Posudek oponenta diplomové práce

Název: Neparametrické regresní odhady
Autor: Martin Měsíček

Shrnutí obsahu práce

Autor v první kapitole představuje lokálně polynomiální odhady a jejich vlastnosti. Zaměřuje se přitom zejména na lokálně lineární odhady. Výklad doplňuje představením „převáženým“ Nadaraya-Watsonovým odhadem, který je v jistém smyslu asymptoticky ekvivalentní lokálně lineárnímu odhadu. Výsledky z této první kapitoly pak využívá ve druhé kapitole, kde se velmi podrobně věnuje problematice odhadu podmíněného rozptylu.

Téma práce. Autor si z vícero možností, které umožňovalo zadání, vybral problematiku odhadování podmíněného rozptylu. Téma je dle mého názoru svou náročností i obsahem přiměřené diplomové práci a bylo autorem splněno.

Vlastní příspěvek. Autor utřídil a sepsal poznatky z nadprůměrně mnoha zdrojů, přičemž na poměry diplomové práce šel opravdu do hloubky. Dále sám dokázal některé vlastnosti odhadů, které dosud nebyly takto podrobně odvozeny. Zajímavé také je teoretické i simulační porovnání odhadů podmíněného rozptylu v „hraničních“ bodech.

Matematická úroveň. Matematická úroveň práce je solidní. Je sympatické, že se autor pustil i do samostatného odvozování, které je místy docela technicky náročné. Bohužel na několika místech si úplně neuvědomil složitost problematiky a některá odvození nejsou úplně matematicky korektní (podrobnosti viz níže).

Práce se zdroji. Autor nic doslovně nepřejímal a snažil se psát vlastními slovy. V práci jsou také odcitovány všechny použité zdroje. Práci by však velmi prospělo, pokud by byl autor podrobnější při citování zdrojů. Pro čtenáře je značně nepříjemné, že většina citací je pouze odkazem na článek nebo na knihu a tudíž se citovaná tvrzení velmi těžko a pracně dohledávají.

Formální úprava. Formální úroveň práce je velmi slušná s jistým drobnými nedokonalostmi a ne nijak zásadně velkým počtem překlepů. Je škoda, že se seznam použitého značení nedostal do obsahu. Navíc tento seznam může čtenář objevit pouze náhodou, protože je uveden až na konci práce.

Celkové hodnocení práce

Předložená práce patří dle mého názoru k nadprůměrným, co se týká rozsahu, hloubky i zaujetí autora pro problematiku. Bohužel však musím říct, že méně by v tomto případě znamenalo více. Je to škoda, protože výsledkem mohla být velmi hezká práce, která se dobře čte. Takto celkový dojem ze zajímavé práce bohužel sráží výše uvedené matematické a formální nedostatky. Pro čtenáře práce je také značně nepříjemné, že přestože je použitý počet symbolů značný, tak se autor příliš nenamáhá připomínat, kde se daný symbol zavedl.

Několik připomínek

1. $\mathbf{6}^4$: Pokud se nepletu, tak symbol $\mathbf{X}$ nebyl zaveden.

2. **7**[7]: Proč leží NW-odhad mezi minimem a maximem pozorované odezvy?

3. **8**: V citované knize Fan a Gijbels (1996) nejsem schopný dohledat uvedený výsledek o vychýlení a rozptylu $\hat{m}_n^{LL}(x_0)$.

4. **9**: V Definici 4 by stálo za vysvětlení, co znamená symbol $\inf_{\hat{m}_n}$

5. **9**: Aby byla Věta 1 v pořádku, tak by se musel mírně modifikovat odhad $\hat{m}_n^{LL}$(viz citovaný článek Fan, 1993).

6. **10**: Myslím, že interpretovat eficienci jako „*kolik procent dostupných dat odhad využívá*“ může být poněkud zavádějící.

7. **11**$_{1-2}$: Nemohu souhlasit s tvrzením, že v případě aplikací má f_X vždy omezený nosič.

8. **12**$_8$: Podmínka $\limsup_{u\to-\infty} |K(u)\,u^5| < \infty$ asi obsahuje nějaký překlep.

