

POSUDEK VEDOUCÍHO DIPLOMOVÉ PRÁCE

Název: Robustní regularizovaná regrese

Autor: Jakub Krett

SHRNUTÍ OBSAHU PRÁCE

Předložená práce se věnuje odhadům parametrů pro lineární regresní model, které kombinují robustnost vůči extrémně odlehlým hodnotám s regularizací. Jako novinka je zde navržen nový odhad LWS-lasso, který představuje L_1 -regularizovanou obdobu odhadu metodou nejmenších vážených čtverců (LWS).

V kapitole 1 je popsán lineární regresní model a běžně používané odhady. V kapitole 2 jsou pak popsány některé robustní odhady a regularizované odhady (lasso odhad). Přirozeným spojením myšlenky robustnosti a regularizace navrhli Alfons a kol. (2017) LTS-lasso odhad. Jeho váženou obdobou je LWS-lasso odhad navržený v předložené práci v kapitole 2.7. Protože odvození jeho teoretických vlastností by bylo příliš složité, je zbytek práce zaměřen experimentálně. Výpočty pak následují v kap. 3 (reálný datový soubor s ekonomickou motivací) a v kap. 4 (simulace).

CELKOVÉ HODNOCENÍ PRÁCE

Téma práce. Téma hodnotím jako přiměřeně náročné. LWS-lasso odhad sice představuje metodologickou novinku, nicméně cílem nebylo jej studovat teoreticky. Autor splnil své zadání návrhem algoritmu pro tento odhad a provedením simulací, které přinášejí argumenty ve prospěch tohoto nově navrženého odhadu.

Vlastní příspěvek. Autorovým přínosem je Algoritmus 1 pro přibližný výpočet LWS-lasso odhadu, který implementoval v prostředí R. Za hlavní výsledek experimentů lze vzít doporučení, že LWS-lasso se zdá být vhodnou (nejvhodnější) metodou zejména pro kontaminovaná data při větším počtu regresorů. Za zajímavou novinku považuji autorovo zjištění, že odhad LTS-lasso implementovaný v balíčku robustHD (jehož vnitřní kód ovšem není veřejně dostupný) dává velmi podivné, nesmyslné výsledky. Následně jsem nezávisle ověřil na jiných datech, že v některých situacích je tomu opravdu tak.

Matematická úroveň. Z hlediska matematické náplně je v práci (v appendixu) rozpracováno odvození bodu selhání LTS-odhadu (podrobné rozepsání důkazu dle článku Alfons a kol. (2017)).

Práce se zdroji. Autor správně cituje použité zdroje, převzaté poznatky doplňuje vlastními postřehy a úvahami.

Formální úprava. Z formálního hlediska hodnotím práci kladně.

PŘIPOMÍNKY

1. Str. 5. Ve vzorci pro $\varepsilon_n^*(\hat{\beta}, Z)$ je zřejmě potřeba nahradit $|\hat{\beta}(\tilde{Z}) - \hat{\beta}(Z)|$ výrazem $\|\hat{\beta}(\tilde{Z}) - \hat{\beta}(Z)\|$.
2. Str. 7, Definice 4. Nesouhlasím, že Def. 4 nedává jednoznačnou definici odlehlého pozorování. Autor má patrně na mysli, že definici nelze v praxi ověřit a je třeba použít některý z existujících přibližných postupů. Nadto by zde ještě bylo vhodné rozlišovat mezi *good leverage points* a *bad leverage points*, jak je zvykem v anglicky psané literatuře.
3. Nejednotné značení u pořádkových statistik: na str. 11 je zavedeno značení, které při $\rho(x) = x^2$ odpovídá $u_{(i:n)}^2$, na str. 13 pak je totéž značeno jako $u_{(i)}^2$.
4. Str. 16. Uvedený vztah pro $\psi_2'(x)$ platí, ale je zde třeba vyloučit situaci $x = \tau$.
5. Str. 17. Vztah pro ψ_4' nesouhlasí: vzhledem k tomu, že platí

$$\frac{d \operatorname{erf}(x)}{dx} = \frac{2}{\sqrt{\pi}} e^{-x^2}, \quad (1)$$

musí pak platit

$$\psi_4'(x) = \frac{d}{dx} (1 - \operatorname{erf}(x)) = -\frac{2}{\sqrt{\pi}} e^{-x^2}. \quad (2)$$

ZÁVĚR

Jakub Krett, který studuje obor Finanční a pojistná matematika, se seznámil se základy robustní statistiky a vypracoval diplomovou práci na zajímavé téma, které má publikační potenciál. V teoretické části se snažil doplňovat

prezentované poznatky svými vlastními komentáři. Celkově Jakub Krett danou problematiku zvládl, v práci neshledávám podstatné chyby a myslím, že ho práce na daném tématu bavila. Práci doporučuji uznat jako diplomovou práci.

RNDr. Jan Kalina, Ph.D.
KPMS MFF UK & Ústav informatiky AV ČR, v.v.i.
22. května 2023