Oponentský posudek na disertační práci:

Přemysl Bejda: Median in some statistical methods

Předložená práce se zabývá použitím mediánu ve vybraných statistických úlohách: ve druhé a třetí kapitole se autor postupně věnuje robustním modifikacím exponenciálního vyhlazování a odhadům polohy pro mnohorozměrná data. Práce je psaná anglicky, na mnoha místech je však patrné, že autor s psaním anglického textu dosud nemá potřebné zkušenosti.

Poznámky ke kapitole 2

Zjednodušeně řečeno, celá druhá kapitola je založena na modelu $y_t = z_t^\top a_t + \varepsilon_t$. V sekci 2.1 se parametry a_t odhadují pomocí vážených regresních kvantilů a v sekci 2.2 pomocí neváženého postupu založeného na klouzavých mediánech a znaménkovém testu. Vlastním přínosem je návrh heuristických algoritmů a odvození některých základních vlastností (např. ekvivariance a breakdown point).

Použití algoritmu navrženého v části 2.1 se mi nezdá být dostatečně dobře zdůvodněné. Osobně bych v tomto případě nejdříve zkusil použít existující funkce pro odhad regresních kvantilů (např. funkce `rq` v knihovně `quantreg`) nebo metodu IRLS (Iteratively Reweighted Least Squares). Zajímavé by pak mohlo být srovnání několika 'přímočarých' algoritmů s postupem navrženým v disertační práci. Zajímalo by mě také, jestli jsou odhady zkoumané v sekci 2.1 srovnatelné například s metodami 'conditional quantile smoothing' viz např. [Abberger, K. (1997). Quantile smoothing in financial time series. Statistical Papers, 38(2), 125-148] nebo CAViaR [Engle, R. F., Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics, 22(4), 367-381].

V sekci 2.2 se používá jiný přístup, který počítá s přítomností skoků. Navržený heuristický algoritmus je nutné vyladit pomocí parametrů, které se volí na základě zkušenosti. Případy $z_t = 1$, obecné z_t a $z_t = (1, t)^\top$ jsou postupně prodiskutované v sekcích 2.1.1 až 2.1.3. Kromě algoritmů je opět odvozena ekvivariance vůči posunutí a breakdown point.

Simulační studie v sekci 2.3 pak ukazuje, že navržené algoritmy v praxi fungují lépe než dva jiné algoritmy a lze je doporučit pro kontaminovaná data a v přítomnosti velkých skoků (algoritmus S). V části 2.4 jsou uvedené algoritmy použité na HDP v Číně.

Poznámky ke kapitole 3

Druhá část předložené práce se týká robustních odhadů polohy a zabývá se především metodami založenými na tzv. geometrickém mediánu. V sekci 3.1 je velice obecně definován nový odhad (u kterého mi místy nebylo jasné, co se vlastně odhaduje). V sekci 3.2

se autor zabývá konzistencí, sekce 3.3 obsahuje simulační studii, popis možného použití navrženého odhadu ke konstrukci modifikovaného boxplotu a bagplotu a návrh nové míry rizika. Sekce 3.4 se zabývá volbou ladicího parametru b za předpokladu normality.

Zajímavá a přínosná se zdá být zejména konstrukce modifikovaného boxplotu a bagplotu: zde je však také dobře vidět, že navržená metoda bude fungovat nejlépe na symetrická rozdělení. Pokud bude zkoumané rozdělení nesymetrické, tak navržený postup symetrii může zamaskovat, protože do 'boxu' nebo 'bagu' zahrnuje pozorování symetricky, pouze na základě jejich vzdálenosti od geometrického mediánu (pro rostoucí počet pozorování by měl tvar 'bagu' konvergovat ke kouli). Tato vlastnost zároveň souvisí s tím, že navržený postup zřejmě není ekvivariantní vůči změně měřítka jedné souřadnice (na rozdíl od podobných postupů založených na poloprostorové nebo simplexové hloubce).

Další připomínky

str. 5 Domnívám se, že Talmud bude ještě o něco starší.

str. 6, odst. 5 Domnívám se, že metody založené na zobecněních mediánu byly navržené i pro obecnější časové řady, než AR(1), viz např. [Franke J, Härdle W, Martin RD, editors Robust and nonlinear time series analysis: Lecture Notes in Statistics, Vol. 26. New York: Springer Verlag; 1984].

s. 8, ř. 8 Binomický koeficient $\binom{0}{0}$ se obvykle definuje jako 1.

s. 8, ř. −1 Co přesně zde znamená symbol $\liminf$? Je to totéž jako v Definici 1.1.6 na následující straně?

s. 9, ř. −4 Co znamená obrat 'estimator of observations'?

s. 10, ř. 8 Co označuje symbol $\overline{T}$ a v jakém smyslu konverguje T_n k $\overline{T}$?

s. 10, ř. −3 Jak se definuje vážený medián?

s. 11, def. 1.1.10 Nemělo by $\pi_t(T_n)$ záviset na p?

s. 18, prop. 2.1.3 Čemu se zde rovná β? Šlo by větu přeformulovat pro $\beta \to 1$?

s. 62, (3.6) Co přesně zde označuje n_0? Nemůže mít $\mathbf{y}$ víc prvků?

s. 64, (3.8) K čemu se vztahuje střední hodnota? Prosím také o kopii citovaného článku Kemperman (1987), který se mi nepodařilo nalézt v dostupných elektronických zdrojích. Možná není zcela vhodné označovat odhadovaný parametr stříškou.

s. 66, ř. −2 Proč nemůže být $\Phi = 1/2$ (pokud by byla hustota nulová v intervalu $(a, a+\varepsilon)$)? Funguje pak důkaz Proposition 3.2.2?

s. 70 Lze na modifikovaném boxplotu nějak poznat nesymetrii?

s. 74, def. 3.3.3 Platí, že za předpokladu normality $a(L) = q_{N(0,1)}(0.75)$? Neměl by se pak modifikovat jmenovatel ve vzorci pro $\beta(L)$?

s. 78, druhý nečíslovaný vzorec Výsledek lze zjednodušit na $b = q(1 - \alpha/2)$.

s. 78, třetí nečíslovaný vzorec Není zde nějaký překlep?

Shrnutí

Hlavním přínosem druhé kapitoly jsou výpočetní algoritmy, které však často nejsou zasazeny do kontextu podobných metod, které lze nalézt v aktuální odborné literatuře. Odvození jsou podle mého názoru korektní, teoretické výsledky se přitom týkají převážně ekvivariance a breakdown pointu navržených odhadů. Zajímavější jsou odhady vyšetřované ve třetí kapitole a jejich použití při konstrukci modifikovaného boxplotu a bagplotu.

V závěru práce mohl být uveden seznam relevantních autorových publikací: takto není zřejmé, které části disertace už jsou publikované v odborné literatuře.

Závěr: Předložená práce prokazuje předpoklady autora k samostatné tvořivé práci a doporučuji ji uznat jako práci disertační za předpokladu, že autor odpoví na všechny výše uvedené připomínky a při obhajobě předloží dostatečně bohatý seznam kvalitních publikací.

Doc. RNDr. Zdeněk Hlávka, Ph.D.