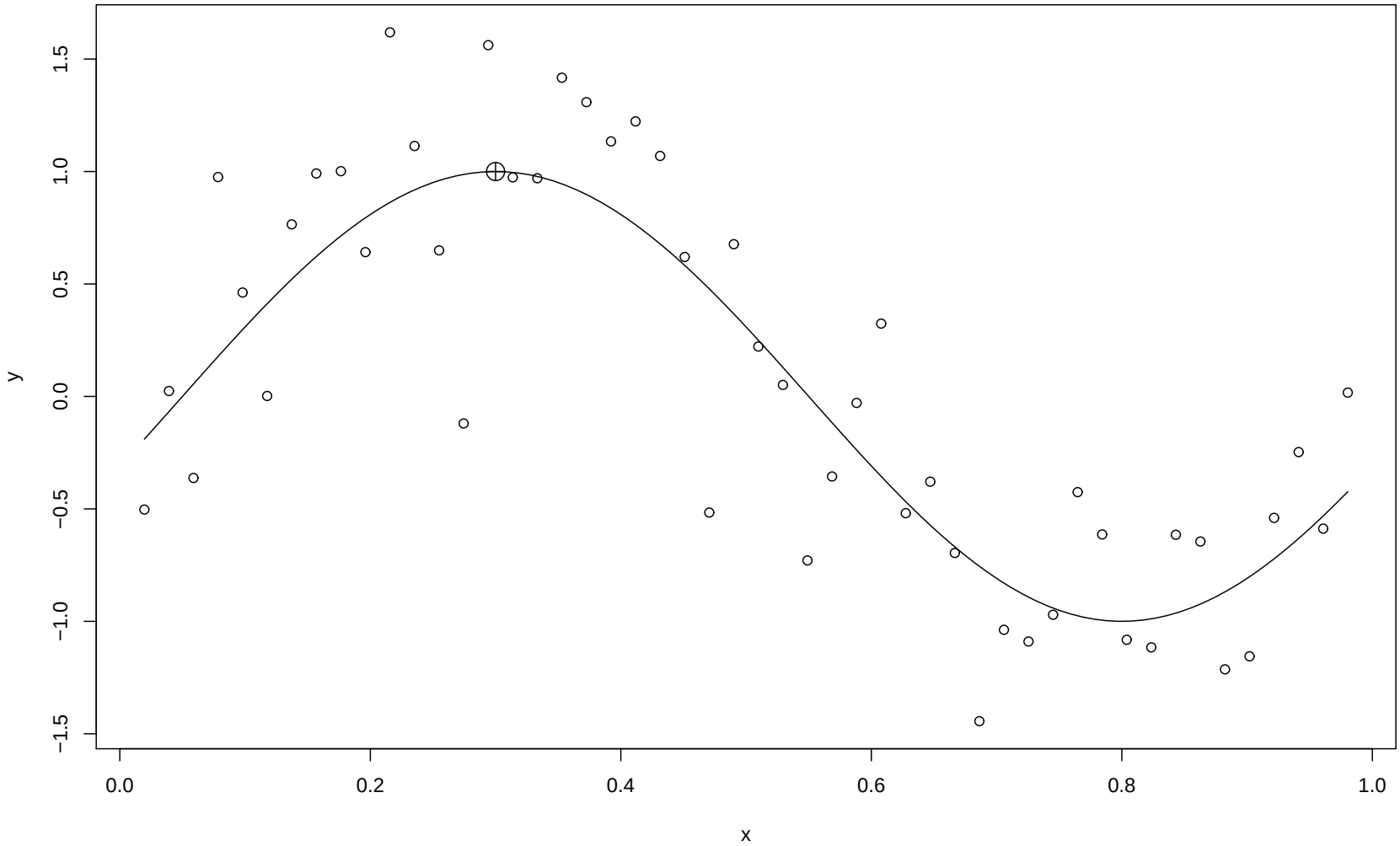
Pokud $x_1 < \cdots < x_n$ a $x_{i+1} - x_i \approx n^{-1}$, pak za "jistých předpokladů":

$$\mathrm{MSE}\{m_h^{GM}(x)\} \approx \frac{1}{nh}\sigma^2 \int K^2(s)ds + \frac{1}{4}h^4\{m''(x)\}^2.$$

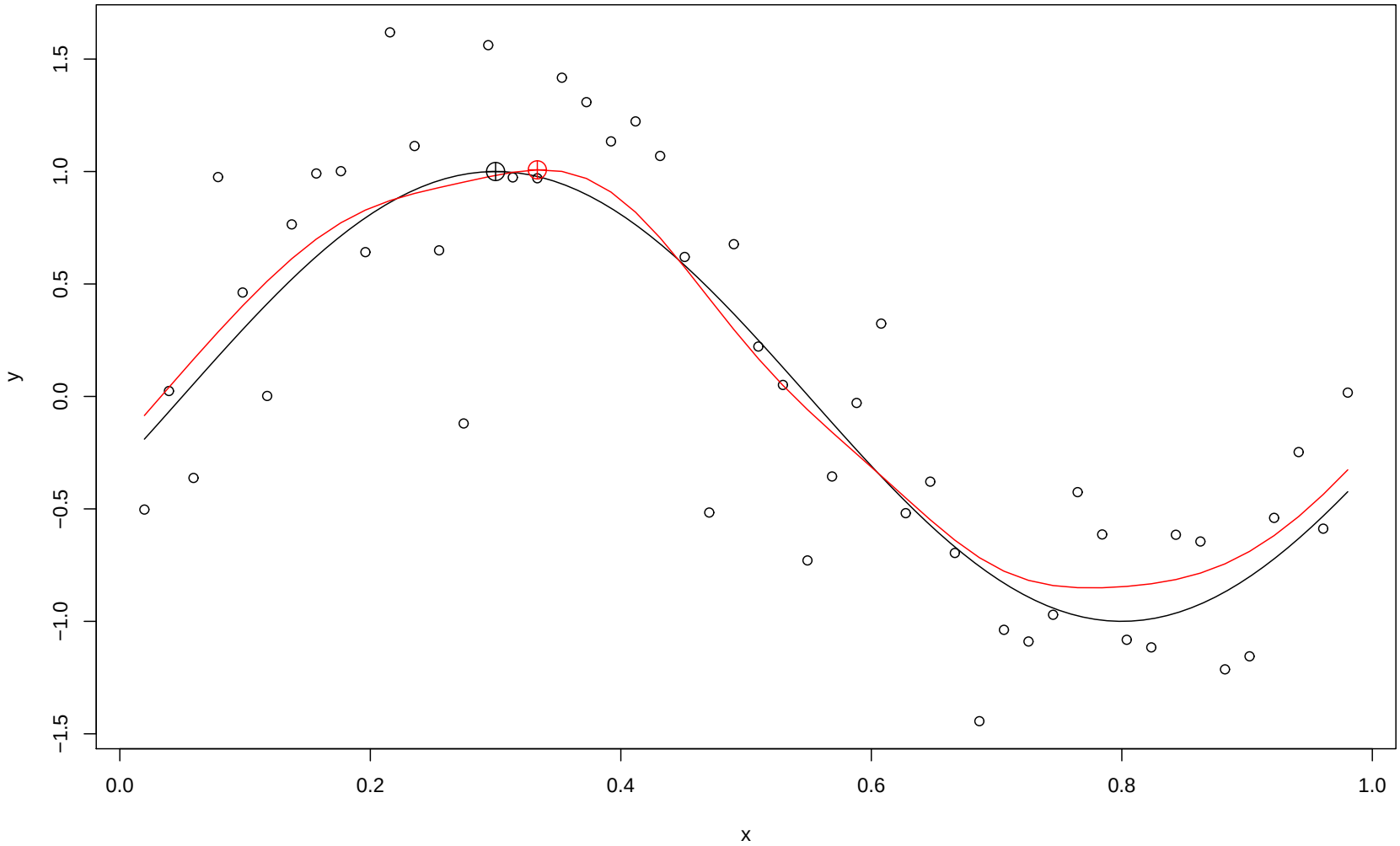
Často nás místo celé regresní křivky zajímá pouze hodnota nebo poloha nějakého zajímavého bodu (například maxima):



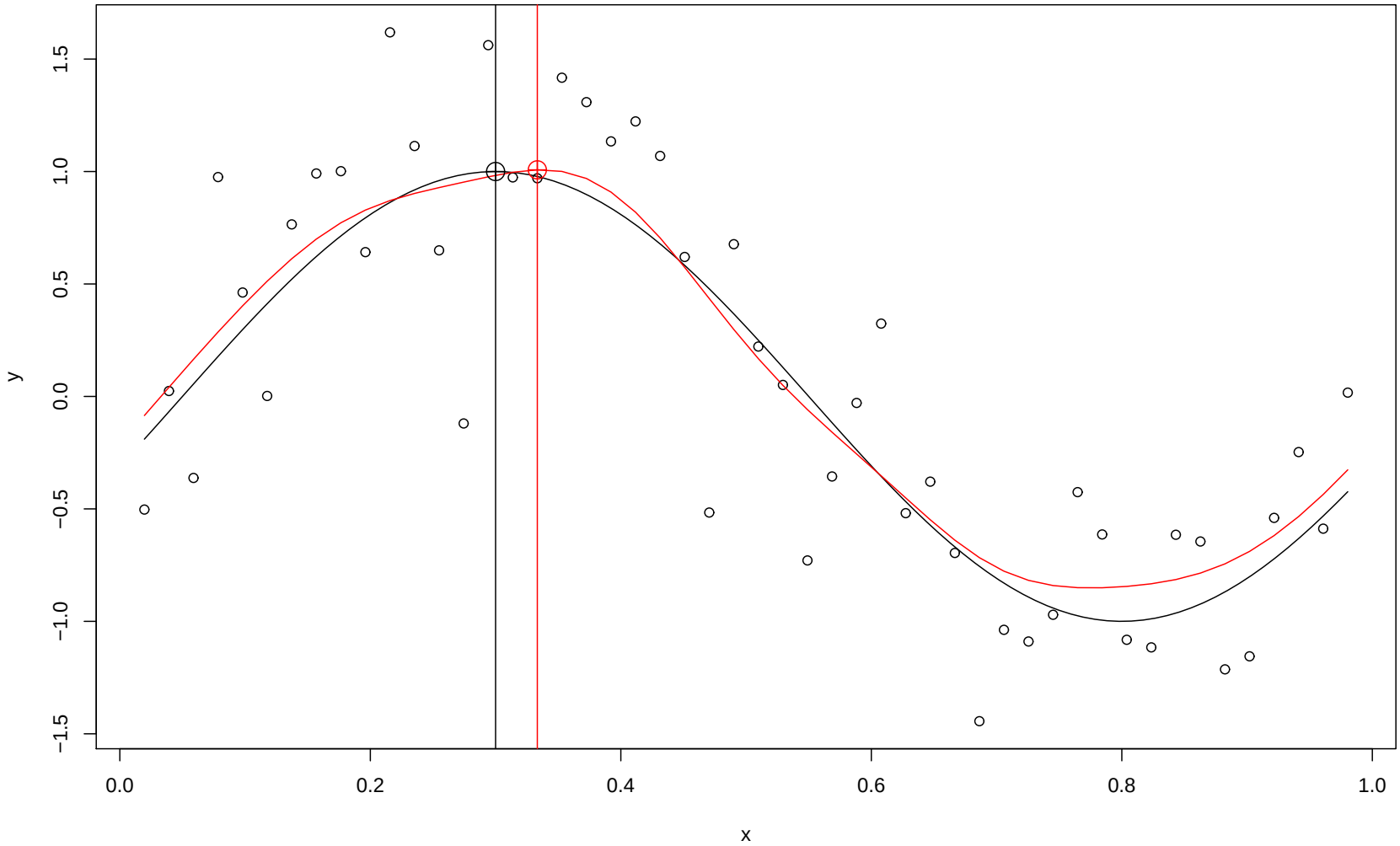
Často nás místo celé regresní křivky zajímá pouze hodnota nebo poloha nějakého zajímavého bodu (například maxima):



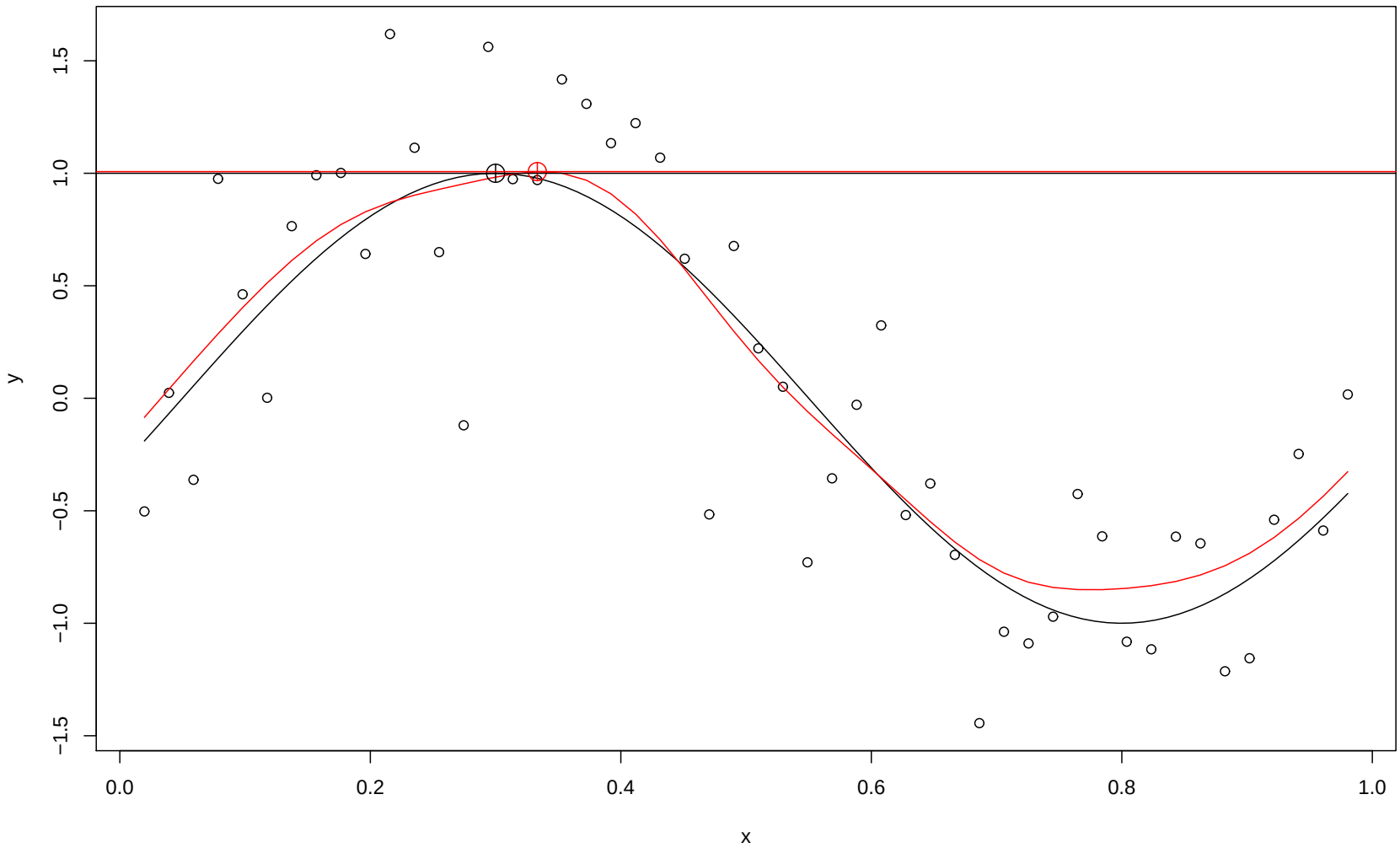
Často nás místo celé regresní křivky zajímá pouze hodnota nebo poloha nějakého zajímavého bodu (například maxima):



Často nás místo celé regresní křivky zajímá pouze hodnota nebo poloha nějakého zajímavého bodu (například maxima):



Často nás místo celé regresní křivky zajímá pouze hodnota nebo poloha nějakého zajímavého bodu (například maxima):



Neparametrický odhad maxima a jeho polohy

Odhadovaný parametr (poloha maxima): $\theta = \arg\max_x m(x)$

Odhad (empirická poloha maxima):
$\theta_n = \inf\left\{x : m_h^{GM}(x) = \max_x m_h^{GM}(x)\right\}$

Odhad maxima regresní funkce je: $m_h^{GM}(\theta_n)$

Vlastnosti θ_n jako odhadu θ a $m_h^{GM}(\theta_n)$ jako odhadu $m(\theta)$ jsou odvozeny v článku Müller (1985).

(Příklad: koncentrace hormonu FSH u pubertálních chlapců. Jak souvisí doba maximální produkce hormonu s růstem nebo s pubertou?)

Müller, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions, *Scand. J. Statist.* 12: 221–232.

