Jádrové odhady regresní funkce pro korelovaná data

Lajdová D., Koláček J., Horová I.

Ústav matematiky a statistiky MÚ Brno

Finanční matematika v praxi III., Podlesí

3.9.-4.9. 2013

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
AMathNet síť pro transfer znalostí v aplikované matematice
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

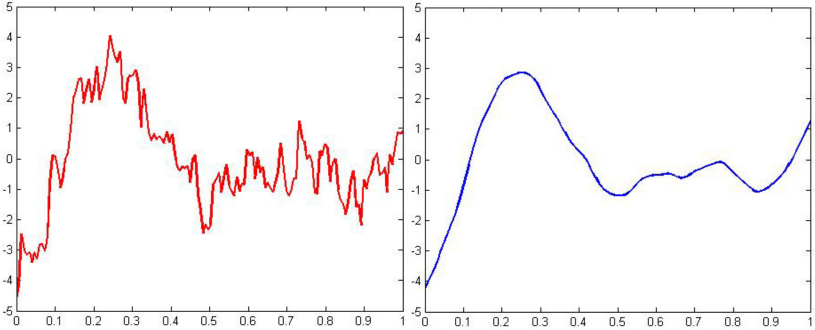
Obsah

Motivace

Co se snažíme získat?

Motivace

Co se snažíme získat?



Regresní model

Standardní regresní model

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde

- m - neznámá regresní funkce
- Y_i - měřená data
- x_i - body, ve kterých se provádí měření
- ε_i - chyby

Jádro

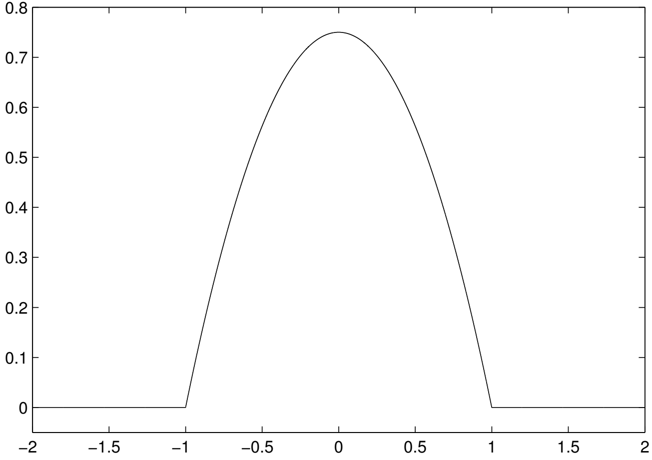
Co je to jádro?

Jádro

Co je to jádro?

- Jádro K je reálná funkce splňující podmínky:
  - $K \in \mathrm{Lip}[-1,1], \quad K(-1) = K(1) = 0$
  - $\mathrm{support}(K) = [-1,1]$
  - $\int_{-1}^{1} x^j K(x)\mathrm{d}x = \begin{cases} 1 & j = 0 \\ 0 & j = 1 \\ \beta_2 \neq 0 & j = 2 \end{cases}$

Epanečnikovo jádro



Jak se používá?

Jak se používá?

Nadaraya-Watsonův odhad regresní funkce

$$\widehat{m}(x,h) = \sum_{i=1}^{n} \frac{K_h(x_i - x)}{\sum\limits_{j=1}^{n} K_h(x_j - x)} Y_i,$$

kde

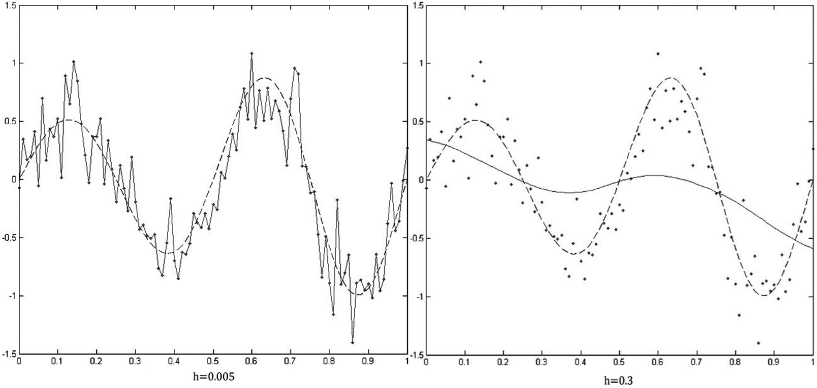
- $K_h(.) = \frac{1}{h} K\left(\frac{.}{h}\right)$
- h - vyhlazovací parametr

Vyhlazovací parametr

Co se stane, když ho zvolíme špatně?

