Jan Ámos Víšek

CHARLES UNIVERSITY, FACULTY OF SOCIAL SCIENCES
INSTITUTE OF ECONOMIC STUDIES,
DEPARTMENT OF MACROECONOMICS AND ECONOMETRICS

Mailing address : Opletalova 26, 110 01 Prague 1, CZECH REPUBLIC
Phone : +420–732–344–585 *Fax* : 42–2–2211-2304 *e-mail* : visek@fsv.cuni.cz

Oponentský posudek dizertace

Mgr. Přemysla Bejdy

MEDIAN IN SOME STATISTICAL METHODS

První kapitola - *Introduction* - je věnována (z velké části) shrnutí toho, co bylo dosud uděláno v oblasti takových robustních metod, kdy je důraz kladen na jednoduchost algoritmu. Zvláštní pozornost je věnována robustním metodám pro časové řady[1]. Současně je naznačeno (a to dost podrobně) co bude cílem následujících dvou kapitol. Odstavec *1.1 Basic notation* je pak věnován zavedení formalizmu a připomenutí některých definic. Oboje - formalizmus i definice (které jsou převážně zobecněním definic pojmů z oblasti "klasické" robustní statistiky) - jsou konstruovány tak, aby pokryly zejména oblast časových řad. K tomu, jak se toto povedlo se ještě vrátíme v části posudku, který naznačí možná vylepšení textu. Ještě však dvě malé poznámky k textu úvodu.

První se týka textu na straně 2 dole. Zde se praví, že k tomu, aby odhad metodou nejmenších čtverců byl nejlepším nestranným odhadem mezi všemi lineárními nestrannými odhady nepotřebujeme normalitu disturbancí. To je sice pravda, ale být nejlepším mezi všemi lineárními nestrannými odhady je jako být jednookým mezi slepými králem. Ve hře je totiž mnoho nelineárních odhadů, které mohou být (a typicky jsou) lepší než lineární, pokud nemáme normalitu disturbancí. Občas se lze v některých (dle mého názoru - špatných) učebnicích dočíst, že oprávnění omezit se na lineární odhady je dáno tím, že studujeme lineární model[2]. Neboť název *"lineární model"* poukazuje pouze na to, že vysvětlovaná veličina je vysvětlena lineární kombinací nějakých vysvětlujících veličin, které ale mohou být velmi různorodými transformacemi (mocninami, logaritmy atd.) původních dat. Jinými slovy, (abychom dali jeden příklad) Stone-Weierstrassova věta (např. Branges, L. De (1959)) zaručuje, že tímto *"lineárním modelem"* můžeme dobře aproximovat téměř jakoukoliv závislost "response variable" na "explanatory variables", byť může být hrubě nelineární. To znamená, že linearita modelu a linearita odhadu nemají spolu nic co do činění. Pak ovšem bychom měli

[1] Toto je možná vhodné místo k tomu, abych čtenáře posudku upozornil na to, že nejsem expert na časové řady a dizertace mi byla svěřena k oponování díky tomu, že jejím podstatným přínosem je robustifikace metod určených na zpracování časových řad. To samozřejmě vede k tomu, že místy jsou moje připomínky naivní, neboť si nejsem vědom tradicí ustálených úzusů a tudíž automaticky přijímaných předpokládů a označení.

[2] Osobně si myslím, že je to brilantní ukázka demagogie.

hledat takový model, ve kterém budou disturbance (aproximované residui) alespoň přibližně normálně rozdělené, abychom zaručili, že odhad metodou nejmenších čtverců je nejlepší mezi všemi nestrannými odhady[3].

