

РЕЦЕНЗИЯ

на дисертационен труд

„Robust Statistical Modeling Trough Trimming"

за присъждане на научната степен **„Доктор на науките"**,

Област 4. Природни науки, математика и информатика,

Професионално направление 4.5. Математика, Научна специалност

„Теория на вероятностите и Математическа статистика"

Автор: Доцент д-р **Нейко М. Неиков**, НИМХ-БАН

Рецензент: Професор д-р Николай М. Янев, ИМИ-БАН

1. Основни сведения.

Представеният дисертационен труд е на английски език и съдържа 200 страници текст, от които 15 стр. са библиография. Освен това е представен Автореферат на български език със заглавие **„Робастно статистическо моделиране чрез тримиране"**, което съответства напълно на английския текст. Авторефератът съдържа 94 стр. текст, от които 14 стр. са библиография. По-нататък ще се придържахме към терминологията, въведена в автореферата.

Авторът д-р Неико М. Неиков работи като доцент в Националния Институт по Метеорология и Хидрология към БАН.

Журито е назначено с заповед 99/26.05.2016 на директора на ИМИ-БАН на основание решение на НС на ИМИ-БАН (Протокол 5 от 20.05.2016). Предзащитата на дисертационния труд е проведена в ИМИ-БАН на 5.05.2016. Документацията и всички процедури по защитата са в съответствие с изисванията на закона и съответните правилници.

2. Структура и анализ на дисертационния труд и автореферата.

Дисертационният труд започва с общ увод относно мотивацията и кратък поглед върху робастната статистика. Дадени са основни дефиниции и кратки исторически бележки. Описана е подробно структурата и съдържанието на дисертацията, която съдържа 8 глави, основаващи се на 11 публикации на автора. Публикациите са съвместни с 5 различни автори - 2 от България (П.Нейчев и Р.Димова) и 3 от чужбина (Ch.Muller,

P.Filzmoser, P.Cizek). Водещата роля на Н.Нейков в тези публикации не подлежи на съмнение. Основната част от публикациите (8 на брой) са в сериозни международни списания и издания (с импакт фактор) като само три от тях са у нас (Доклади БАН и Плиска). Всички публикации са получили широко международно признание, както се вижда от приложения списък с цитирания — 214 на брой, подредени по съответните глави. Прави впечатление, че те са предимно в списания с импакт фактор, книги, а също и в дисертации в чужбина и у нас.

Авторефератът е доста подробен и правилно отразява основните дефиниции, модели и резултати на дисертацията. Освен това той е снабден с допълнителни коментари, които дават една по-добра представа и за самата дисертация.

3. Научни и научно-приложни приноси.

Най-общо казано, дисертацията разглежда въпроси за робастно оценяване при тримиране (т.е. орязване, отстраняване) на наблюдения, несъответстващи по някакъв начин на основния разглеждан модел. Тези идеи при високи прагови точки се илюстрират в дисертацията върху обобщени линейни модели, можествена линейна квантилна регресия, линейна експоненциална фамилия, екстремални разпределения и в някои задачи с големи размерности. Праговата точка всъщност съответства на максималния процент наблюдения, които могат да бъдат произволно заместени без сериозно влошаване на съответната оценка. Предложени са различни техники за характеризирание на прагови точки за някои класове оценки. За въведените в дисертацията робастни оценки е изследвано поведението им във връзка със съответните класически оценки. Освен строгите математически доказателства на съответните теореми, в дисертацията има интересни примери с графични представяния и симулации.

Основните резултати и приноси в отделните глави на дисертацията са добре представени в кратки резюмета (Summary) в началото на всяка глава. В автореферата тези резюмета (Summary) са наречени Въведения, а в края на всяка глава са предствени и допълнителни коментари, които са обозначени като Резюмета. Всички те дават една добра представа за получените резултати и приноси в сравнение с постигнатото от други автори или от дисертанта в негови предишни изследвания.