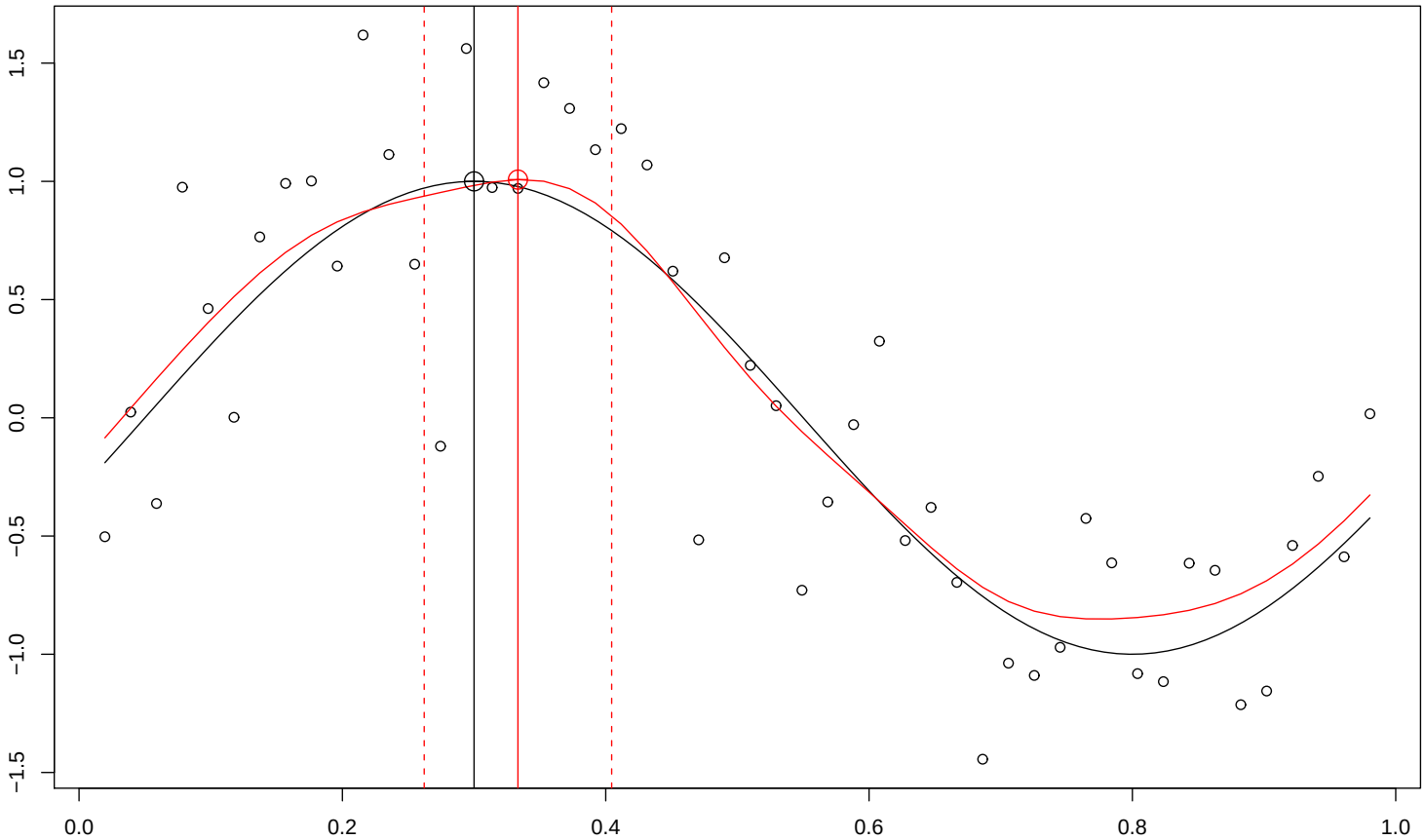
1. Funkce $m(.)$ je 4×spojitě diferencovatelná. Jádrová funkce $K(.)$ je 2×spojitě diferencovatelná a taková, že $K^{(2)}(.)$ je Lipschitzovská.
2. Předpokládejme, že $\liminf_{n\to\infty} nh_n^4 > 0$, $nh_n^5/\log n \to \infty$ a pro nějaké $r > 2$ platí $E|\varepsilon_1|^r < \infty$ a $\liminf_{n\to\infty} h_n n^{1-2/r}/\log n > 0$.
3. Existují $x_l < \theta < x_u$, $c > 0$ a $\rho \geq 1$ tak, že $m(.)$ je rostoucí $[x_l, \theta]$ a klesající na $[\theta, x_u]$ a $|m(t) - m(\theta)| > c|t - \theta|^\rho$ pro $t \in [x_l, x_u,]$.

Müller (1985), Věta 3.1: Předpokládejme, že platí podmínky 1–3 a body měření jsou rovnoměrně rozložené na $(0, 1)$. Pokud $nh_n^7 \to d^2 \geq 0$, pak

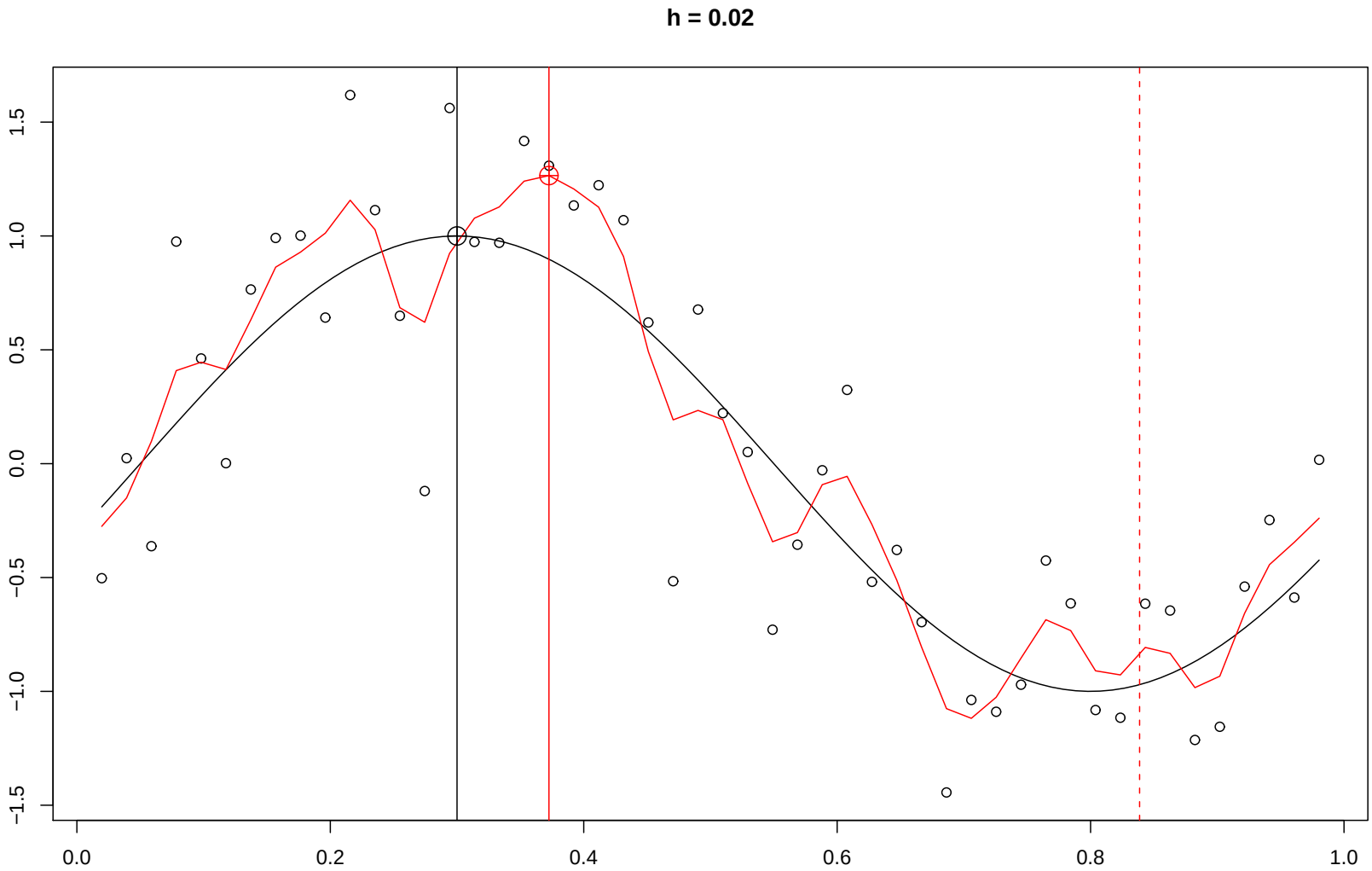
$$(nh_n^3)^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{D}} N\left(-\frac{dm^{(3)}(\theta)B_2}{m^{(2)}(\theta)}, \frac{\sigma^2 V'}{\{m^{(2)}(\theta)\}^2}\right),$$

kde $B_2 = (1/2)\int K(s)s^2 ds$ a $V' = \int \{K^{(1)}(s)\}^2 ds$.

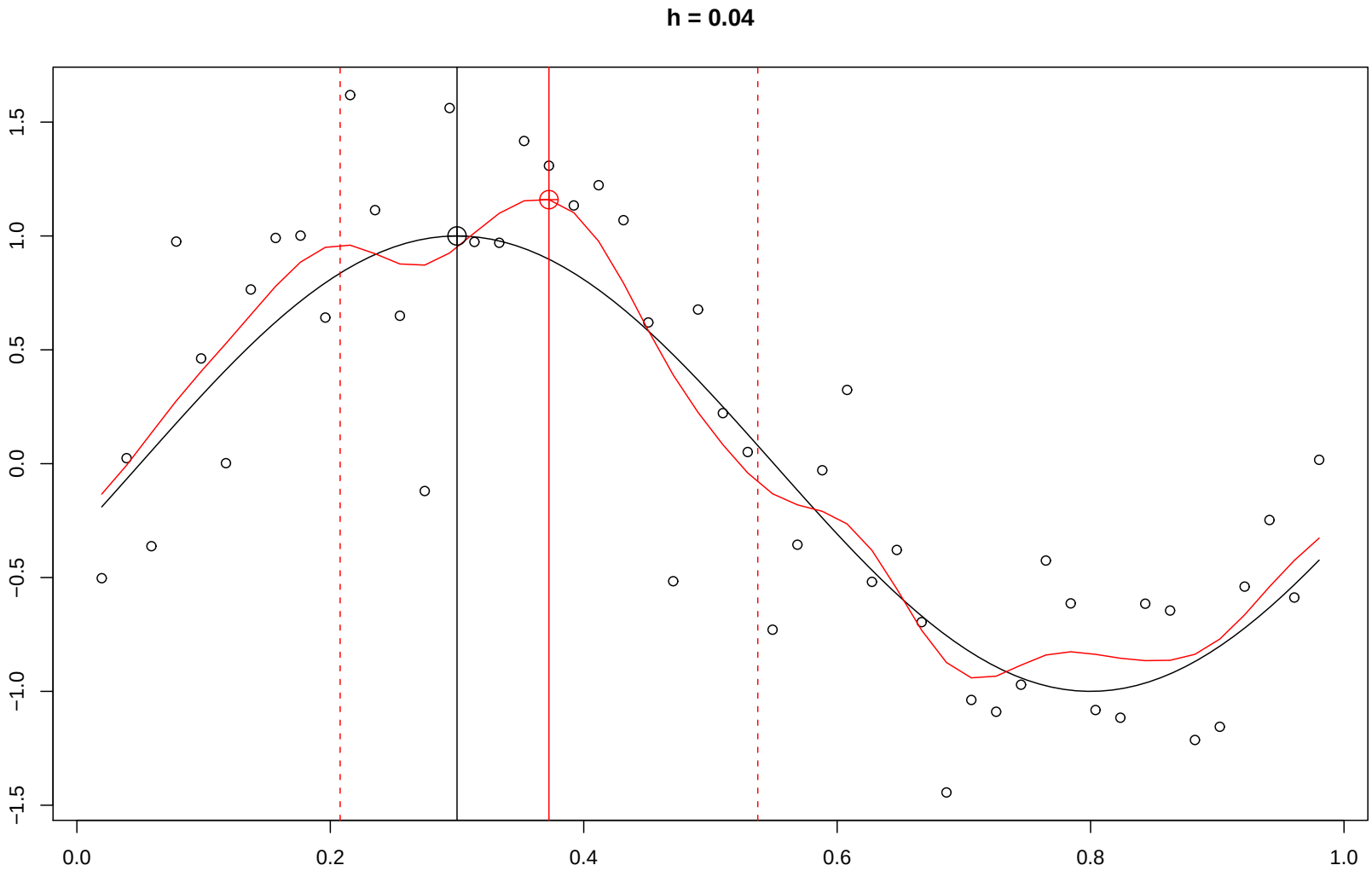
$$P\left\{\left(\theta_n - \frac{u_{1-\alpha/2}}{\sqrt{nh_n^3}}\frac{\sigma\sqrt{V'}}{m^{(2)}(\theta)}, \theta_n + \frac{u_{1-\alpha/2}}{\sqrt{nh_n^3}}\frac{\sigma\sqrt{V'}}{m^{(2)}(\theta)}\right) \ni \theta\right\} \approx 1-\alpha$$