Druhá poznámka směřuje k textu na straně 6[18]. Zde je vyslovena myšlenka založení odhadu na datech, která - pomoci metody s vysokým bodem selhání - označíme jako *"nekontaminovaná"* (v dizertaci se pro ně používá označení *"unperturbated"*). Jinými slovy, tato myšlenka říká: *"Pokud totiž - byť za cenu (velké) ztráty eficience - nalezneme "skutečný underlying" model, můžeme celou odhadovací metodu následně zeficientnit použitím např. jednokrokového odhadu"*. To je myšlenka, která se asi poprvé objevila v článku Petera Bickela (1975). Jednalo se tedy o to, zda bychom mohli nalézt odhad regresních koeficientů, který by měl 50% bod selhání (podobně jako medián v úloze odhadu parametru polohy) a ten pak použít jako vodítko k nalezení - pomocí už podstatně eficientnějšího odhadu - "skutečného underlying" modelu. Tuto myšlenku (patrně poprvé) zpochybnil článek Thomase Hettmanspergera a Simona Sheathera (1992), který ukázal, že malá změna jedné souřadnice jednoho pozorování může vést k velké změněn odhadu metodou *"the least median of squares"*[4]. Snadno se navíc najde (akademický) příklad dat, kde aplikace metody *"the least median of squares"* a metody *"the least trimmed squares"* dá dva ortogonální regresní modely, ač oba odhady (tak jak jsou v takovém případě nastaveny jejich parametry, viz Víšek (2000)) mají 50% bod selhání. To znamená, že postup založený na myšlence "očistění" dat v prvním kroce od kontaminace a následném "doladění" odhadu může zklamat.

Podstatné výsledky dizertace jsou obsaženy v druhé a třetí kapitole. Obě kapitoly (a ostatně i část kapitoly první) jsou napsány poněkud netradičním stylem, který kombinuje (více meně) klasický styl matematického textu s uváděním algoritmů (u kterých by ale čtenář, který sám nic v životě neprogramoval, byl bez vysvětlení - které tam není - ztracen, nebo by se jen dohadoval, co který prvek vlastně znamená; nejspíše by se dohadoval správně, ale to neznamená, že by tam to vysvětlení nemělo být). Navíc snadnějšímu porozumnění algoritmům by pomohl nějaký heuristický komentář či jejich uvedení v blokovém schematu. Takto se čtenář (přinejmenším u složitějších algoritmů) musí domýšlet myšleku, která stojí v jejich pozadí. Neuškodilo by rovněž pár slov o tom, proč byla zvolena tato netradiční forma výkladu, tj. nějaký komentář vysvětlující, proč jsou odhady neznámých parametrů definovány pomocí algoritmů a nikoliv jako řešení extremálních problémů (tak jak jsme tomu zvyklí z klasické i robustní statistiky). Definování odhadů pomocí algoritmů totiž pro čtenáře zvyklého na definice pomocí extremálních problémů, značně ztěžuje porozumění textu, neboť definice pomocí extremálního problému téměř okamžitě napoví, na jaké myšlence (heuristice) je odhad

[3]To samozřejmě také záleží na filozofii, se kterou přistupujeme ke zpracování dat - zda zastáváme spíše pozici *kritického realizmu* (který více méně nepřipouští nějaké manipulace s daty, např. jejich transformace a považuje hledaný regresní model daný od Pana Boha) nebo je nám bližší *instrumentalizmus* a pak připouštíme, že tvar vysvětlujícího modelu je zcela na nás, na naší invenci a zručnosti.

[4]Výsledky uvedené v článku byly sice špatně - díky špatnému algoritmu, který použili, viz Boček & Lachout (1993) a případně Víšek (1994), ale upozornili na tento problém.

založen a tak si jej člověk snadno zapamatuje. Navíc, místy dost možná může tato forma definování odhadů vést k přehlédnutí nekorektnosti zavedení toho či onoho odhadu. Trochu mi to připomíná situaci, kdy na přelomu 70. a 80. let, kdy se počítače (byť stále ještě sálové) staly přátelštější, někteří matematici začali experimentovat s matematickými metodami připravenými ad hoc pro tu či onu situaci (čeština má pro tuto činnost přiléhavé slovo *bastlit*). Paul Halmos (autor excelentní knihy *Measure Theory*) na to reaguje článkem *Applied mathematics is a bad mathematics*, ve kterém vysvětluje, že dobrá matematika by se měla přidržet takové formy, která byla založena pracemi Thaléta z Milétu, Pythagora ze Samu, Euklida z Alexandrie či Archiméda ze Syrakus a na níž navázali (a podstatně obohatili nejen co do výsledků, ale tako co do pojetí) Bernard Bolzáno, Georg Cantor, David Hilbert, Kurt Gődel. Pro statistiku a teorii pravděpodobnosti pak Pierre-Simon Laplace, Carl Friedrich Gauss, Adrien-Marie Legendre, Aylmer Fisher, Andrej Nikolajevič Kolmogorov a u nás Jaroslav Hájek (a mnozí další).