9. **13**: V Definici 9 má zřejmě být $\mathrm{ef}[\hat{m}_n^*(x_0)|\boldsymbol{X}]$.

10. **14**$_1$: Ve vzorci pro $h_{LL,n}^{opt}(x_0)$ je překlep.

11. **16**: Funkce ψ pro Riceho metodu očividně nesplňuje Definici 10.

12. **16**: Vynechání Věty 6 (a jejího důkazu v druhé kapitole) je dle mého názoru pěkný příklad toho, kde autor mohl šetřit síly. Věta je zformulována bez předpokladů. A také vágně, protože z kontextu není úplně jasné, co je podmíněné asymptotické vychýlení funkce $P(h_n)$. Dále je nešťastné napsáno, že věta se dokáže později, ale není řečeno kde. Samotný důkaz této věty pak není matematicky úplně korektní. A přitom tato věta není důležitá ani pro teoretickou ani pro simulační část práce.

13. **17**[10]: Myslím, že symbol h_n^{aopt} nebyl zaveden.

14. **17**$_{14}$: Akaikeho

15. **18**: Metoda křížového ověřování by se dala najít i leckde jinde než pouze v zápiscích z přednášky NMST434.

16. 2. Kapitola: V této kapitole autor značí derivace funkce m tečkou, ale v předchozí kapitole používal čárku.

17. **20** Podmínka 1(d): Asi by bylo lepší, aby se autor již zde omezil na nezávislé stejně rozdělená pozorování, než aby používal termíny jako mixing koeficient, který ani nevysvětlí. Podobné omezení by také jistě stačilo při formulaci podmínek v kapitole 2.3.

18. **21**[2]: Není úplně zřejmé, co autor myslí pojmem „asymptoticky ekvivalentní“.

19. **22**: Použití dvojteček za výrazy pro vychýlení a rozptyl odhadu $\widehat{\sigma}^2_{d:LL+LL,n}(x)$ bude zřejmě nějaký překlep.

20. **22**$_5$: Pokud tomu správně rozumím, tak autor přešel od značení $\hat{v}_n^{LL}(x)$ ke značení $\hat{v}_{LL}(x)$.

21. **23**: Je nějaký důvod, proč je kriteriální funkce optimalizační úlohy psána jinak než v kapitole 1.3.3?

22. **25**: Věta 8 zformulovaná pro f_X s nosičem [a,b]. Věta 16, která s ní úzce souvisí, pak pro nosič [0,1].

23. **26**: Pokud se autor namáhal s formulováním Věty 9, tak by bylo dobré podrobněji vysvětlit, k čemu je tato věta užitečná. Navíc ve formulaci věty není jasné, jak rozumět symbolu h v druhé části věty.

24. **27**: Nepodařilo se mi zjistit, proč autor uvažuje odhad podmíněného rozptylu v pevných bodech $x_1, \ldots, x_n$ a nikoliv pouze v jednom pevném bodě. Poněkud to komplikuje přehlednost následujících důkazů.

25. **28**: S využitím

$$[|Y_n - g(X_n)| > \varepsilon] \subset [|Y_n - \mathsf{E}\ Y_n| > \tfrac{\varepsilon}{2}] \cup [|\ \mathsf{E}\ Y_n - g(X_n)| > \tfrac{\varepsilon}{2}]$$

mohl být důkaz Lemmatu 11 kratší a přehlednější.

26. **30** Lemma 14: Stálo by za vysvětlení, co rozumí autor symbolem $O_P(\frac{1}{nh_{n,2}})_{n\times n}$, resp. symbolem $O_P(\frac{1}{nh_{n,2}})_{n\times 1}$. Na kolik rozumím důkazu Lemmatu 14, tak tento symbol značí, že se jedná o matici řádu $n \times n$ (resp. n rozměrný vektor), jejíž (jehož) každý prvek je řádu $O_P(\frac{1}{nh_{n,2}})$. To je však kámen úrazu v důkazu Vět 15 a 16.

27. **33**: Poslední rovnost v důkazu Věty 6 by si zasloužila podrobnější vysvětlení. Navíc zde bude taky problém, popsaný v následující připomínce.