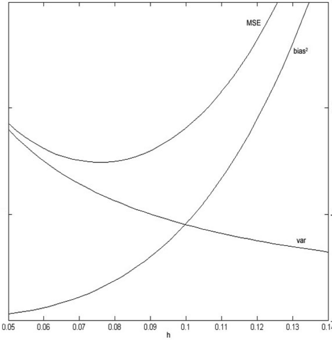
Vyhlazovací parametr

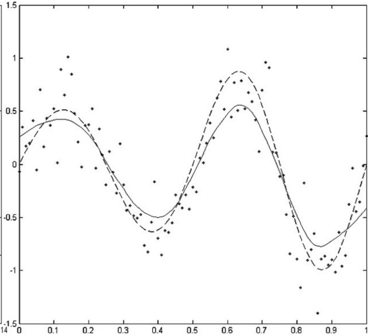
Co se stane, když ho zvolíme špatně?



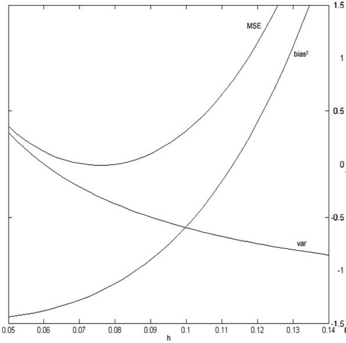
Jaká je optimální hodnota?

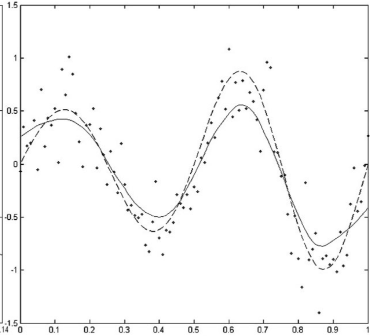
Jaká je optimální hodnota?





Jaká je optimální hodnota?





$$h_{opt} = \left(\frac{\sigma^2 \int_{-1}^{1} K^2(x)\mathrm{d}x}{n\beta_2^2 A_2} \right)^{1/5}, \quad A_2 = \int_0^1 m''(x)^2 \mathrm{d}x$$

Krosvalidační metoda

Jak zvolit správné h?

Krosvalidační metoda

Jak zvolit správné h?

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}\left(\widehat{m}_{-j}(x_j, h) - Y_j\right)^2 \longrightarrow min,$$

kde $\widehat{m}_{-j}(x_j, h)$ je odhad $m(x_j, h)$, kde je x_j smazáno

Regresní model s korelovanými chybami

Jak se náš model změní?

Regresní model s korelovanými chybami

Jak se náš model změní?

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

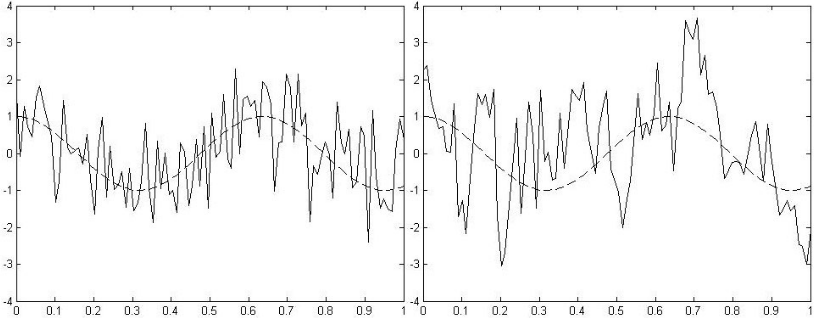
kde

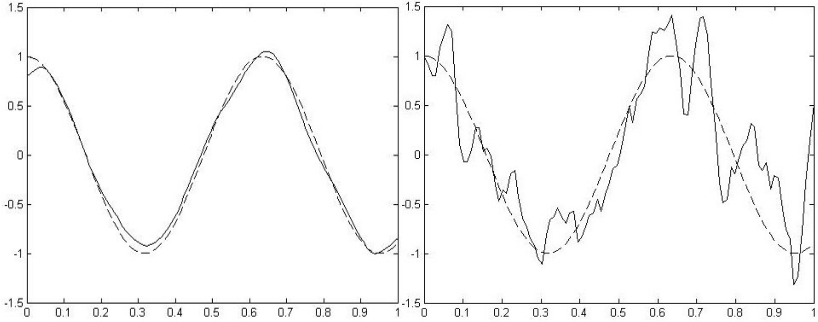
- ε_i - neznámý kauzální ARMA proces, neboli