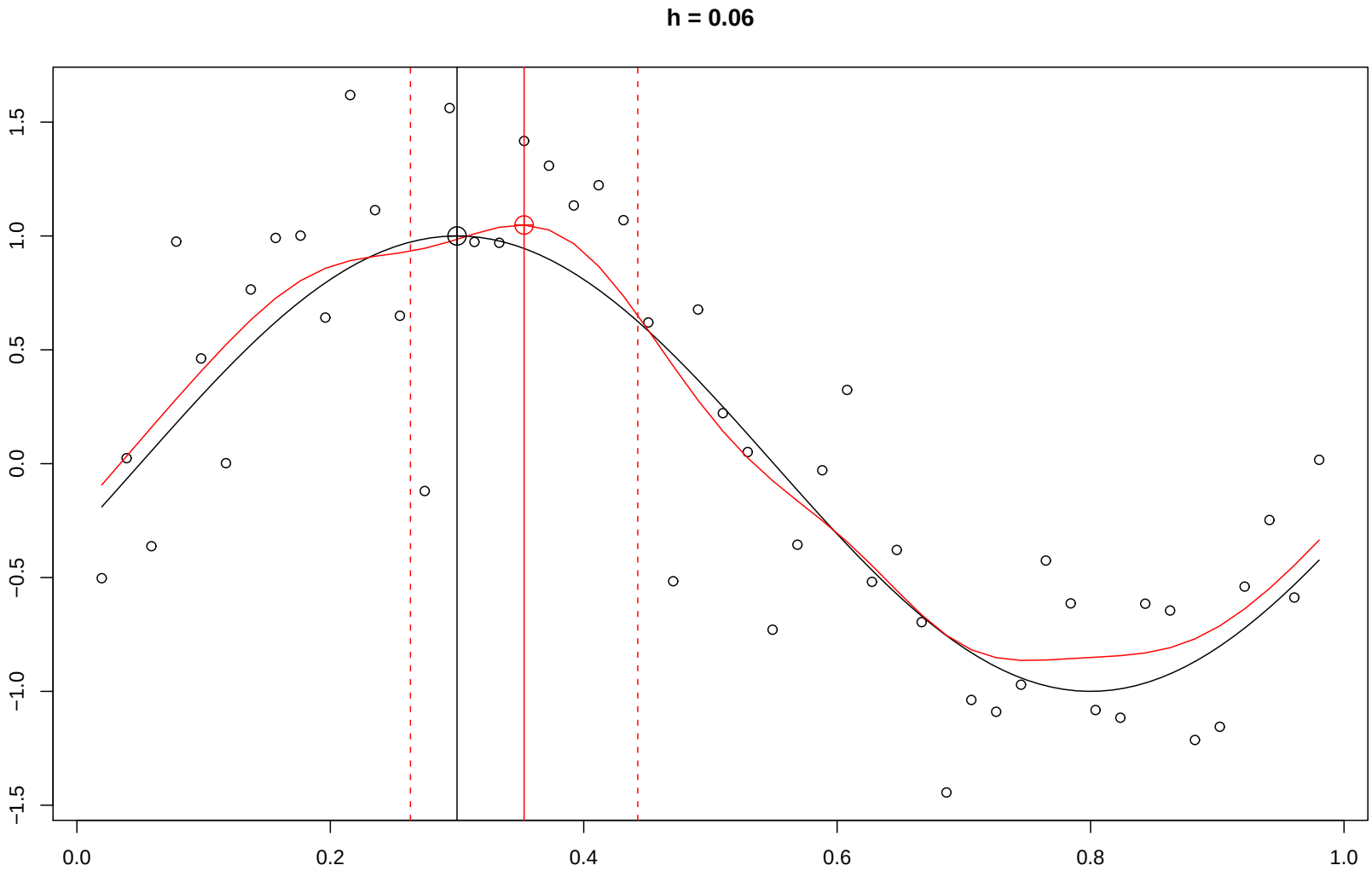
Závislost na bandwidth h_n



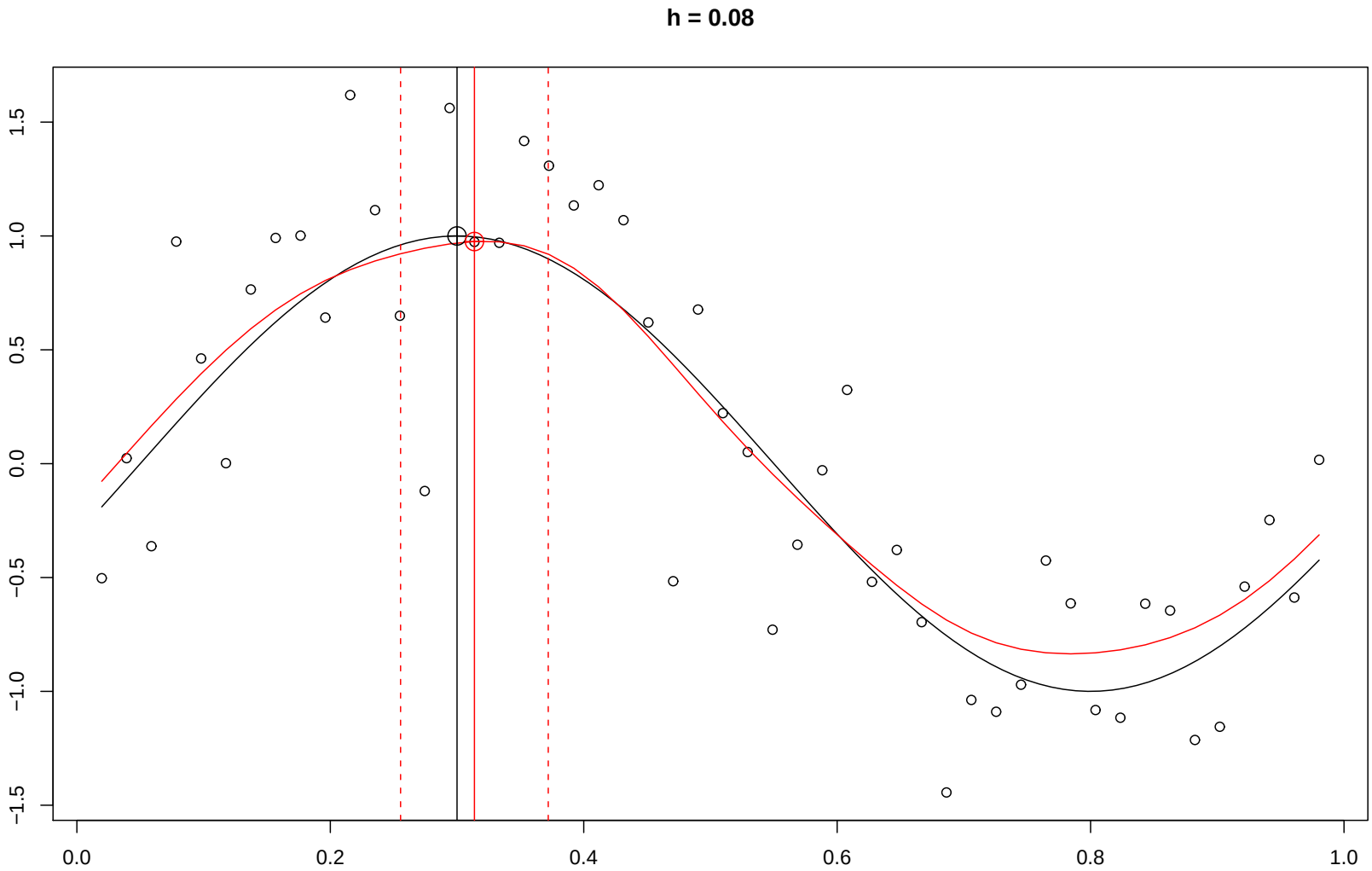
Závislost na bandwidth h_n



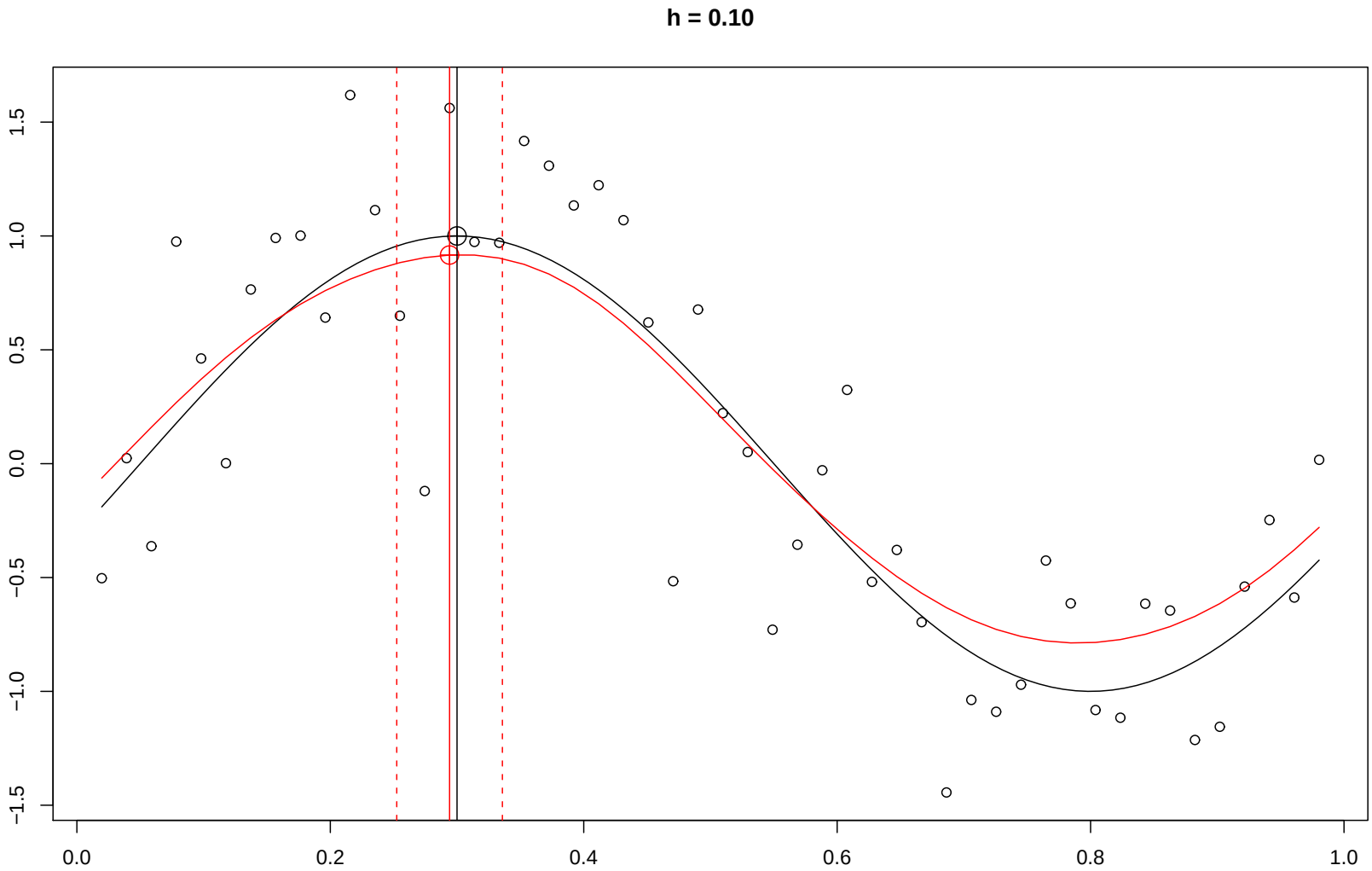
Závislost na bandwidth h_n



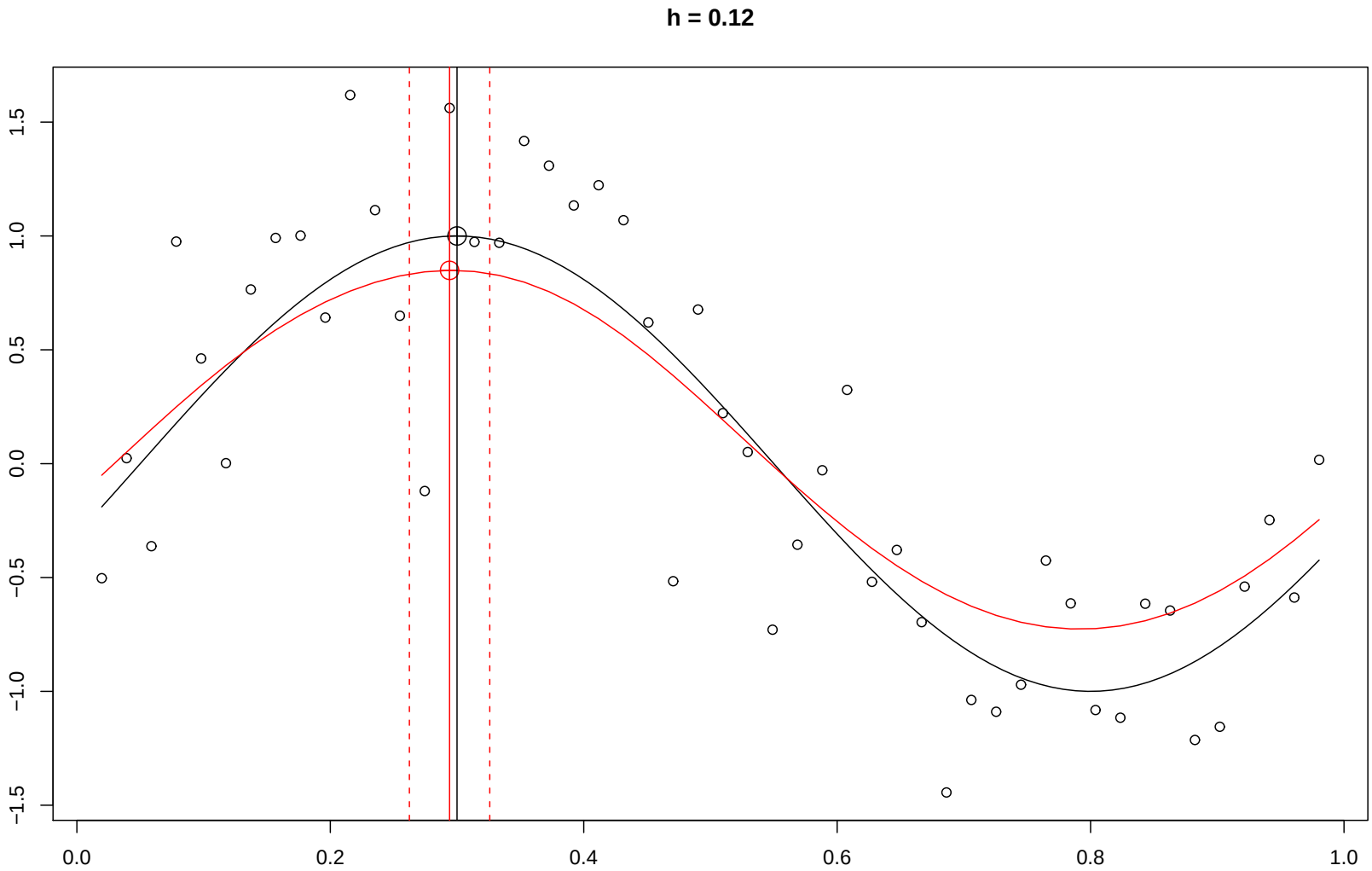
Závislost na bandwidth h_n



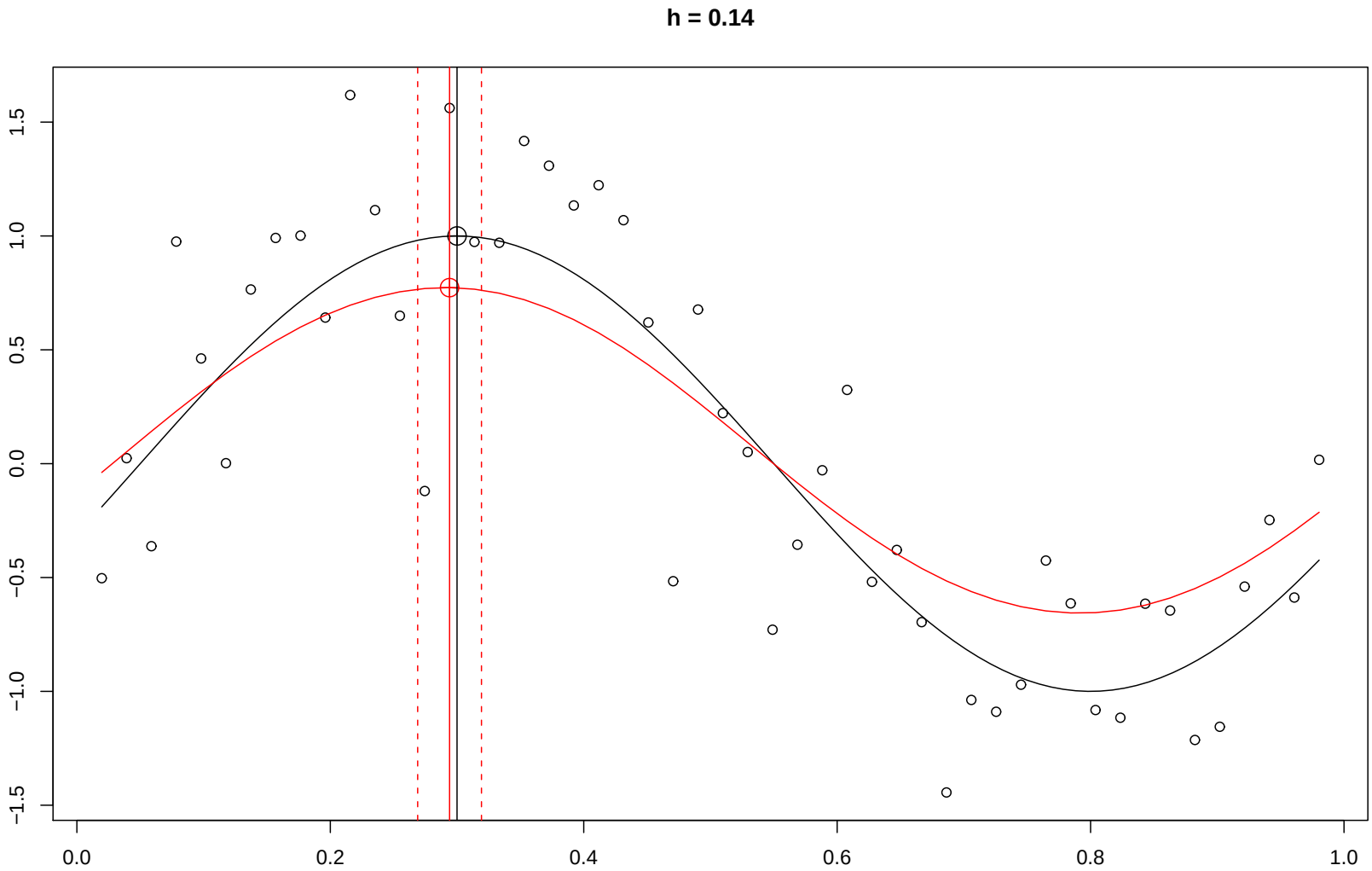
Závislost na bandwidth h_n