Pokud se chceme tuto tradici oputit měli bychom pro to mít silné důvody a ty bychom měly v úvodu k takovému pokusu dobře vysvětlit.

Druhá kapitola je věnována *Rekursivním adaptivním metodám*. Jsou zde diskutovány dvě verze robustifikace exponenciálního vyrovnávání:

- využití regresních kvantilů v exponenciálním vyrovnávání

a

- exponenciální vyrovnávání kombinované s klasickým znaménkovým testem.

V odstavci *2.1 Exponential smoothing based on regression quantiles* jsou v klasickém exponenciálním vyrovnávání prvního řádu čtverce residuí nahrazeny hodnotami, které získáme dosazením residuí do funkce ρ_α, zavedené Rogerem Koenkerem a Gilbertem Bassettem (1978). (Domnívám se, že příslušné robustní vyrovnání mohlo být definováno jako argument, který minimalizuje výraz v (2.3) a hodnota (či hodnoty) nalezená pomocí Algoritmu 2.1.1 by pak byla aproximací tohoto odhadu.) Následuje diskuze o algoritmu, která je rozdělena na případ jednorozměrné a vícerozměrné regrese.

Pro případ jednorozměrné regrese je nejprve popsán algoritmus a poté jsou dokazovány vlastnosti odhadu definovaného jako výsledek tohoto algoritmu - ekvivariance, případně škálová ekvivariance. Mimochodem, vzhledem k tomu, že text - zejména v prvé kapitole - zavádí celou řadu pojmů pomocí definic, je trochu nedůsledné zavést jeden z (hlavních) výsledků dizertace, totiž *kvantilový odhad* $\tilde{a}_t^\alpha$, viz (2.6), jen konstatováním na str.16_{15}: *We construct the quantile estimate in this way.*). Výklad je doplněn příkladem a důkazem *inkonzistence* takto získaného odhadu regresního koeficientu - srovnej s Hawkins *&* Olive (2003). Již tyto úvahy jsou poměrně komplikované a proto je nepochybně dobře, že autor uvažuje nejprve tento "jednoduchý" případ.

Lemma 2.1.4 uvedená v závěru této pasáže stanovuje velikost bodu selhání. Validitu jejího

důkazu však lze ověřit - dle mého názoru - jen intuitivně, neboť by bylo nejprve třeba upřesnit znění *Definice 1.1.7*, která se odvolává na *Definici 1.1.6*. Ta je možná v pořádku, ale pokud ano, neodpovídá běžně používané definici v robustní statistice (a proto by to stálo za nějaký komentář, nejlépe bezprostředně za ní).

Následuje výše zmíněný výklad pro obecný případ, tj. kdy stále ještě vyrovnáváme časovou řadu exponenciálním vyrovnáváním prvého řádu, ale předpokládáme, že časová řada byla generována vícerozměrnou regresí. Algoritmus výpočtu odhadu je založen na stejné myšlence jako algoritmus pro jednorozměrnou regresi, jen jeho realizace je podstatně složitější a vyžaduje si dost komplikované geometrické (viz str. 20_2) a kombinatorické (viz tabulka B_t) úvahy.

Odstavec 2.1 je zakončen návrhem α-winsorizovaného odhadu a jeho algoritmu. Odkazem na Portnoy *&* Koenker[5] (1997) je vlastně připomenuto to, že návrh regresních kvantilů byl dobře přijat statistickou a zejména ekonometrickou komunitou, bylo dáno tím, že prakticky současně s jejich zavedením objevila cesta, jak je spolehlivě spočítat pomocí metod lineárního programování.