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2$$
$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \gamma_{|i-j|} = \sigma^2 \rho_{|i-j|}, \quad \mathrm{corr}(\varepsilon_i, \varepsilon_j) = \rho_{|i-j|}$$

Proč prostě nepoužijeme předchozí metodu?

Proč prostě nepoužijeme předchozí metodu?




1.5
1
0.5
0
-0.5
-1
-1.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Plug-in přístup

Můžeme nějak využít optimální hodnotu h?

Plug-in přístup

Můžeme nějak využít optimální hodnotu h?

$$h_{opt} = \left(\frac{S \int_{-1}^{1} K^2(x) \mathrm{d}x}{n \beta_2^2 A_2} \right)^{1/5},$$

$$S = \sum_{k=-\infty}^{\infty} \gamma_k, \; A_2 = \int_0^1 m''(x)^2 \mathrm{d}x$$

Plug-in přístup

Můžeme nějak využít optimální hodnotu h?

$$h_{opt} = \left(\frac{S \int_{-1}^{1} K^2(x)\mathrm{d}x}{n\beta_2^2 A_2} \right)^{1/5},$$

$$S = \sum_{k=-\infty}^{\infty} \gamma_k, \; A_2 = \int_0^1 m''(x)^2 \mathrm{d}x$$

$$\widehat{h}_{PI} = \left(\frac{\hat{S} \int_{-1}^{1} K^2(x)\mathrm{d}x}{n\beta_2^2 \widehat{A}_2} \right)^{1/5}$$

Jak odhadnout S a A_2?

Jak odhadnout S a A_2?

- odhad S

Jak odhadnout S a A_2?

- odhad S
    - odhadnutím chyb pomocí nějakého počátečního vyhlazení (průměr, přehlazení, ...)

Jak odhadnout S a A_2?

- odhad S
  - odhadnutím chyb pomocí nějakého počátečního vyhlazení (průměr, přehlazení, ...) - nefunguje

Jak odhadnout S a A_2?

- odhad S
  - odhadnutím chyb pomocí nějakého počátečního vyhlazení (průměr, přehlazení, ...) - nefunguje
  - odhadnutím spektrální hustoty

$$f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} e^{-it\lambda} \gamma_t$$

Jak odhadnout S a A_2?

- odhad S
  - odhadnutím chyb pomocí nějakého počátečního vyhlazení (průměr, přehlazení, ...) - nefunguje
  - odhadnutím spektrální hustoty

$$f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} e^{-it\lambda} \gamma_t$$

- odhad A_2

Jak odhadnout S a A_2?

- odhad S
  - odhadnutím chyb pomocí nějakého počátečního vyhlazení (průměr, přehlazení, ...) - nefunguje
  - odhadnutím spektrální hustoty

$$f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} e^{-it\lambda} \gamma_t$$

- odhad A_2
  - pracuje se na tom - viz Koláček (2008)

Částečná krosvalidace (PCV)

Dá se nějak obejít korelace?

Částečná krosvalidace (PCV)

Dá se nějak obejít korelace?

- rozdělení pozorování do g skupin tak, že vezmeme každé g-té pozorování
- minimalizace průměru obyčejných CV každé skupiny

$$CV^*(h) = g^{-1} \sum_{k=1}^{g} CV_k(h) \longrightarrow min$$

Částečná krosvalidace (PCV)

Dá se nějak obejít korelace?

- rozdělení pozorování do g skupin tak, že vezmeme každé g-té pozorování
- minimalizace průměru obyčejných CV každé skupiny

$$CV^*(h) = g^{-1} \sum_{k=1}^{g} CV_k(h) \longrightarrow min$$

h minimalizující $CV^*(h)$

$$\widehat{h}^*_{CV} = \arg\min CV^*(h) \Longrightarrow \widehat{h}_{PCV(g)} = g^{-1/5} h^*_{CV}$$

Částečná krosvalidace (PCV)

Dá se nějak obejít korelace?