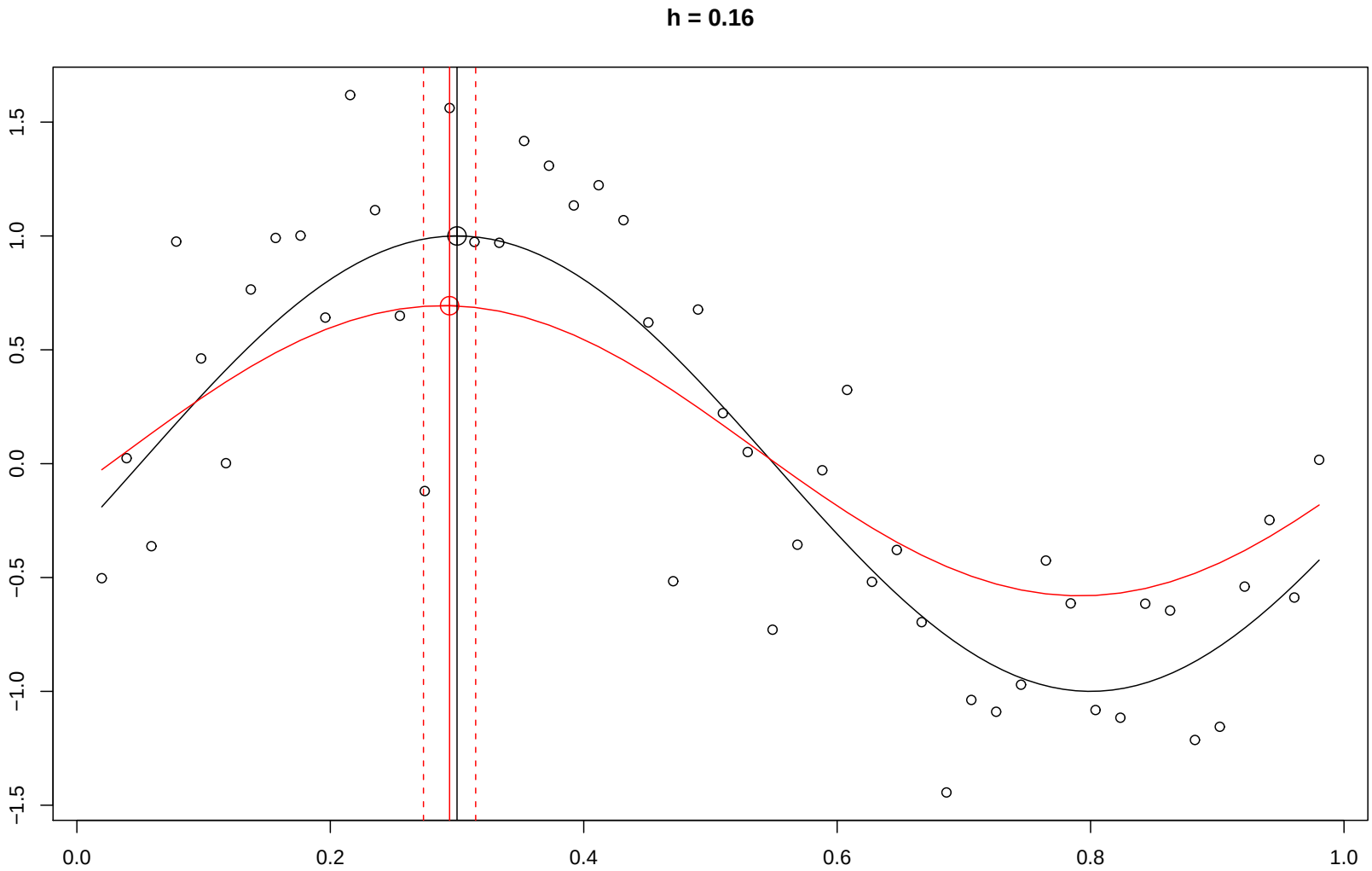
Závislost na bandwidth h_n



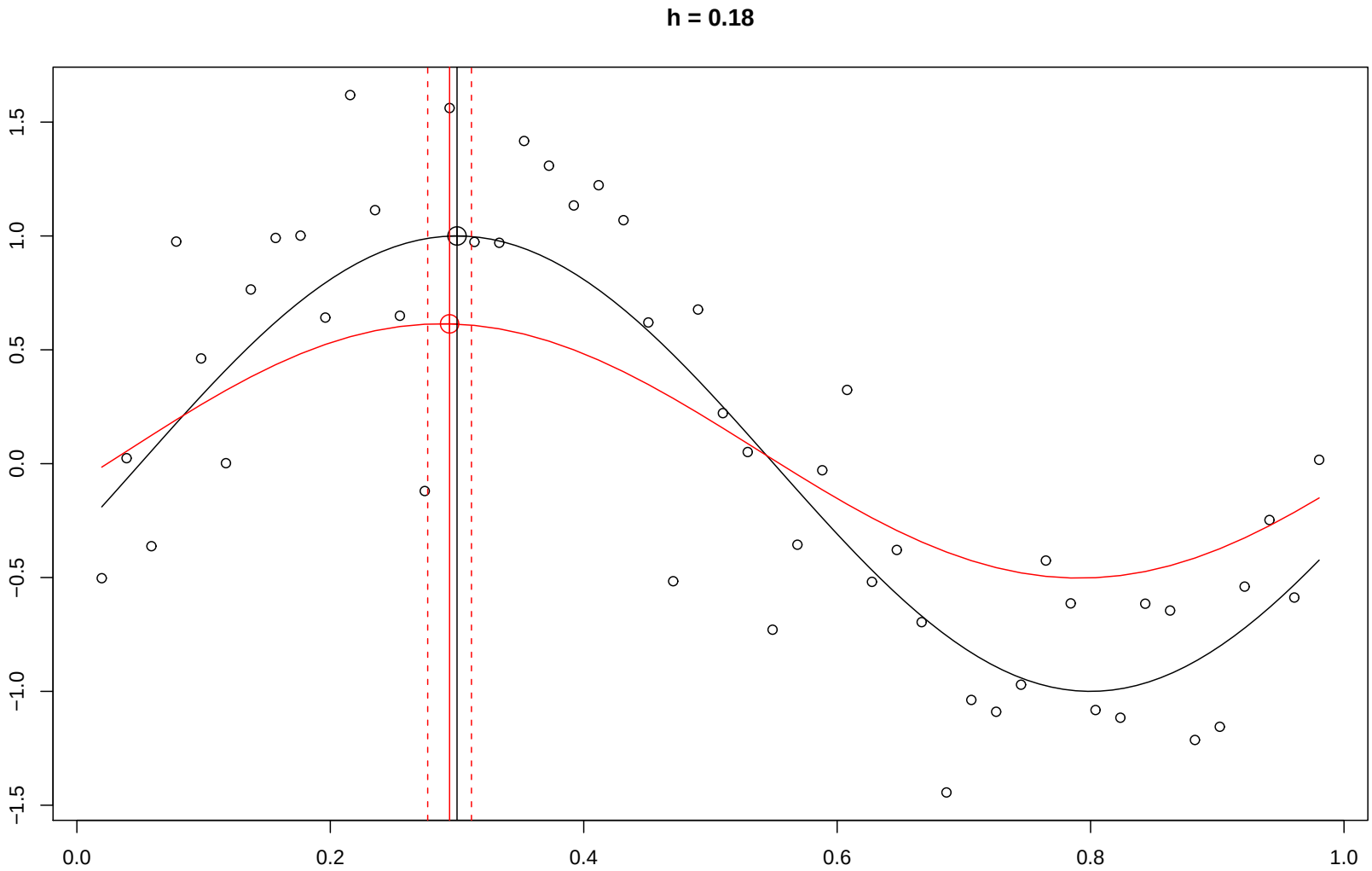
Závislost na bandwidth h_n



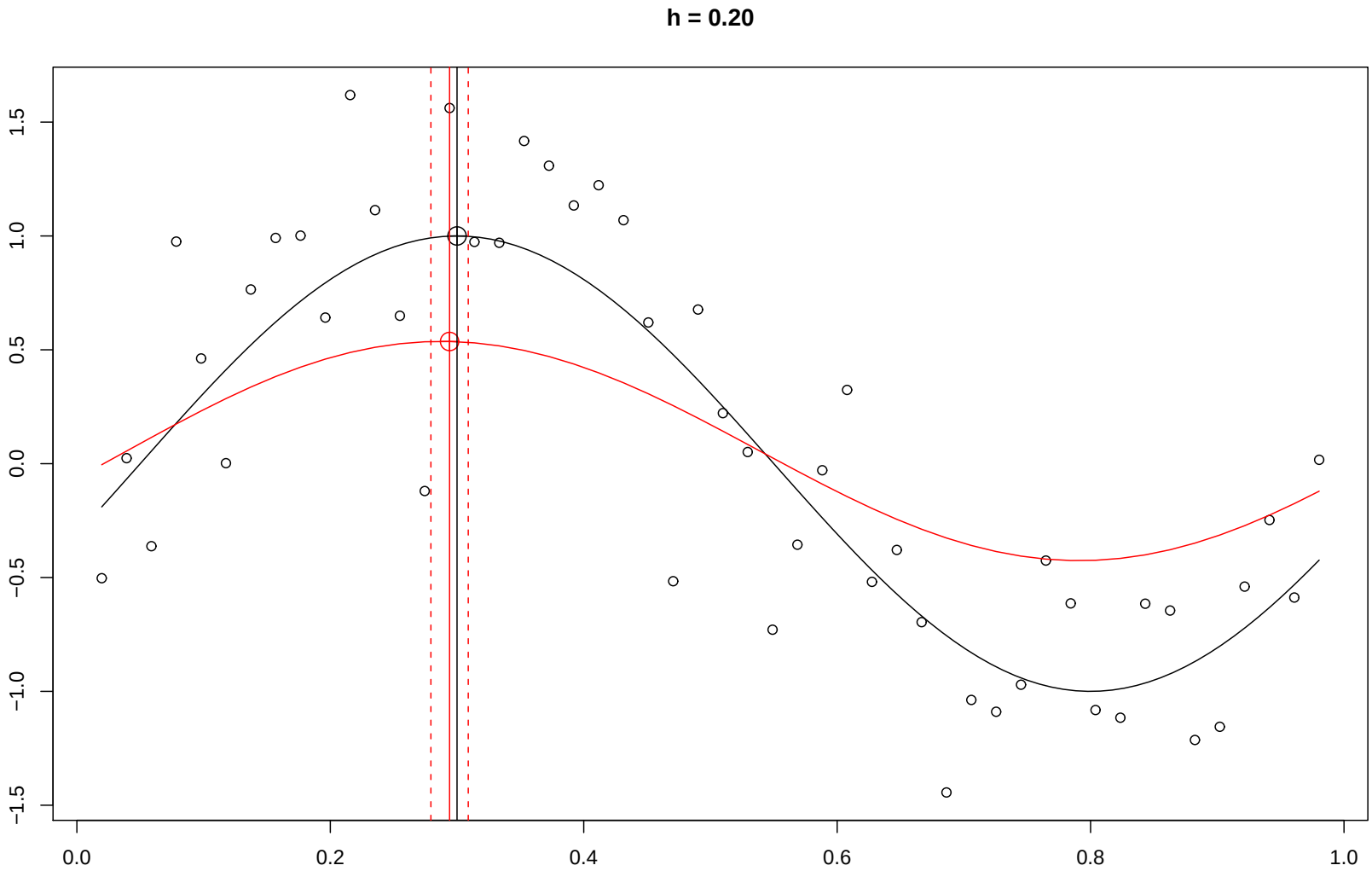
Závislost na bandwidth h_n



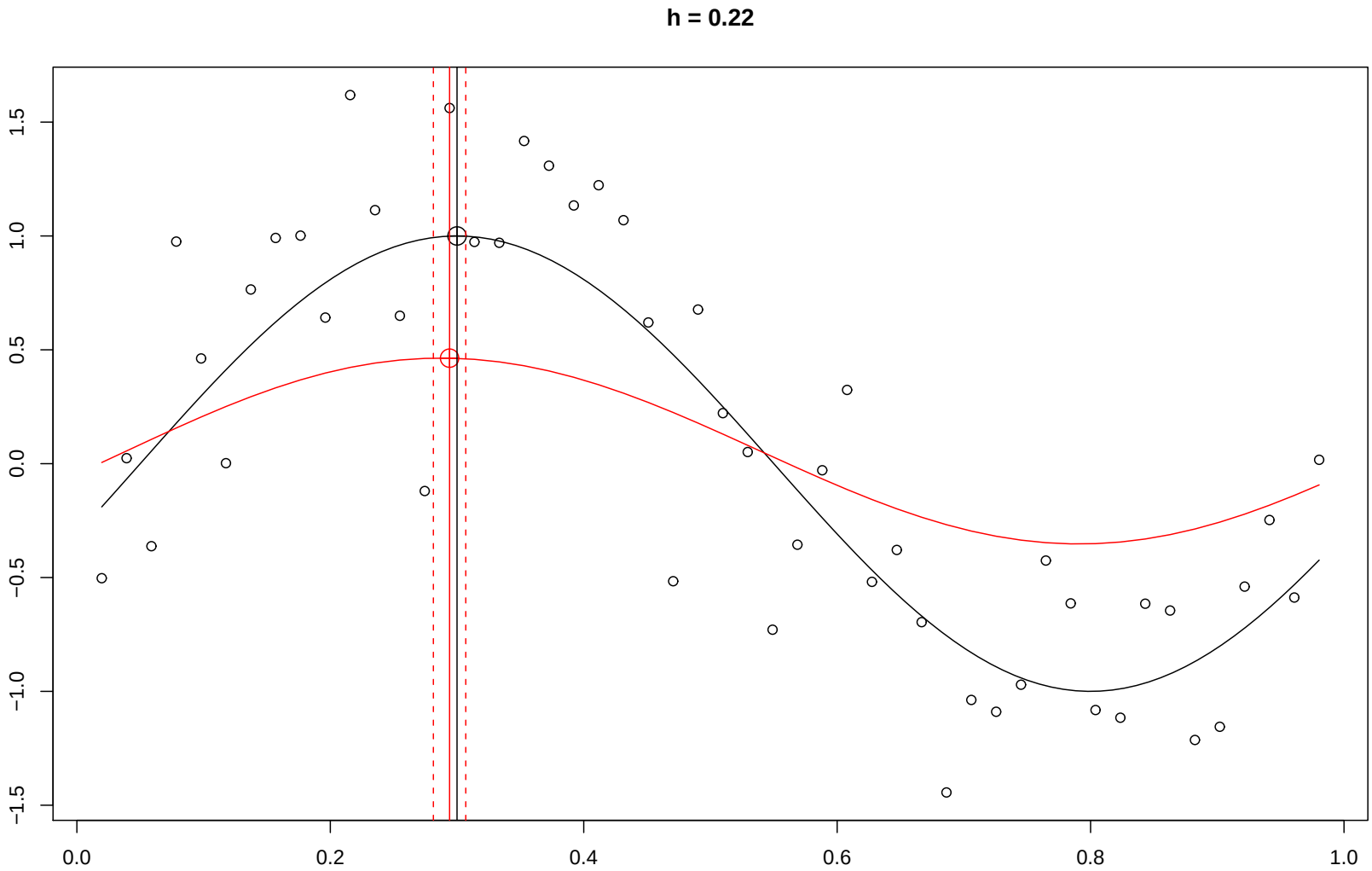
Závislost na bandwidth h_n



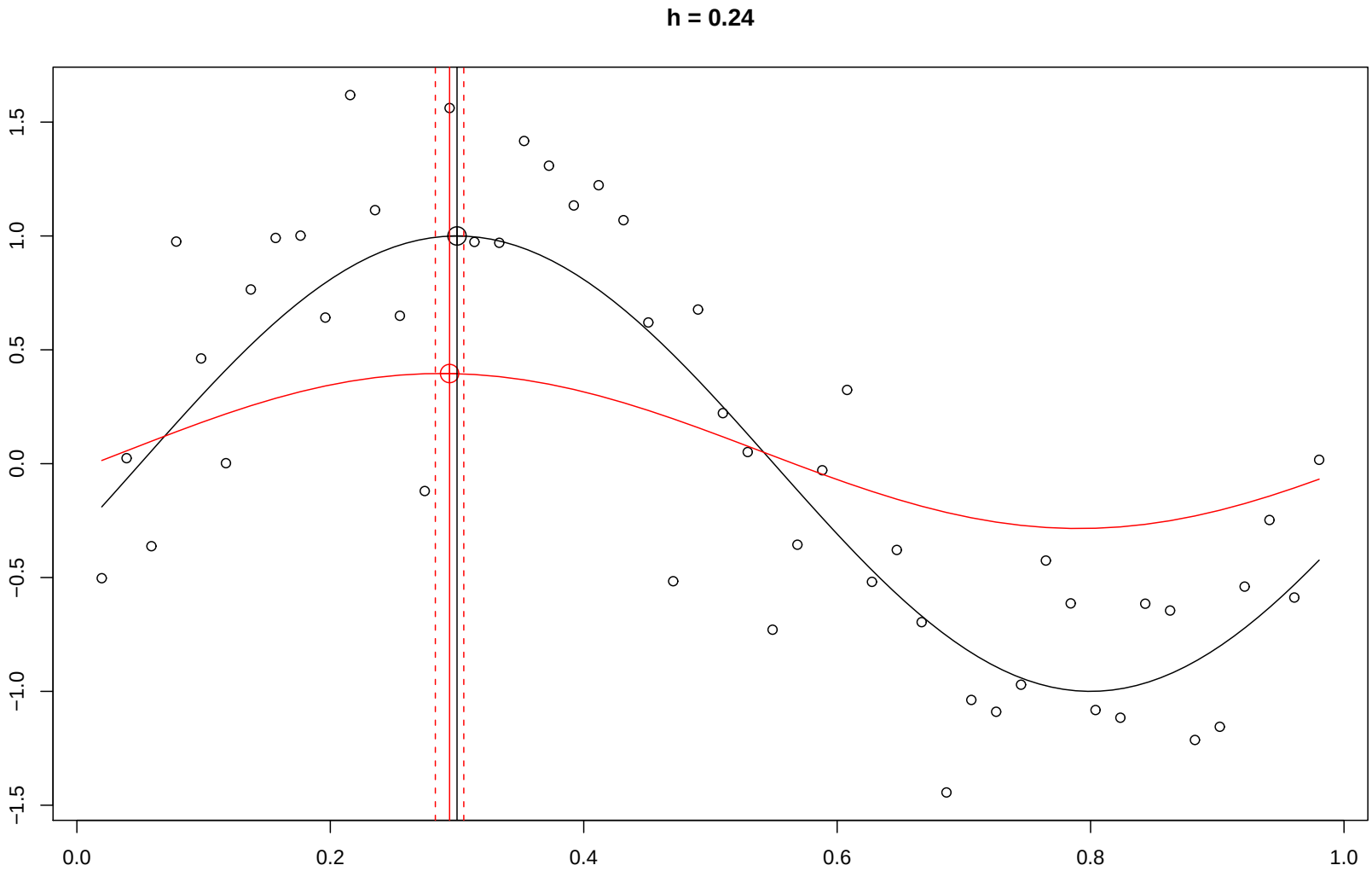
Závislost na bandwidth h_n



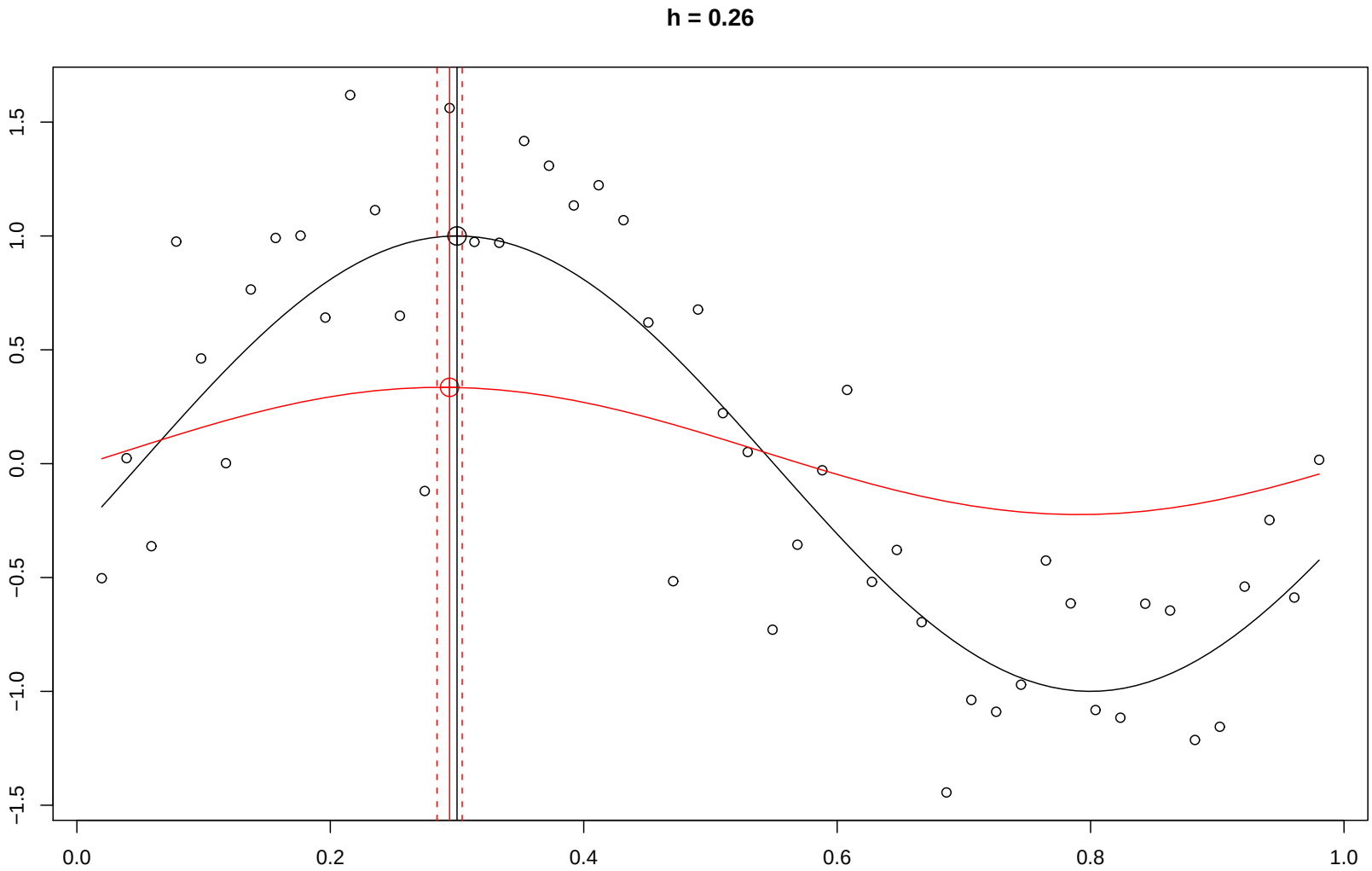