Druhá kapitola pokračuje návrhem robustního vyrovnávání časové řady, které je založeno na znaménkovém testu. Robustifikace spočívá v použití váženého mediánu. Tato část dizertace místy přechází do přehledového textu. Taková shrnutí by mělo být spíše v úvodu dizertace než v částech, kde se rozvíjí vlastní teorie (mimochodem odkaz na Pollard (2012) pro případ i.i.d., viz 31_{17}, je trochu chozením s kanónem na vrabce, neboť odkaz na Anděl (1993) by splnil stejný účel - citace Pollarda je možná oprávněná tím, že některé excelentní knížky - Anděl (1993), Cipra (1986) či Štepán (1987) - jsou bohužel napsány česky a nedá se na ně tak úplně dobře odkazovat).

Úvahy o použití znaménkového testu jsou rozděleny do několika odstavců. Nejprve je výklad zaměřen na časové řady s konstantním trendem a to (podobně jako výše) pro situaci, kdy je časová řada generována jen konstantním členem (interceptem). Zobecnění, tj. situace, kdy je řada s konstantním trendem generována vícerozměrnou regresí pak velmi stručně zobecní zde nalezené výsledky (či spíše nanačí, jak je zobecnit). Následují výsledky dosažené pro časovou řadu s linearním trendem. Na konci odstavce *2.2.* je pak navržena modifikace algoritmu (studovaného v předchozím odstavci), která se opírá o myšlenku Andrewa Siegela (1988). Odhad navržený Andrewem Siegelem *the repeated median* je výpočetně tak složitý (složitost je $\mathcal{O}(n^p)$, kde p je dimenze regresního modelu), že jej lze implementovat (pokud vím) jen pro jednoduchou regresi[6]. To je však právě situace studovaná v odstavci 2.2.3 a 2.2.4, kde vystačíme s jednoduchou regresí.

[5]Nevím, zda příslušná metoda a její implementace byla někdy publikována, ale vím, že jednou přijel (domnívám se, že ještě před sametovou revoluci) do Prahy Roger Koenker, aby se poučil o tom, jak navrhl Jaromír Antoch počítat regresní kvantily.

[6]Navíc má podobné, nepříliš dobré vlastnosti jako *the least median of squares*, tj. daleko menší eficiencí a pomalejší konvergencí než např. S-odhady, viz Rousseeuw *&* Yohai (1984) (návdavkem k tomu je fakt, že asymptotické rozdělení není *normální*). Konec konců v dizertaci je to potvrzeno např. na str. 61^8.

Z textu druhé kapitoly je patrné, že autor dizertace se seznámil s mnoha výsledky jiných autorů a na řadě míst v dizertaci navrhl jejich využití při zpracování dat algoritmy jím navrženými.

Rád bych ještě zdůraznil jednu věc. Již v době, kdy se objevila práce Hawkins & Olive (2003), začala diskuze o tom, zda je vůbec smysluplné zamýšlet se nad konzistencí "odhadů" vypočtených algoritmy popsanými např. v Rousseeuw & Leroy (1986), Boček & Lachout (1993), Hawkins (1994), Hawkins & Olive (1999), Víšek (1994), (2000), Klouda (2007), (2015) či zda vlastně takové diskuze jsou implicitně založené na nepochopení toho, co dané algoritmy počítají. To, po čem volal Paul Halmos, totiž že naše úvahy mají splňovat odhady (či obecně metody) definované ve formě extremálních problémů (např. mají být konzistentní), zatímco navržené algoritmy vyčíslují (pokud možno těsné) aproximace řešení těchto extremálních problémů. Jejich vlastnosti je pak třeba zmapovat pomocí dobře koncipovaných numerických studií, viz např. Marazzi (1992) či Klouda (2009). Proto je třeba ocenit, že většina výsledků dizertace se věnuje vlastnostem algoritmů pro konečné výběry dat, zejména jejich ekvivarianci či invarianci. Zcela v souladu s myšlenkami v Marazzi (1992) druhá kapitola v závěru nabízí dvě numerické studie. Prvá je založena na simulovaných datech a opět je nejprve uvažován konstantní trend a pak lineární trend. Celá kapitola je uzavřena aplikací znaménkového testu a modelu s lineárním trendem na data o ročních výsledcích čínské ekonomiky - representovaných *hrubým domácím produktem* - v letech 1952 až 2014.