28. **34**: Nakolik se v úvahách autora orientuji, tak důkaz Věty 15 od spojení „*Stačí ukázat, že zbylé členy ve ...*“ přestává být korektní. Problém je v tom, že autor pracuje s maticemi typu $O_P(\frac{1}{r_n})_{n\times n}$, resp. vektory typu $O_P(\frac{1}{r_n})_{n\times 1}$ (kde $r_n \to \infty$) jako by pro maxima absolutních hodnot prvků těchto matic (vektorů) platil stejný řád konvergence. To však pro používané matice (vektory) nebylo dokázáno. Podobně pro matice typu $o_P(\frac{1}{r_n})_{n\times n}$, resp. vektory typu $o_P(\frac{1}{r_n})_{n\times 1}$. Stejný problém je také v důkazu Věty 16.

29. **37**: Je škoda, že autor nijak neupozornil na souvislost Věty 16 a Věty 8.

30. **37 – 39**: V důkazu Věty 16 vypadla ve značení pro $\hat{\sigma}^2_{XP}$ druhá mocnina.

31. **43**: Domnívám se, že z hlediska aplikací by bylo zajímavější uvažovat následující modifikaci FY odhadu:

$$\widetilde{\sigma}^2_{FYM,n}(x) = \begin{cases} \hat{\sigma}^2_{FY,n}(x), & \hat{\sigma}^2_{FY,n}(x) \geq 0, \\ \hat{\sigma}^2_{XP,n}(x), & \hat{\sigma}^2_{FY,n}(x) < 0. \end{cases}$$

Ono mít v aplikacích nulový odhad rozptylu je také krajně podezřelé.

32. **44**: S Větou 18 může být zádrhel v tom, že není jasné, zda příslušné rozptyly a MSE existují. Navíc je tato věta asi celkem nadbytečná, protože z konstrukce a z nezápornosti podmíněného rozptylu je zřejmé, že

$$|\hat{\sigma}^2_{FYM,n}(x) - \sigma^2(x)| \leq |\hat{\sigma}^2_{FY,n}(x) - \sigma^2(x)|.$$

V tomto smyslu by bylo třeba také upravit poznámky k Větě 18 na následující straně.

33. Kapitola 2.7 a 2.8: Oceňuji, že autor vložil do přílohy práce také kód použitý v těchto částech práce. A to i přes těžkosti, které jsou s tím spojené díky Opatření rektora č. 13/2017. Je ovšem škoda, že kód v případě ilustračních dat vyžaduje zásah uživatele, protože obsahuje cestu k adresáři na počítači autora. Pokud data z nějakého důvodu nemohla být přidána do přílohy, tak mohl zdrojový kód obsahovat příkazy k instalace potřebného balíčku, který data obsahuje.

34. **50** Kapitola 2.8: Trochu mě zaráží, že zatímco v teoretické části pracuje autor s náhodným designem, tak v simulační studii použil (bez jakéhokoliv vysvětlení) ekvidistantní pevný design. Bylo by zajímavé zjistit, jak by dopadlo porovnání v případech náhodného designu, když f_X není hustota rovnoměrného rozdělení.

35. **57 – 59**: Práce obsahuje drobné formální nedokonalosti v seznamu použité literatury. Tak např. u Fan a kol. (1995) chybí bližší informace o článku. Rušivě také působí, že u článku Hall and Marron (1991) autor použil zkrácený název časopisu, zatímco u všech ostatních časopiseckých zdrojů jsou názvy časopisů uváděný nezkrácené. U časopisu *Journal of the Royal Statistical Society* chybí rozlišení, o kterou *Series* (A, B, C, D) se jedná. U internetového zdroje by to chtělo uvádět také datum přístupu...

Otázky k obhajobě

1. Odpovězte na připomínku č. 4.
2. Odpovězte na připomínku č. 18.
3. Odpovězte na připomínku č. 23.

Závěr

I přez výše uvedené výhrady se domnívám, že předložená práce splňuje všechny požadavky kladené na diplomovou práci na oboru Pravděpodobnost, matematická statistika a ekonometrie a doporučuji ji za ni uznat.

Ing. Marek Omelka, Ph.D.
KPMS MFF UK
4. srpna 2017