Závislost na bandwidth h_n



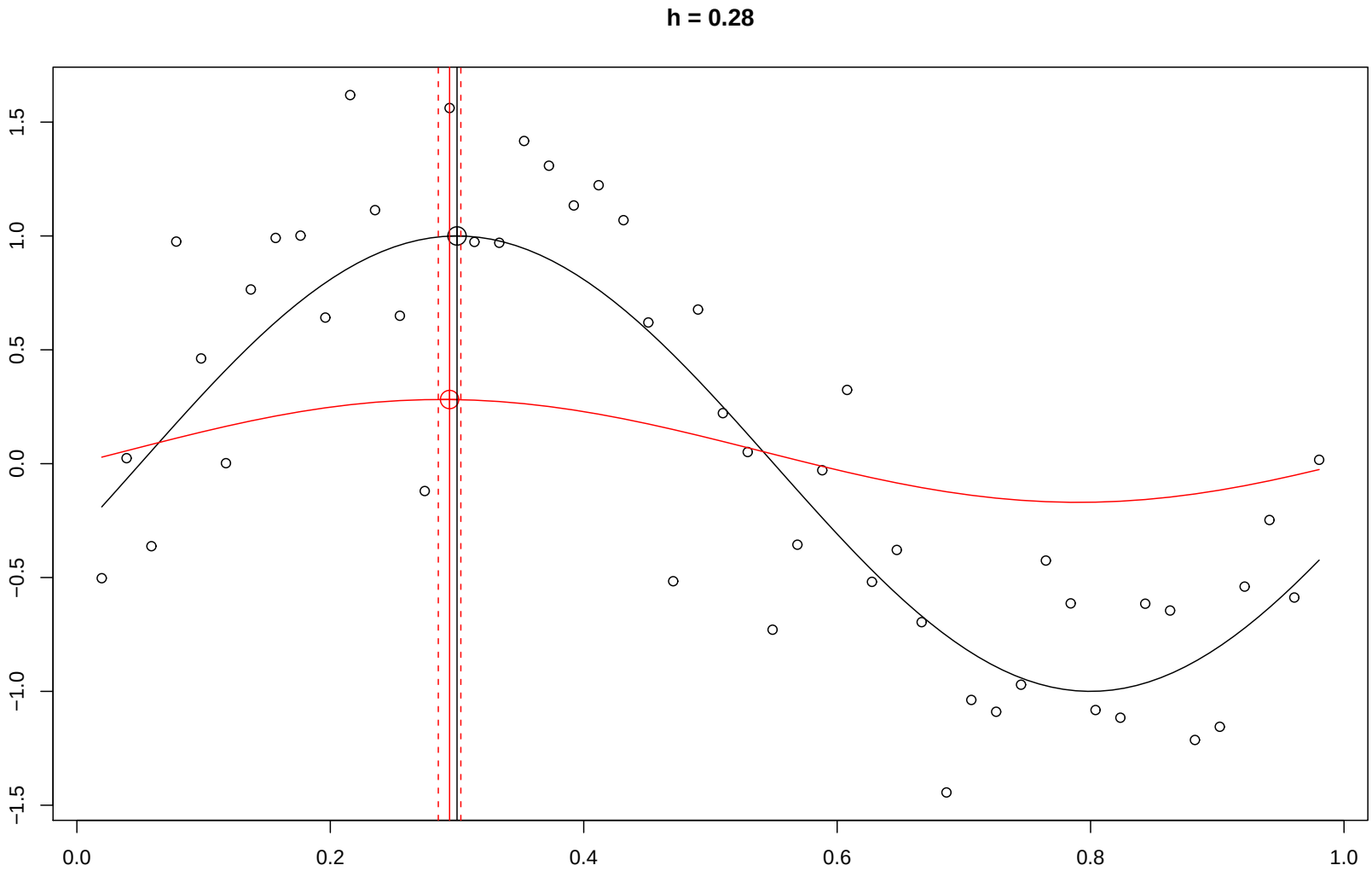
Závislost na bandwidth h_n



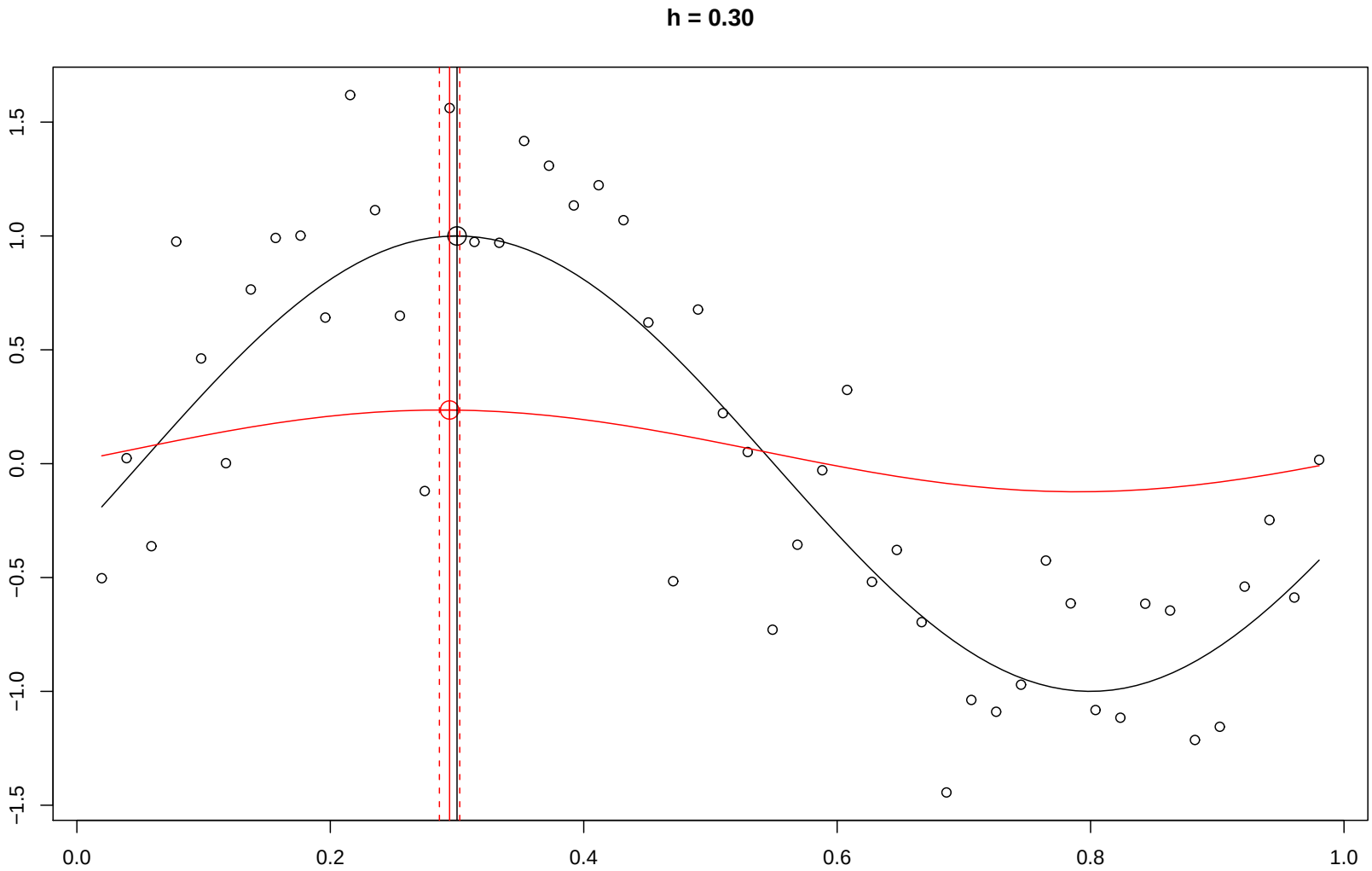
Závislost na bandwidth h_n



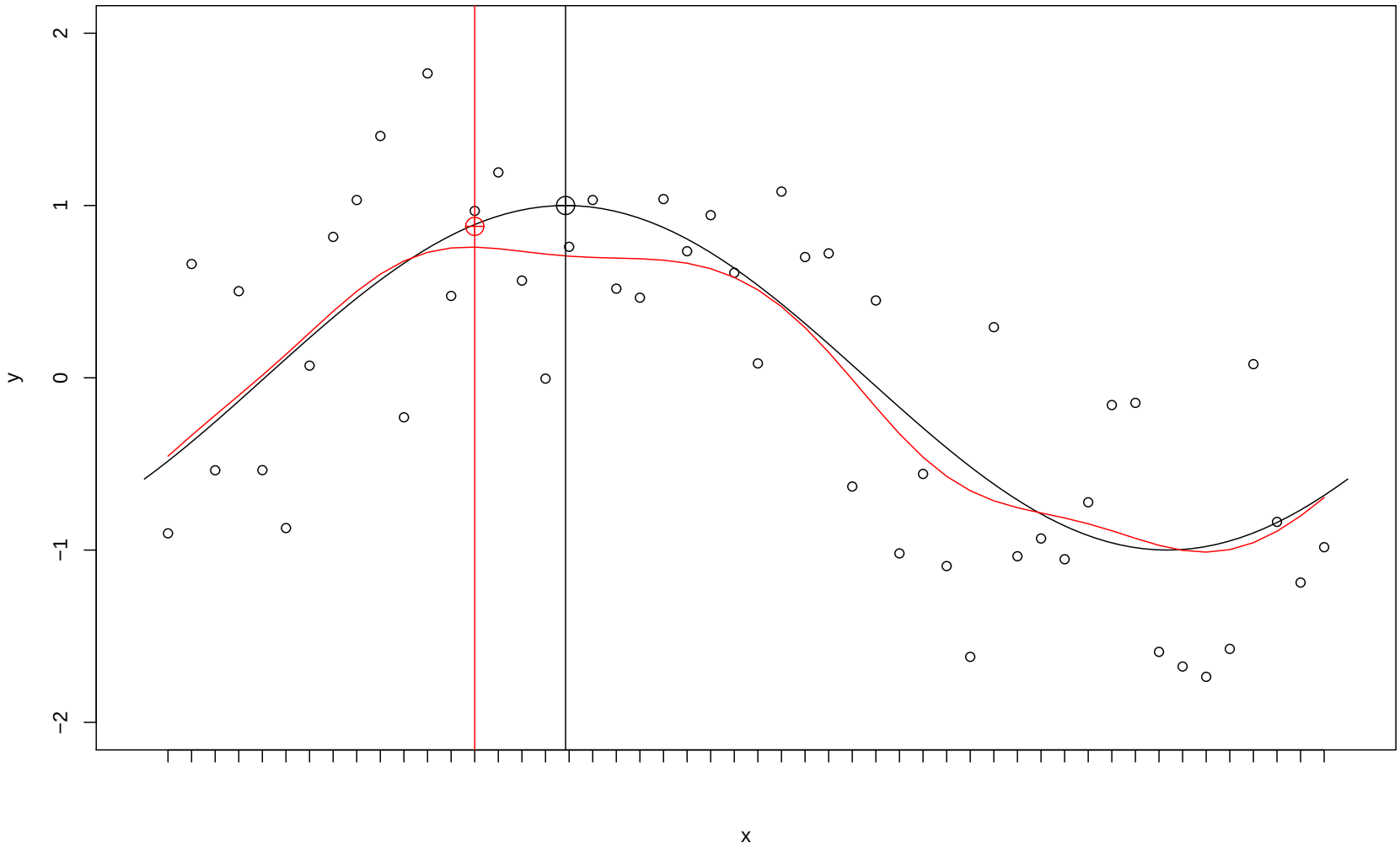
Závislost na bandwidth h_n



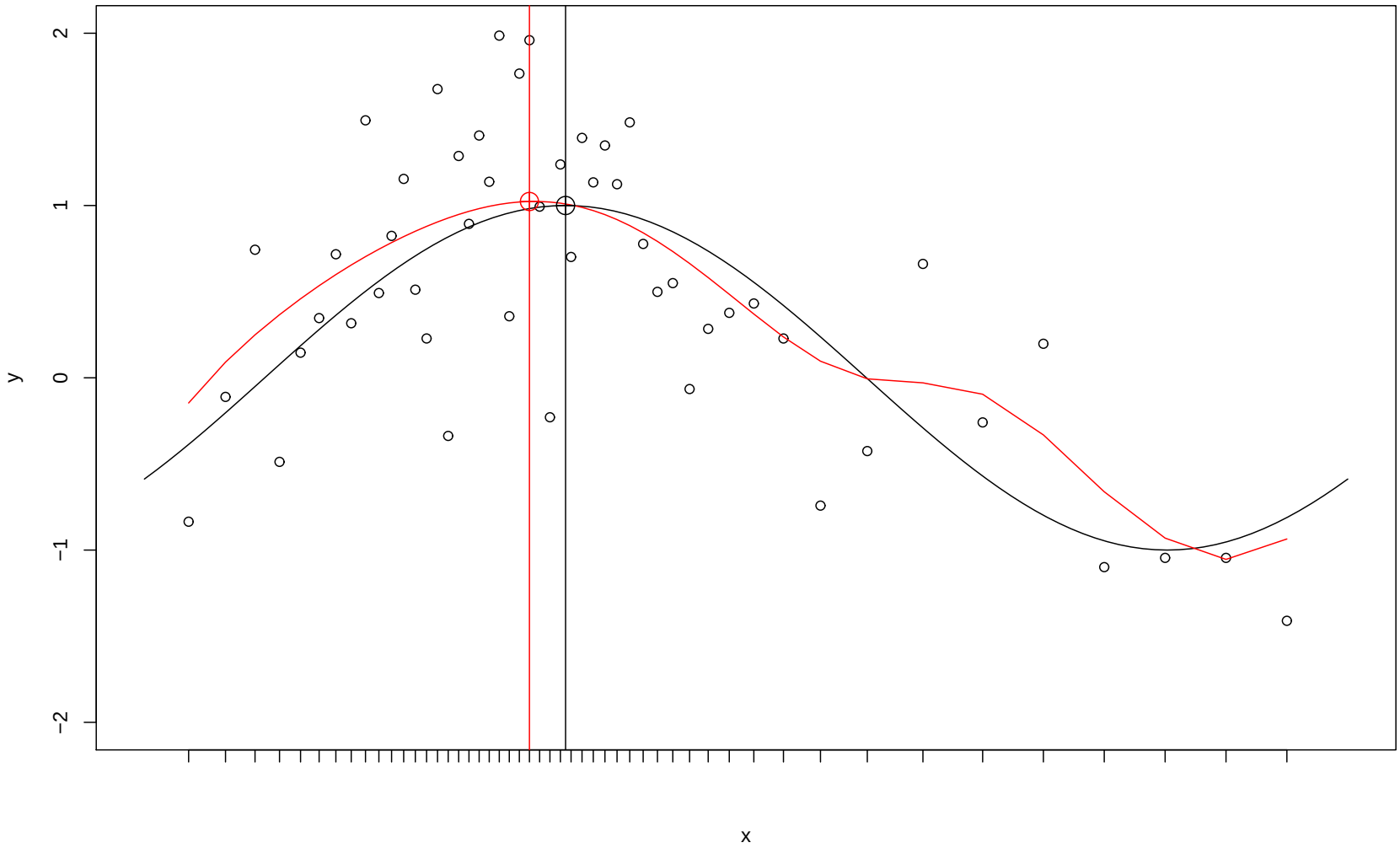
Závislost na bandwidth h_n



Jak se změní vlastnosti GM odhadu při změně $f_X(x)$?