Třetí kapitola se věnuje studiu odhadů parametru polohy a vychází přitom z geometrického mediánu. Je zde odvozen jednoduchý nástroj na vyhodnocení velikosti bodu selhání takového odhadu, který je "složením" dvou zobrazení, přičemž se předpokládá, že u prvého zobrazení známe bod selhání, zatímco druhé umíme majorizovat vhodným "marginálním" zobrazením. Třetí kapitola představuje na jedné straně samostatný soubor výsledků, na druhé straně ji lze chápat jako kolekci výchozích odhadů, které lze použít v metodách studovaných v druhé kapitole. Musím přiznat, že jsem výklad ve třetí kapitole prošel jen z části, neboť místy byl použito označení, které nebylo nikde před tím (ani potom, alespň si to myslím) vysvětleno a jen jsem se mohl domnívat, co znamená (připomínky k textu, jsem poskytl přímo autorovi).

Dovolím si ještě jednu poznámku. Na začátku třetí kapitoly je malý přehled robustních odhadů - MVE, S-odhady, τ-odhady, atd. a je tam poznamenáno, že jejich výpočet bývá časově náročný. Vzhledem k tomu, že to je oblast, ve které je stále veden výzkum, lze tyto odhady počítat v rozumném čase a to i pro dosti rozsáhlá data. Např. výpočet S-odhadů lze spočítat (velice rychle) dle stejného vzorečku jako *the ordinary least squares*, tj. $(X'X)^{-1}XY$ (viz Cohen-Freue et al. (2013), Desborges & Verardi (2012)), kde příslušné kovarianční matice odhadneme pomocí S-odhadů (viz Campbell et al. (1998)). Rychlost výpočetní techniky se navíc zvyšuje (možná) rychleji než složitost našich výpočtů. Takže dokonce i iterativní algoritmus Hawkinsova typu použitý pro výpočet *"the least trimmed squares"* (LTS) pro

iterativně hledaný odhad gravitačního modelu pro cca 130 států (tj. designová matice měla téměř jak 17 000 řádků a LTS byl řádově desetkrát opakováně počítán, viz Egger & Víšek (2014)) trval jen několik minut.

Celá dizertace je pak zakončena stručným shrnutím dosažených výsledků a docela pěknou nabídkou možných dalších badatelských problémů.

Ačkoliv je to možná překvapivé, že po cca dvou či dokonce třech desetiletích od navržení robustních odhadů typu LTS či S-odhadů se stále bádá nad algoritmy, jak je spočítat - viz např. Desborges & Verardi (2012) - je tomu tak. To ilustruje to, že výsledky uvedené v dizertaci - návrhy algoritmů pro výpočet robustních odhadů - jsou aktuální. Jak plyne z některých připomínek, je na formální stránce dizertace co zlepšovat, než bude např. text použitelný pro publikování v nějakém časopise. To je však dáno tím, že se autor pustil do novátorské práce, tj. pokusil se o novou formu definování odhadů a proto se domnívám, že to lze zcela pochopit a omluvit. To, zda tento počin bude opravdu signifikantní, tj. přinášející závažnou inovaci ve statistice, ukáže až čas. Samozřejmě bych to autorovi přál. To že většina výsledků je nových je evidentní. Jsem přesvědčen, že autor dizertace prokázal, že je schopen samostatné vědecké práce a proto doporučuji přijmout dizertaci k obhajobě.

V průběhu obhajoby bych uvítal, kdyby autor vysvětlil:

- Proč se uvažování normovaného vektorového prostoru X lépe hodí pro studium časových řad než $\mathbb{R}^d$ (viz str. 7^3).
- Jakou výhodu - kromě v jistém smyslu větší obecnosti - mají množinové odhady neznámých parametrů. Vzhledem k tomu, že obecně se nejedná o intervalové odhady s předepsanou pravděpodobností pokrytí skutečné hodnoty těchto parametrů, není mi jasné, jak si uživatel nakonec vybere mezi obecně několika diskrétními veličinami, které mohou být (třeba) od sebe velmi vzdáleny.