Глава 1 има фундаментален характер и в нея се обобщават предишни резултати на Върндев и Нейков (1998) относно праговата точка на така въведените S -оценки, които имат по-общ характер от преди това изследваните. Един от основните резултати е получен в Лема 1.2 и Теорема 1.1. За един подклас на линейната експоненциална фамилия в Теорема 1.3 е намерена интересна връзка между индекса на d -пълнота, въведен от Върндев (1993), и т.н. параметър $N(X)$, въведен от Muller (1997). По този начин са определени праговите точки за линеен логистичен регресионен модел (Теорема 1.4), лог-линеен регресионен модел (Теорема 1.5) и линеен модел с разпределение на Лаплас от ред q (което обхваща редица други известни разпределения) - Теорема 1.6. Като илюстрация на получените теоретични резултати в § 1.4 е разгледан един пример с реални данни, представени първоначално от канадски автори.

Глава 2 започва с въведение, в което се разглеждат т.н. WTL (Weighted Trimmed Likelihood - Претеглени Орязани Максимално-правдоподобни) оценки, а в §2.2 е представен т.н. FAST-TLE алгоритъм за приближеното им пресмятане в случай на обобщени линейни модели. Използвана е удачно идеята за повторни извадки (resampling techniques). В §2.3 са разгледани подробно три примера, които добре илюстрират алгоритъма като резултатите са представени в таблици и фигури.

Глава 3 е посветена на въведената от Върндев (1993) теория, основаваща се на понятието d -пълнота на множества от функции. Във въведението е добре представен исторически този проблем, а в §3.2 - неговото сегашно обобщение. Доказателствата на основните резултати Lemma 3.1, Proposition 3.1 и Proposition 3.2 са дадени в 3.4. Appendix. При доказателството на Lemma 3.1 остава известно съмнение, тъй като при граничен преход може само да се твърди, че $g_{\ast}$ не надминава C , а строгото неравенство трябва да се обоснове. В §3.3 получените резултати се прилагат в случай на обобщен логистичен регресионен модел и в Следствие 3.2 е намерена праговата точка.

В Глава 4 се разглеждат отново т.н. TLE (тримирани максимално-правдоподобни оценки), но сега в случай на GEV (обобщено разпределение на екстремалните стойности). Основната цел е изследване поведението на МПО и TLE(K) оценките по данни със и без замърсяване чрез Монте Карло симулационен експеримент. Резултатите са илюстрирани чрез 5 фигури и са коментирани в §4.4.

В Глава 5 се изследват пак TLE(K) оценки, но сега в случай на крайна смес от вероятностни разпределения. За оценяване на неизвестните параметри се развива идеята на ЕМ алгоритъм, а също и на класификационен ЕМ алгоритъм. Разгледан е въпросът за праговата точка при някои допълнителни условия. В §5.3 са разгледани подробно три интересни примера с използване на FAST-TLE алгоритъм. Тези резултатите са добре представени графично и таблично. Заслужава специално внимание фактът, че резултатите от тази глава са широко цитирани и използвани при различни приложения.

В Глава 6 се разглеждат линейни модели с разпределения от експоненциалната дисперсионна фамилия, общата форма на която е представена в автореферата. Целта е да се оценят неизвестните параметри, свързани с математическото очакване и дисперсията. Основната идея се състои в използване на функцията EQL (Extended Quasi-Likelihood - модифицирана функция на квази-правдоподобие). В §6.2 се разглеждат т.н. тримиранни квази-правдоподобни оценки (ETQL), за които в Теорема 6.1 е намерена праговата точка. В §6.3 е предложена процедура за апроксимация на оценките wGTE (претеглени обобщени тримиранни). В §6.4 е разгледан един реален експеримент като данните са взети от цитирания там сайт. Резултатите са представени в три фигури. В §6.5 се използват симулационни методи за изследване поведението на оценките EQL и ETQL. Подробно са изследвани три различни частни модели като резултатите са дискутирани и представени чрез Фигури 6.4 - 6.15. Показано е предимството на ETQL оценките.