Jak se změní vlastnosti GM odhadu při změně $f_X(x)$?



Nerovnoměrně rozložená měření

4. Hustota měření, $f_X(.)$, je spojitě diferencovatelná a existuje $\delta_0 > 0$ tak, že $f_X(x) \geq \delta_0$, $x \in [0, 1]$.

Předpokládejme, že jsou splněny podmínky 1–4 a že body měření jsou rozloženy podle hustoty $f_X(.)$. Pokud $nh_n^7 \to d^2 \geq 0$, pak

$$\mathrm{Var}\{m^{GM}(x)\} = \frac{\sigma^2}{nh_n}\left[\int K^2(s)ds\{f_X(x)\}^{-1} + o(1)\right].$$

Jaké rozložení bodů na x-ové ose je nejlepší?

Müller, H.-G. (1984). Optimal designs for nonparametric kernel regression, *Statistics & Probability Letters* 2: 285–290.

AIMSE optimální rozložení měření

Zvolíme pravděpodobnostní míru H s kladnou a spojitou hustotou $h(.)$ na intervalu $[0, 1]$,

$$IMSE = E \int \{m^{GM}(x) - m(x)\}^2 dH(x) \approx \frac{1}{nh_n} \int K^2(s)ds \int \frac{h(x)}{f_X(x)} dx$$

Müller (1984): odvození AIMSE optimálního rozložení $f_X^*(.)$ za různých předpokladů (např. heteroskedasticita, nekonstantní bandwidth apod.).

Za našich předpokladů (homoskedasticita, konstantní bandwidth):

$$f_X^*(x) = \frac{h(x)^{1/2}}{\int h(x)^{1/2} dx} \propto h(x)^{1/2}$$

Nevýhoda: není jasné, co je ta míra H a jak by se měla volit.

Apriorní informace o poloze maxima

V praxi často můžeme získat alespoň přibližnou apriorní informaci o poloze maxima díky

- dřívějším experimentům,
- zkušenostem a literatuře.

Očekávanou polohu maxima můžeme vyjádřit pomocí apriorní hustoty $a(.)$ odhadovaného parametru θ.

5. Předpokládejme, že apriorní hustota $a(.)$ je spojitě diferencovatelná a že existuje $0 < \delta < K$ tak, že $\delta < a(x) < K$, $x \in [0, 1]$.

Odhad polohy maxima: nerovnoměrný design

Předpokládejme, že jsou splněny podmínky 1–4 a že body měření jsou rozložené podle hustoty $f_X(.)$. Pokud $nh_n^7 \to d^2 \geq 0$, pak

$$(nh_n^3)^{1/2}(\theta_n - \theta) \xrightarrow{\mathcal{D}} N\left(-\frac{dm^{(3)}(\theta)B_2}{m^{(2)}(\theta)}, \frac{\sigma^2 V'}{\{m^{(2)}(\theta)\}^2 f_X(\theta)}\right).$$

Zvolíme-li bandwidth h_n tak, že $nh_n^7 \to 0$, získáme $1-\alpha$ konfidenční interval pro polohu maxima:

$$\left(\theta_n - \frac{u_{1-\alpha/2}}{\sqrt{nh_n^3}} \frac{\sigma\sqrt{V'}}{|m^{(2)}(\theta)| f_X^{1/2}(\theta)}, \theta_n + \frac{u_{1-\alpha/2}}{\sqrt{nh_n^3}} \frac{\sigma\sqrt{V'}}{|m^{(2)}(\theta)| f_X^{1/2}(\theta)}\right).$$

Délka intervalu spolehlivosti je přímo úměrná $f_X^{-1/2}(\theta)$.

Pokud je parametr θ tam, kde je víc měření, tak výsledný konfidenční interval bude kratší.

Optimální rozložení bodů měření

Předpokládejme, že jsou splněny podmínky 1–5 a že navíc $m^{(2)}(\theta)$ nezávisí na θ:

1. hustota bodů měření, která minimalizuje $\int \mathrm{Var}(\theta_n|\theta = u)a(u)du$ je:

$$f_X^*(x) \propto a^{1/2}(x),$$

Optimální rozložení bodů měření

Předpokládejme, že jsou splněny podmínky 1–5 a že navíc $m^{(2)}(\theta)$ nezávisí na θ:

1. hustota bodů měření, která minimalizuje $\int \mathrm{Var}(\theta_n|\theta = u)a(u)du$ je:

$$f_X^*(x) \propto a^{1/2}(x),$$

2. hustota bodů měření, která minimalizuje očekávanou délku konfidenčního intervalu pro polohu maxima je:

$$f_X^{*CI}(x) \propto a^{2/3}(x).$$

Důkaz: Víme, že $\mathrm{Var}(\theta_n|\theta = x) = cf_X^{-1}(x)$, kde $f_X(.)$ je hustota určující rozložení bodů měření a c je známá konstanta, která závisí na n, h_n, σ^2 a $m^{(2)}(\theta)$. Řešení minimalizačního problému

$$\begin{aligned} f_X^* &= \arg\min_{f_X} \int_0^1 \mathrm{Var}(\theta_n|\theta) a(\theta) d\theta \\ &= \arg\min_{f_X} \int_0^1 cf_X^{-1}(\theta) a(\theta) d\theta \end{aligned}$$

musí splňovat Eulerovu diferenciální rovnici, která se v tomto případě zjednoduší na

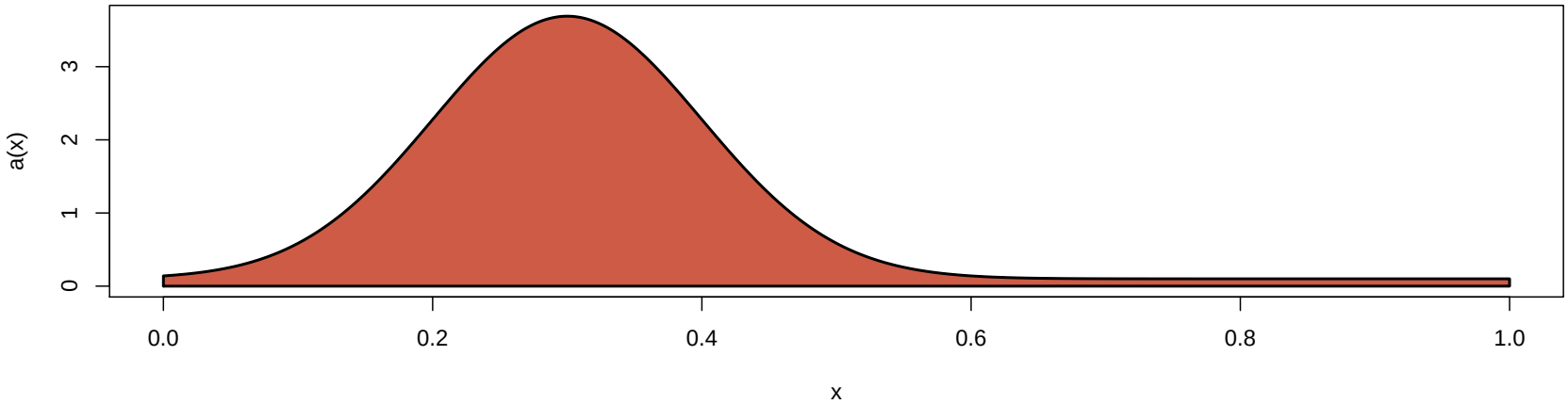
$$\{f_X^*(x)\}^{-2} a(x) = \text{konstanta}$$

a tedy

$$f_X^*(x) = \text{konstanta} \times a^{1/2}(x).$$

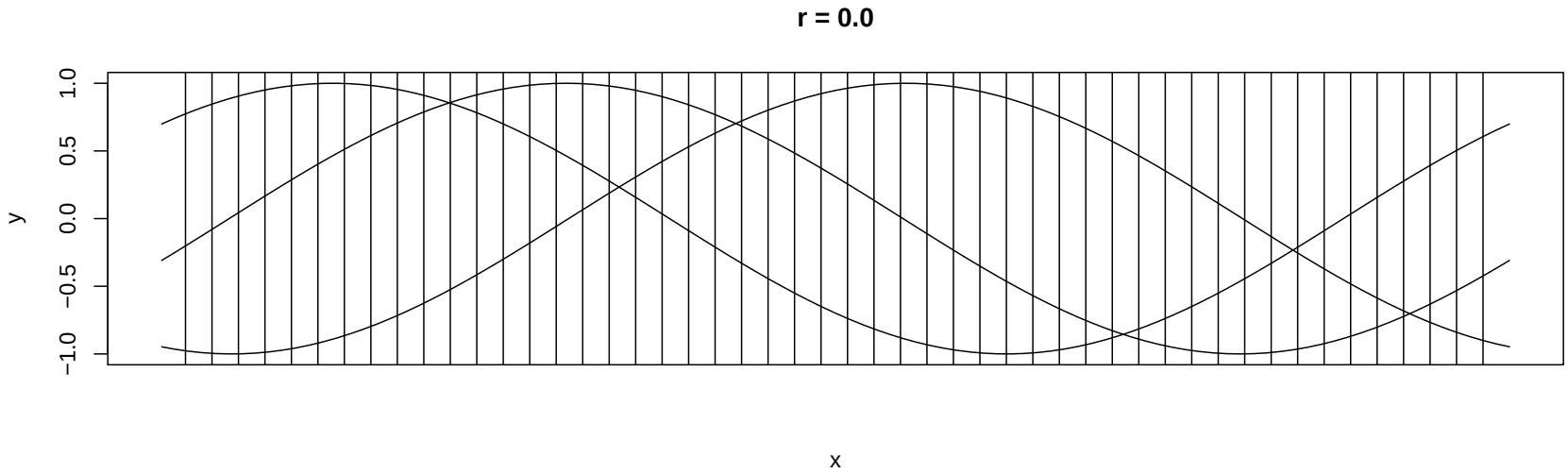
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



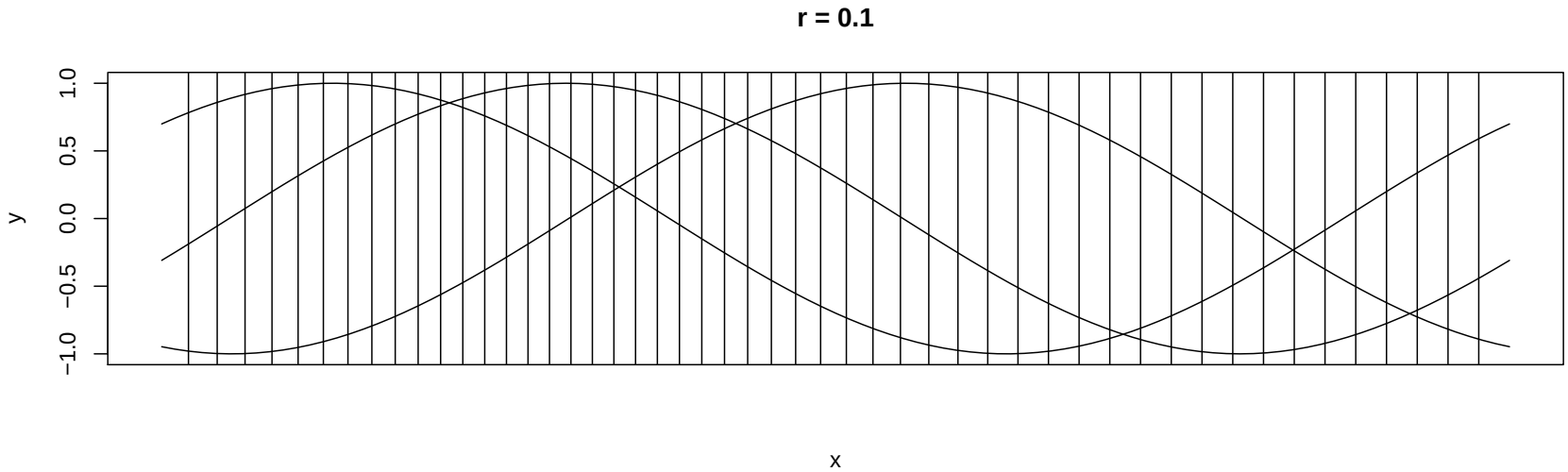
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



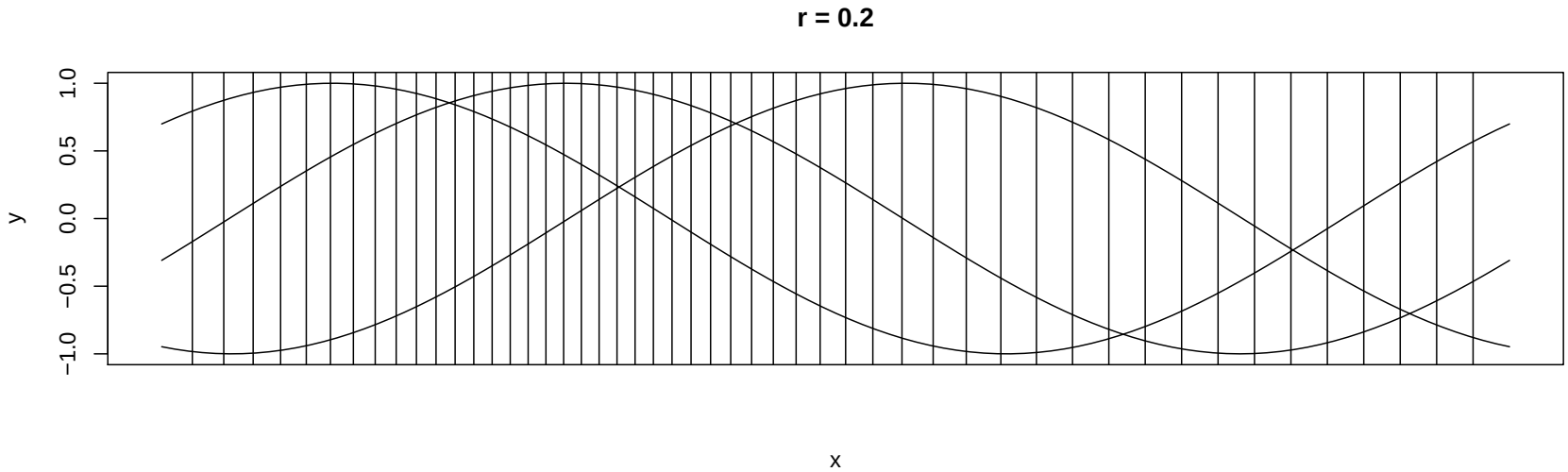
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



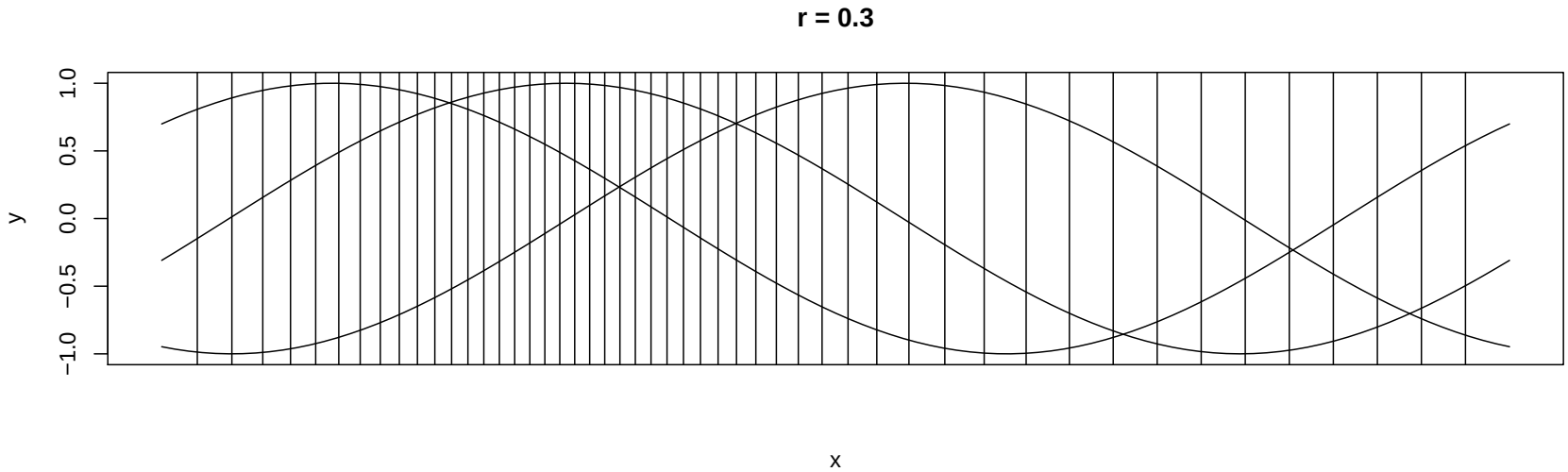
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