Literatura

- Anděl, J. (1993): *Statistické metody*, MATFYZPRESS, Praha, 1993.
- Bickel, P. J. (1975): One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70, 428–433.*
- Boček, P., P. Lachout (1993): Linear programming approach to LMS-estimation. *Memorial volume of Comput. Statist. & Data Analysis 19(1995), 129 - 134.*
- Branges, L. De (1959): The Stone-Weierstrass Theorem. *Proceedings of the American Mathematical Society, Vol. 10, No. 5 (Oct., 1959), pp. 822-824 .*
- Campbell, N. A., Lopuhaa, H. P., Rousseeuw, P. J. (1998): On calculation of a robust S-estimator of a covariance matrix. *Statistics in medcine, 17, 2685 - 2695.*

- Cipra, T. (1986): *Analýza časových řad s aplikacemi v ekonomii.* SNTL/ALFA 1986.
- Cohen-Freue, G. V., Ortiz-Molina, H., Zamar, R. H. (2013): Natural robustification of the ordinary instrumental variables estimator. *Biometrics 69 641 - 650.*
- Desborges, R., Verardi, V. (2012): A robust instrumental-variable estimator. *The Stata Journal (2012) 12, 169 -181.*
- Egger, P. *&* Víšek, J. Á. (2014): Structural Least-trimmed Squares Estimation of Gravity Models. (bohužel stale ještě draft).
- Halmos, P. R. (1981): Applied mathematics is a bad mathematics. *Mathematics tomorrow*, Springer Verlag, New York, 9 - 20.
- Hawkins, D. M. (1994): The feasible solution algorithm for least trimmed squares regression. Computational Statistics and Data Analysis 17, 185 - 196.
- Hawkins, D. M., D. J. Olive (1999): Improved feasible solution algorithms for breakdown estimation. *Computational Statistics & Data Analysis, Volume 30, Number 1, 1 - 12.*
- Hawkins, D. M., D. J. Olive (2003): Inonsistency of resampling algorithm for high breakdown regression estimation and a new algorithm. *Journal of the American Statistical Association 97, 136-159.*
- Hettmansperger, T. P., S. J. Sheather (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*
- Huber, P. J. (1964): Robust estimation of a location parameter. *Ann. Math. Statist. 35, pp. 73–101.*
- Klouda, K. (2007): Algorithms for computing robust regression estimates. *Diploma thesis, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague.*
- Klouda, K. (2009): Comments to Random Solution Algorithm. *Preprint submitted to Elsevier Science.*
- Klouda, K. (2015): An exact polynomial time algorithm for computing the least trimmed squares estimate. *Computational Statistics & Data Analysis 84, 27-40.*
- Koenker, R., G. Bassett (1978): Regression quantiles. *Econometrica, 46, 33-50.*
- Marazzi, A. (1992): *Algorithms, Routines and S Functions for Robust Statistics.* Wadsworth *&* Brooks/Cole Publishing Company, Belmont, California, 1992.

- Pollard, D. (2012): *Convergence of Stochastic Processes.* Springer Sciences & Business Media.
- Portnoy, S., Koenker, R. (1997): The gaussian hare and laplacian tortoise: computability of squared-error versus absolute-error estimator. *Statistical Sciences 12(č), 279 - 300.*
- Rousseeuw, P. J., A. M. Leroy (1987): *Robust Regression and Outlier Detection.* New York: J.Wiley & Sons.
- Rousseeuw, P. J., V. Yohai (1984): Robust regression by means of S-estimators. In: *Robust and Nonlinear Time Series Analysis. eds. J. Franke, W. Hărdle and R. D. Martin, Lecture Notes in Statistics No. 26 Springer Verlag, New York, 256-272.*
- Siegel, A. F. (1982): Robust regression using repeated medians. *Biometrica, 69, 242 - 244.*
- Štepán, J. (1987): *Teorie pravděpodobnosti.* Academia, Praha 1987.
- Víšek, J. Á. (1994): A cautionary note on the method of Least Median of Squares reconsidered, *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Lachout, P., Víšek, J.Á. (eds), 1994, Academy of Sciences of the Czech Republic, Prague, 1994, pp. 254 - 259.*
- Víšek, J. Á. (2000): On the diversity of estimates. *Computational Statistics and Data Analysis 34, (2000) 67 - 89.*

Praha, 21. listopad, 2017