В Глава 7 се разглежда случай на линейна квантилна регресия. Проблемът в исторически план е добре представен в §7.1. Introduction. Цел на тази глава е въвеждане и изследване т.н. LTQR оценки (Least Trimmed Quantile Regression). Както е отбелязано, тези оценки могат да бъдат намерени като частен случай на оценките GTE (Generalized Trimmed Estimator), предложени от Върндов и Нейков (1998). Това дава възможност в Теорема 7.1 да бъде намерена съответната прагова точка. В Теорема 7.2 е доказана състоятелност. Доказателството на този безспорно интересен резултат е трудно проверяемо, тъй като авторът използва доста факти от чужди автори, проследяването на които никак не е лесно. Освен това в §7.5 на автореферата се посочва и асимптотична нормалност, каквато в дисертацията не се вижда. В §7.5.1 е разгледан един реален експеримент

(звезден клъстър) и резултатите са представени във Фигура 7.1. В §7.5.2 са проведени Монте Карло симулационни експерименти за изследване и сравнение на LTQR и QR оценките. Резултатите са представени чрез Фигури 7.2 - 7.9. В §7.5.3 LTQR оценките са сравнени с робастни оценки на други автори и резултатите са представени в Таблица 7.1. Отбелязани са предимствата на LTQR оценките.

Последната Глава 8 разглежда някои задачи с висока размерност, когато броят на промеливите p е съизмерим или по-голям от броя на наблюденията n . А при задачи с ултра висока размерност (при $p \gg n$) се предполага, че например $\log(p) = O(n^a)$, $0 < a < 1$. В §8.1 се въвеждат някои форми на пенализация, при които моделите съществено се опростяват (най-общо казано). В §8.2 се въвеждат т.н. тримирани правдоподобни оценки с пенализация - PMTLE (Penalized Maximum Trimmed Likelihood Estimator). В §8.3 при ултра високи размерности се въвеждат и изследват т.н. процедури SIS (Sure Independence Screening) и ISIS (Iterative SIS) в случай на тримиране. Тези резултати се прилагат в §8.4 върху симулирани данни при множествена регресия и Поасонова (лог-нормална) линейна регресия. Резултатите са представени в Таблицы 8.1 - 8.7 и са подробно коментирани.

Като един общ извод, може да се отбележи, че анализът на научните и научно-приложните приноси в дисертацията показва, че те покриват и дори надхвърлят посочените в автореферата „Предмет и цели на дисертационния труд" (§0.5, стр. 13).

4. Бележки и препоръки.

Освен посочените по-горе забележки, биха могли да се посочат още някои очевидно поправими грешки и недостатъци, които обаче не затрудняват сериозно читателя. Дисертацията има определено монографичен характер и изложението е ясно, но доста стегнато, на места дори прекалено. ДТ изобилства с многобройни съкращения, които затрудняват читателя. Един индексен списък би било голямо улеснение.

Имам непосредствени отлични впечатления от докладите на дисертанта пред Националния семинар по Стохастика.

5. ЗАКЛЮЧЕНИЕ

Представеният дисертационен труд има всички качества на сериозна международна монография в областта на статистиката. Получените интересни резултати определят ново направление в областта на робастното оценяване. Една част от приносите имат определено фундаментален статистически характер, а друга част дават нови методи и алгоритми за оптимално оценяване при тримиране. Всички тези резултати са публикувани в престижни издания в последните 13 години, но имат вече много сериозен международен отзвук (над 214 цитирания). Освен това те са докладвани на редица престижни международни конференции и симпозиуми, като една част от докладите са по специална покана. Всичко това дава пълно основание да заключим, че са налице всички условия от ЗРАСРБ и неговия правилник, както и този на ИМИ-БАН, които се прилагат за получаване на научната степен „Доктор на науките“ по съответната специалност. Ето защо определено предлагам на почитаемото жури да оцени подобаващо високо представения дисертационен труд, а на неговия автор доцент д-р Нейко М. Нейков да бъде присъдена научната степен „ДОКТОР НА НАУКИТЕ“ в Област 4. Природни науки, математика и информатика, Професионално направление 4.5. Математика, Научна специалност „Теория на вероятностите и Математическа статистика“.

15.08.2016

Резензент:

/професор дмн Николай М. Янев/