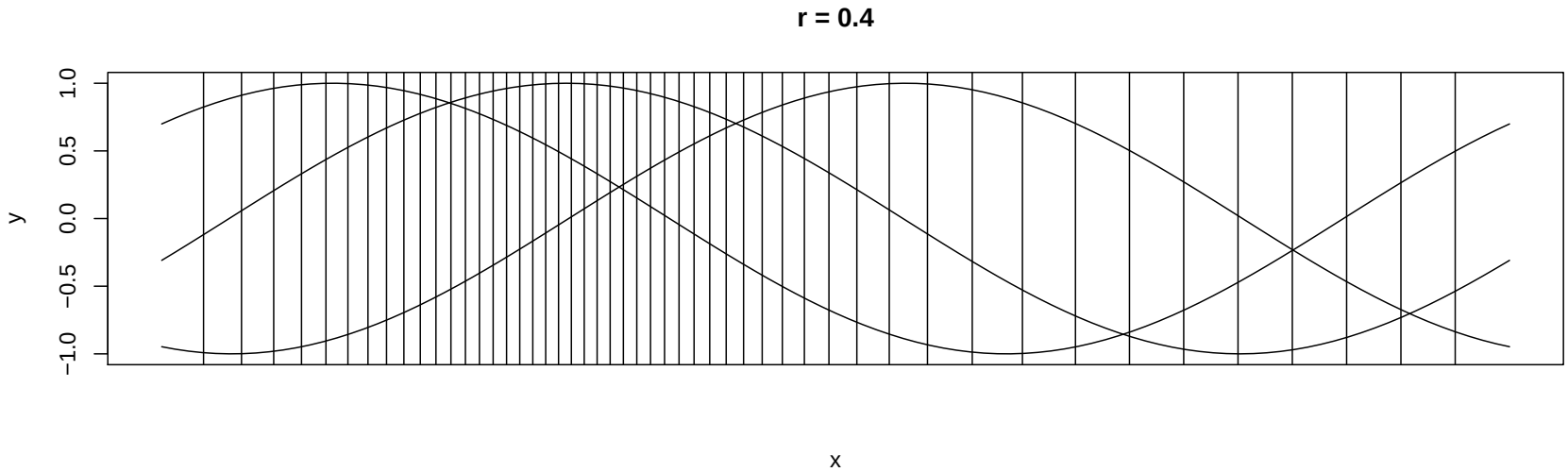
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



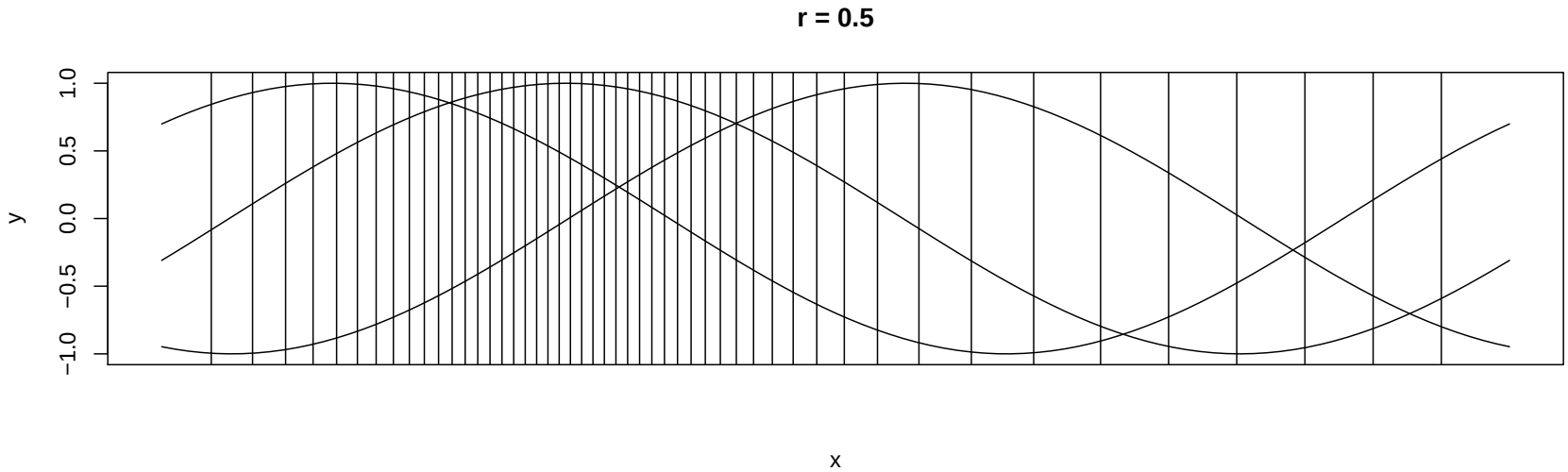
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



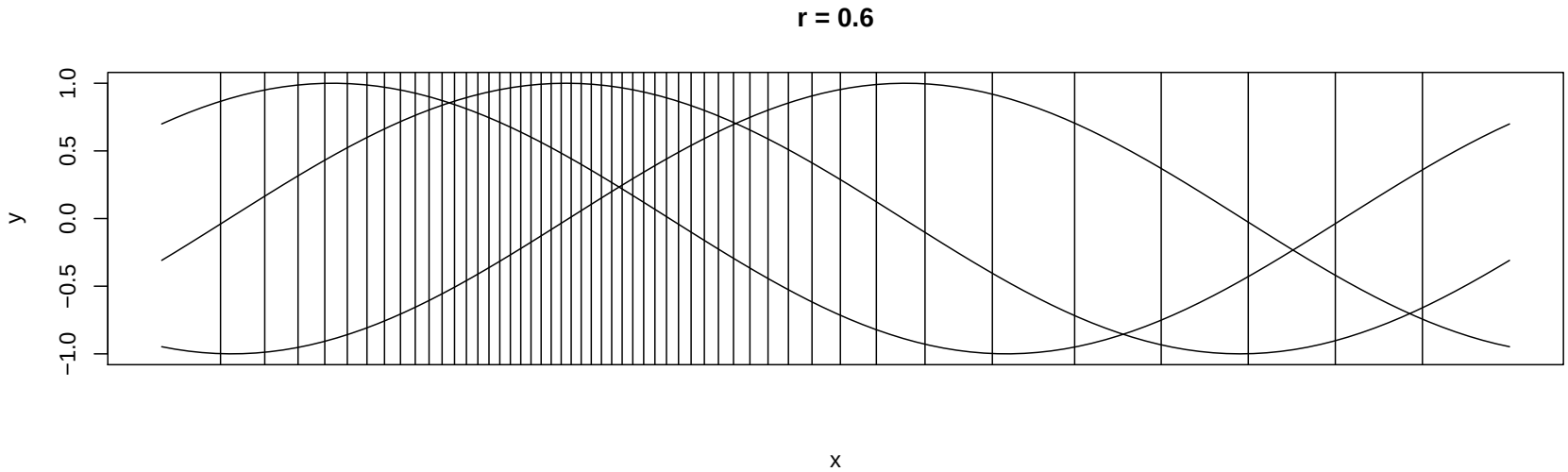
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



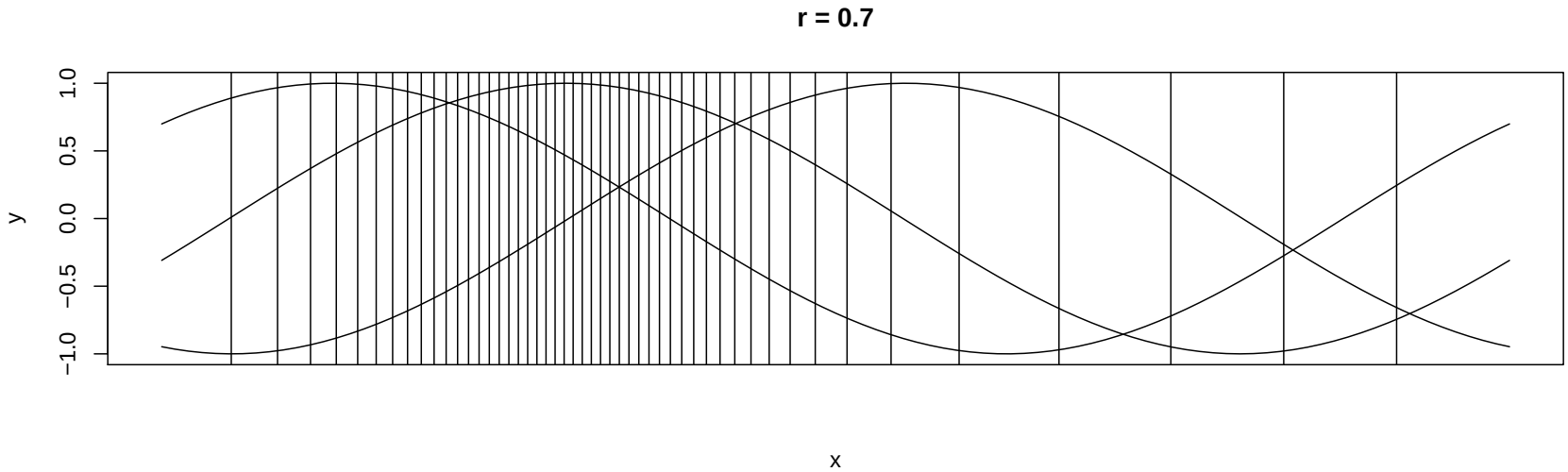
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



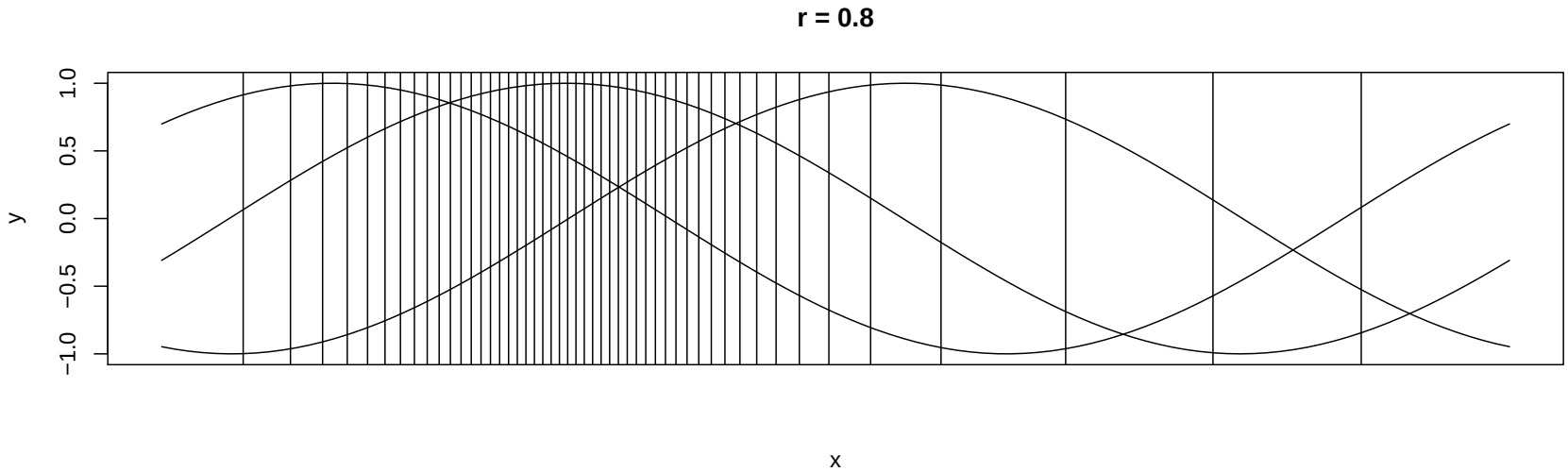
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



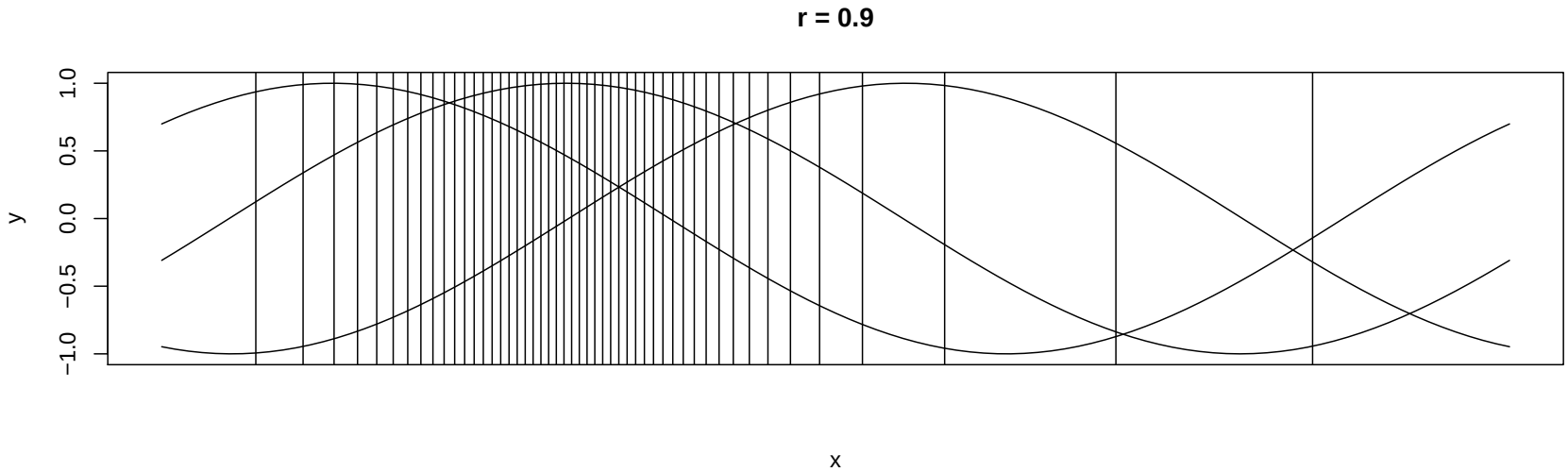
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



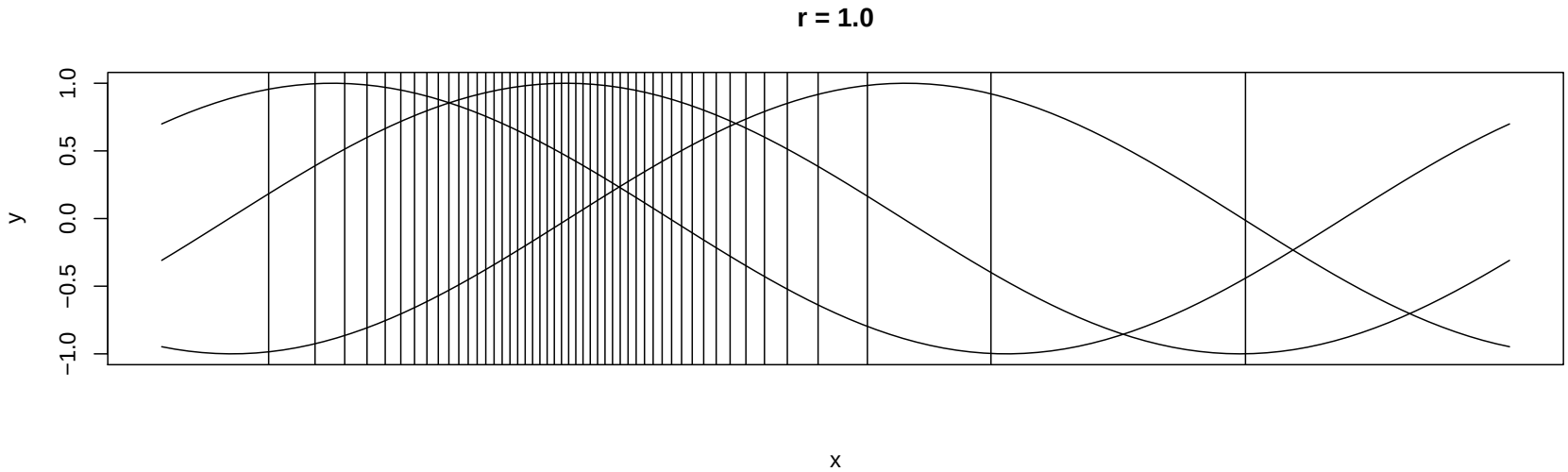
Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



Simulace

1. Apriorní rozdělení polohy maxima: $a(\theta) = p + (1-p)N(\mu_\theta, \sigma_\theta^2)$.
2. Odhadovaná funkce: $m(x) = \cos\{2\pi(x-\theta)\}$, $x \in (0,1)$.
3. Měření v bodech určených hustotami $f_X(x) \propto \{a(x)\}^r$, $r \in [0,1]$.



Parametry simulace

Nastavení:

- podíl rovnoměrného rozdělení $p = 0$,
- mocnina $r \in [0, 1]$,
- $h_n \in [0.005, 0.1]$,
- rozdělení polohy maxima $\mu_\theta = 0.3$, $\sigma_\theta = 0.1$,
- počet pozorování $n \in \{50, 100, 200, 500, 1000\}$,
- $N = 10000$ simulací.