- rozdělení pozorování do g skupin tak, že vezmeme každé g-té pozorování
- minimalizace průměru obyčejných CV každé skupiny

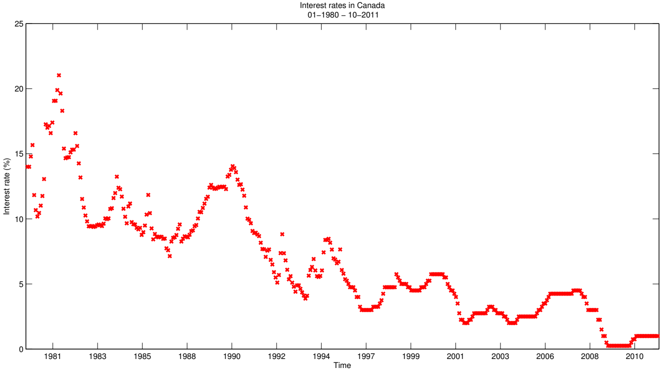
$$CV^*(h) = g^{-1} \sum_{k=1}^{g} CV_k(h) \longrightarrow min$$

h minimalizující $CV^*(h)$

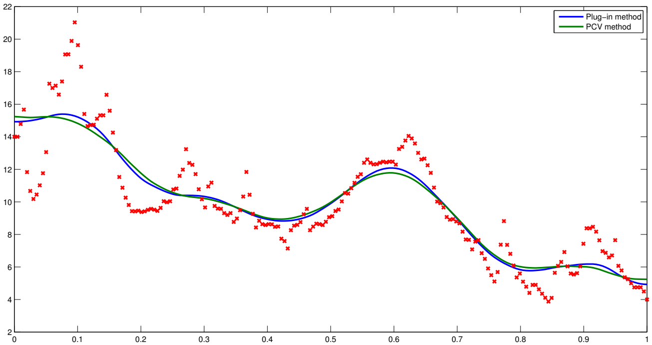
$$\widehat{h}^*_{CV} = \arg\min CV^*(h) \Longrightarrow \widehat{h}_{PCV(g)} = g^{-1/5} h^*_{CV}$$

Ale volba g závisí γ_k, K, A_2

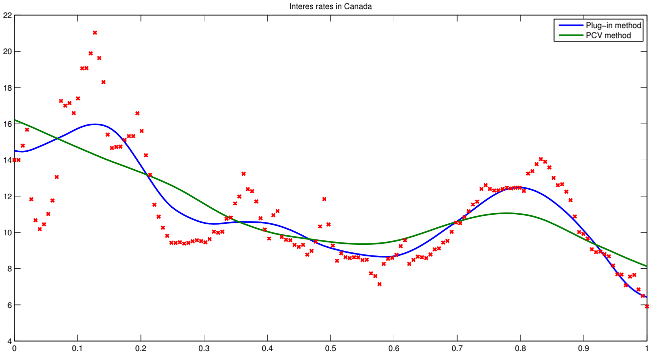
Úroková míra CB Kanda - 1980-2011



Prvních 200 měsíců



Prvních 150 měsíců



References

Altman, N.S.
Kernel Smoothing of Data With Correlated Errors
Journal of the American Statistical Association, Vol.85, No.411 (Sep., 1990), pp.749-759

Bowman, A.W. - Azzalini, A.
Applied Smoothing Techniques for Data Analysis
Clarendon Press, Oxford, 1997

Chu, C.-K. - Marron, J.S.
Comparison of two bandwidth selectors with dependent errors
The Annals of statistics, Vol.19, No.4 (1991), pp.1906-1918

Härdle, W. - Vieu, P.
Kernel Regression Smoothing of Time Series
Journal of Time Series Analysis, Vol.13, No.3, (May, 1992), pp.209–232

References

Herrmann, E. - Gasser, T. - Kneip, A.
Choice of Bandwidth for Kernel Regression when Residuals are Correlated
Biometrika, Vol.79, No.4 (Dec., 1992), pp.783-795

Koláček, J.
Plug-in method for nonparametric regression
Computational Statistics, Vol.23, No.1 (Jan., 2008), pp.63-78

Opsomer, J. - Wang, Y. - Yang, Y.
Nonparametric Regression with Correlated Errors
Statistical Science, Vol.16, No.2 (May, 2001), pp.134-153

Děkuji za pozornost