Sledujeme:

- MAD (stř. abs. odchylka, měla by být nejmenší pro $r = 2/3$)
- MSE (stř. čtv. chyba, měla by být nejmenší pro $r = 1/2$)

Parametry simulace

Nastavení:

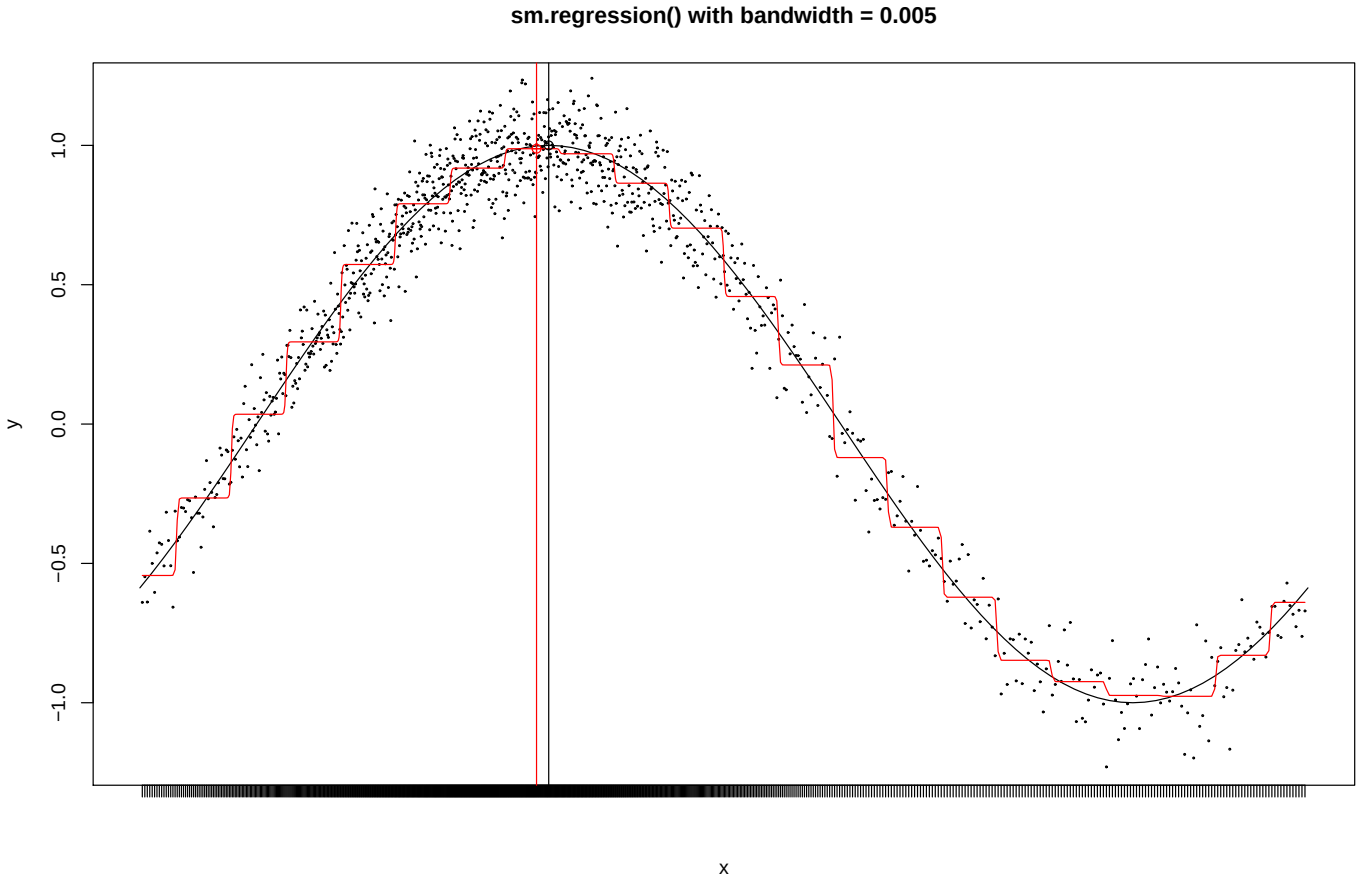
- podíl rovnoměrného rozdělení $p = 0$,
- mocnina $r \in [0, 1]$,
- ~~$h_n \in [0.005, 0.1]$~~, $h_n \in [0.03, 0.1]$
- rozdělení polohy maxima $\mu_\theta = 0.3$, $\sigma_\theta = 0.1$,
- počet pozorování $n \in \{50, 100, 200, 500, 1000\}$,
- $N = 10000$ simulací.

Sledujeme:

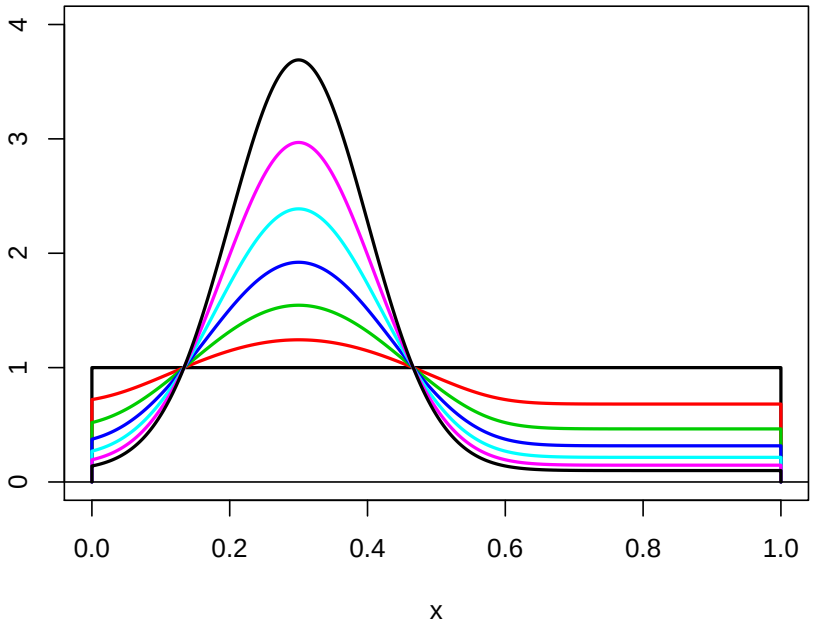
- MAD (stř. abs. odchylka, měla by být nejmenší pro $r = 2/3$)
- MSE (stř. čtv. chyba, měla by být nejmenší pro $r = 1/2$)

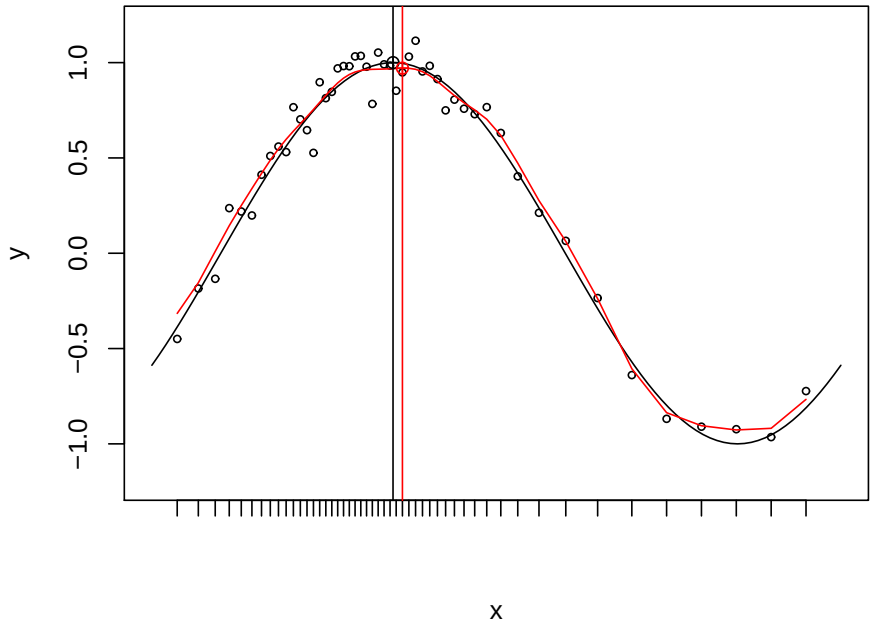
Výpočetní problémy

Funkce `sm.regression()` se *chová divně* pro $h_n \leq 0.02$.



Rozložení bodů na x-ové ose (parametr $r \in [0, 1]$); $p = 0.1$





p = 0 & n = 50

h_n	$r_{opt}(MAD)$	$r_{opt}(MSE)$	$r_{opt}(MAD)$	$r_{opt}(MSE)$
	$\sigma = 0.1$		$\sigma = 1.0$	
0.03	0.42 (1.5189)	0.25 (0.0383)	1.00 (7.3298)	1.00 (0.8670)
0.04	0.17 (1.1932)	0.17 (0.0235)	1.00 (7.0714)	1.00 (0.8114)
0.05	0.12 (0.9978)	0.08 (0.0169)	0.17 (6.6515)	0.96 (0.7799)
0.06	0.04 (0.8850)	0.08 (0.0136)	0.17 (6.3145)	0.25 (0.7614)
0.07	0.04 (0.8322)	0.08 (0.0143)	0.00 (5.8194)	0.17 (0.6646)
0.08	0.00 (0.7665)	0.08 (0.0176)	0.00 (5.5060)	0.12 (0.6213)
0.09	0.00 (0.7570)	0.04 (0.0204)	0.00 (4.9989)	0.12 (0.5413)
0.10	0.00 (0.7295)	0.08 (0.0297)	0.00 (4.6180)	0.12 (0.5103)

p = 0 & n = 100

h_n	$r_{opt}(MAD)$	$r_{opt}(MSE)$	$r_{opt}(MAD)$	$r_{opt}(MSE)$
	$\sigma = 0.1$		$\sigma = 1.0$	
0.03	0.21 (1.1482)	0.21 (0.0213)	0.29 (6.3702)	0.29 (0.7748)
0.04	0.21 (0.8553)	0.21 (0.0127)	0.21 (5.8114)	0.33 (0.6399)
0.05	0.12 (0.7273)	0.12 (0.0096)	0.21 (5.3655)	0.25 (0.5429)
0.06	0.08 (0.6619)	0.12 (0.0105)	0.17 (4.8736)	0.25 (0.4430)
0.07	0.04 (0.6236)	0.12 (0.0138)	0.17 (4.4651)	0.17 (0.4148)
0.08	0.00 (0.5149)	0.12 (0.0211)	0.00 (4.0580)	0.21 (0.3559)
0.09	0.00 (0.4679)	0.08 (0.0172)	0.00 (3.5234)	0.17 (0.3179)
0.10	0.00 (0.4482)	0.12 (0.0445)	0.00 (3.1862)	0.17 (0.3050)

p = 0 & n = 200

h_n	$r_{opt}(MAD)$	$r_{opt}(MSE)$	$r_{opt}(MAD)$	$r_{opt}(MSE)$
	$\sigma = 0.1$		$\sigma = 1.0$	
0.03	0.33 (0.8604)	0.29 (0.0127)	0.33 (5.1893)	0.33 (0.5093)
0.04	0.17 (0.6615)	0.17 (0.0079)	0.25 (4.6271)	0.33 (0.4227)
0.05	0.08 (0.5844)	0.17 (0.0091)	0.29 (4.1369)	0.29 (0.3198)
0.06	0.00 (0.5109)	0.17 (0.0129)	0.25 (3.7794)	0.25 (0.2644)
0.07	0.00 (0.4305)	0.17 (0.0197)	0.04 (3.3601)	0.25 (0.2427)
0.08	0.00 (0.3257)	0.17 (0.0301)	0.04 (2.9379)	0.21 (0.2282)
0.09	0.00 (0.3327)	0.17 (0.0469)	0.00 (2.5716)	0.21 (0.2160)

p = 0 & n = 500

h_n	$r_{opt}(MAD)$	$r_{opt}(MSE)$	$r_{opt}(MAD)$	$r_{opt}(MSE)$
	$\sigma = 0.1$		$\sigma = 1.0$	
0.03	0.21 (0.5723)	0.21 (0.0057)	0.33 (3.7893)	0.42 (0.3098)
0.04	0.17 (0.4831)	0.21 (0.0049)	0.33 (3.3214)	0.38 (0.2080)
0.05	0.00 (0.4736)	0.21 (0.0077)	0.25 (2.8827)	0.33 (0.1716)
0.06	0.00 (0.2734)	0.21 (0.0133)	0.29 (2.6086)	0.29 (0.1269)
0.07	0.00 (0.3653)	0.21 (0.0237)	0.00 (2.2568)	0.25 (0.1291)
0.08	0.00 (0.2374)	0.21 (0.0374)	0.00 (1.9130)	0.29 (0.1430)
0.09	0.00 (0.2236)	0.21 (0.0610)	0.00 (1.5797)	0.25 (0.1411)
0.10	0.00 (0.1395)	0.00 (0.0003)	0.00 (1.4060)	0.00 (0.1237)

p = 0 & n = 1000

h_n	$r_{opt}(MAD)$	$r_{opt}(MSE)$	$r_{opt}(MAD)$	$r_{opt}(MSE)$
	$\sigma = 0.1$		$\sigma = 1.0$	
0.03	0.21 (0.4227)	0.21 (0.0033)	0.42 (3.0237)	0.42 (0.1892)
0.04	0.00 (0.4287)	0.25 (0.0045)	0.38 (2.4993)	0.38 (0.1236)
0.05	0.00 (0.3148)	0.25 (0.0083)	0.33 (2.1529)	0.38 (0.0967)
0.06	0.00 (0.2171)	0.25 (0.0166)	0.00 (1.9071)	0.33 (0.0851)
0.07	0.00 (0.2832)	0.25 (0.0293)	0.00 (1.5748)	0.33 (0.0884)
0.08	0.00 (0.1612)	0.00 (0.0402)	0.00 (1.3904)	0.29 (0.0960)
0.09	0.00 (0.1769)	0.00 (0.0700)	0.00 (1.2926)	0.29 (0.1376)
0.10	0.00 (0.1658)	0.00 (0.0699)	0.00 (1.0508)	0.00 (0.1302)

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.
- Pro menší σ a velké n se nerovnoměrná hustota měření může vyplatit.

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.
- Pro menší σ a velké n se nerovnoměrná hustota měření může vyplatit.

Chtělo by to další simulace:

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.
- Pro menší σ a velké n se nerovnoměrná hustota měření může vyplatit.

Chtělo by to další simulace:

1. menší σ a(nebo) větší n,

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.
- Pro menší σ a velké n se nerovnoměrná hustota měření může vyplatit.

Chtělo by to další simulace:

1. menší σ a(nebo) větší n,
2. komplikovanější funkce $m(.)$,

Simulace: shrnutí výsledků

Shrnutí:

- Rovnoměrná hustota měření většinou dává dobré výsledky.
- Nejvíce záleží na volbě bandwidth h_n.
- Pro menší σ a velké n se nerovnoměrná hustota měření může vyplatit.

Chtělo by to další simulace:

1. menší σ a(nebo) větší n,
2. komplikovanější funkce $m(.)$,
3. nepoužívat `sm.regression()`.

2000 simulations, $n = 5000$, $h_n = 0.002$, $\sigma = 0.01$, $p = 0.1$

r	MAD$\times 100$	MSE$\times 10^4$
0.000	0.489	0.360
0.083	0.459	0.327
0.167	0.474	0.340
0.250	0.447	0.308
0.333	0.432	0.288
0.417	0.435	0.288
0.500	0.435	0.291
0.583	0.442	0.299
0.667	0.444	0.303
0.750	0.431	0.291
0.833	0.423	0.280
0.917	0.426	0.300
1.000	0.433	0.312

2000 simulations, $n = 100$, $h_n = 0.04$, $\sigma = 0.01$, $p = 0.1$

r	MAD$\times$100	MSE$\times 10^4$
0.000	0.424	0.281
0.083	0.395	0.247
0.167	0.379	0.230
0.250	0.365	0.209
0.333	0.360	0.210
0.417	0.360	0.215
0.500	0.342	0.195
0.583	0.343	0.206
0.667	0.360	0.242
0.750	0.378	0.305
0.833	0.422	0.498
0.917	0.448	0.731
1.000	0.493	1.181

Je GM odhad vhodný?

Gasser-Müllerův odhad:

$$
\begin{aligned}
m_h^{GM}(x) &= \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(x-u)du Y_i \\
&\approx \sum_{i=1}^{n} (s_i - s_{i-1}) K_h(x - x_i) = \sum_{i=1}^{n} (x_{i+1} - x_{i-1}) K_h(x - x_i)/2
\end{aligned}
$$

Modifikovaný GM odhad, např. Chu (1985).

Chu, C. K. (1985). A new version of the Gasser-Mueller estimator, *Journal of Nonparametric Statistics*, 3: 187–193.

Závěr

- Připomněli jsme GM odhad, který se používá pro neparametrickou regresi s pevným designem.

Závěr

- Připomněli jsme GM odhad, který se používá pro neparametrickou regresi s pevným designem.
- Odvodili jsme optimální rozložení měření pro odhad polohy maxima, které vyjde (pro MSE) stejně jako AIMSE (Müller, 1984).

Závěr

- Připomněli jsme GM odhad, který se používá pro neparametrickou regresi s pevným designem.
- Odvodili jsme optimální rozložení měření pro odhad polohy maxima, které vyjde (pro MSE) stejně jako AIMSE (Müller, 1984).
- Obdobně lze odvodit optimální rozložení měření pro odhad polohy nulových bodů derivací funkce $m(.)$ (odhad polohy maxima nebo minima $\approx$ odhad nulového bodu první derivace).

Literatura

Gasser, Th. & Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method, *Scand. J. Statist.* 11: 171–185.

Müller, H.-G. (1984). Optimal designs for nonparametric kernel regression, *Statistics & Probability Letters* 2: 285–290.

Müller, H.-G. (1984). Smooth optimum kernel estimators of regression curves, densities and modes, *Ann. Statist.* 12: 766–774.

Müller, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions, *Scand. J. Statist.* 12: 221–232.

Chu, C. K. (1985). A new version of the Gasser-Mueller estimator, *Journal of Nonparametric Statistics*, 3: 187–193.