

РЕЦЕНЗИЯ

от проф. д-мн Евгения Стоименова

на дисертационен труд **„Robust Statistical Modelling Through Trimming“** с
автор доц. д-р Нейко Матеев Нейков, НИМХ - БАН

за получаване на научна степен „доктор на математическите науки“ в
професионално направление: 4.5 Математика, научна специалност: „Теория на
вероятностите и математическа статистика“

Представеният дисертационен труд е на английски език. Той е в обем от 200 страници и включва увод, 8 глави и литература от 175 източници. Дисертацията се основава на 11 научни статии на автора, публикувани след 2003 година, които не са използвани в други процедури.

В дисертацията са представени изследванията на автора върху свойствата на статистически оценки с висока прагова точка, основани на тримиране. Това са оценки устойчиви на определен вид нарушения на условията на избрания статистически модел. Нестрого казано прагова точка на една оценка на параметър е максималното допустимо „замърсяване“ в данните при което статистическите оценки са робастни. Естествено е да се очаква, че прилагането на стандартни статистически методи, когато условията на модела са нарушени може да доведе до неточни или погрешни статистически изводи. Като цяло методите на робастната статистика за моделиране и анализ на данни са насочени към построяване на устойчиви оценки на параметрите и свързаните с тях критерии и доверителни интервали не само когато данните се подчиняват на предполагаемото разпределение, но и когато това е изпълнено само частично. Областта е актуална и бурно развиваща се през последните четири десетилетия.

В увода на дисертацията са въведени основните понятия: прагова точка, тримирани регресионни оценки, тримирани максимално правдоподобни оценки и някои основни предхождащи изследването теореми. Праговата точка е максималната част от наблюдения на извадката, които могат да бъдат заместени с произволни стойности така, че абсолютното отклонение на оценката от оценявания параметър да не може да приема произволно големи стойности. Тримирането се отнася за отстраняване (тримиране) на част от наблюденията с най-големи и най-малки стойности. Ненаправен е обзор на получените в дисертацията резултати по глави.

В Първа глава се изследват прагови точки на тримирани правдоподобни оценки и свързаните с тях оценки в обобщени линейни модели. Тук е дефиниран класът $\mathcal{S}$ от статистически оценки, основани на минориране и мажориране от $f_{\mathcal{S}}$ -тия елемент на d -пълно множество от функции. Долната граница на една прагова точка на тримирани правдоподобни оценки най-общо се представя чрез индекса на d -пълнота, въведен

от Д. Въндев (Statist. Probab. Lett., 1993), а по-късно развит от Въндев и Нейков (Statistics, 1998). Понятието d -пълнота е често използвано в тази дисертация и ще го приведем за яснота на това изложение. *Едно крайно множество от функции се нарича d -пълно, ако за всяко негово подмножество с размер d супремума на това множество е субкомпактна функция.*

В Първа глава са получени резултати за праговата стойност на обобщените тримираны оценки $wGTE$, дефинирани за d -пълно множество от функции. Основен резултат в тази глава е Теорема 1.1, определяща долна граница за праговата точка на S оценките. Резултатът е обобщение на резултат от Vandev и Neykov (1998). Намерена е по-ниска долна граница за праговата точка на $wGTE(k)$ оценките. Направено е съответствие между индекса на d -пълнота и параметъра $N(X)$, дефиниран от Muller (1997) и дефиниран чрез максималния брой повторни наблюдения в пространството на предикторите при множествен линеен регресионен модел. Резултатите са приложени за класическия множествен линеен регресионен модел с нормално разпределение на грешката, линейната логистична регресия, лог-линейната (Поасонова) регресия, линеен регресионен модел с разпределение на Лаплас от ред q с дисперсионен параметър. Като следствие от това са намерени горни граници на праговите точки на претеглените тримираны правдоподобни оценки $WTLE(k)$ за тези вероятностни модели. С техниката на d -пълнота е определена горна граница праговата точка на линейните регресионни S_c оценки на Rousseeuw and Leroy (1987). Резултатите от тази глава са публикувани в [1].

Във Втора глава е предложен и изследван алгоритъм FAST-TLE за приближено пресмятане на тримираны правдоподобни оценки $TLE(k)$ оценките за обобщени линейни модели. Доказана е локална сходимост на алгоритъма (Proposition 2.1). За да съществува решение на оптимизационната задача, се предполага, че множеството от функции е d -пълно. В случая на линеен регресионен модел с нормално разпределение на грешката алгоритъмът е еквивалентен на FAST-LTS алгоритъма, предложен от Rousseeuw and Van Driessen (1999). Представени са три примера за приложение на метода за линейната логистична регресия и Поасонова регресия, където са илюстрирани свойството на TLE оценките. Резултатите от тази глава са публикувани в [2].

В Трета глава е направено обобщение на техниката на d -пълнота, предложена от Vandev (1993) и разширена от Muller and Neykov (2003) за изучаване на праговата точка на обобщените тримираны оценки $wGTE(k)$ в ситуации на сложни вероятностни модели. Основните резултати са Proposition 3.1 и Proposition 3.2, които дават необходими условия за съществуване на решение на съответните оптимизационни задачи и намират долна граница на праговата точка на $wGTE(k)$. Твърденията са доказани при условия за множеството от функции, които са по-слаби от изискването за субкомпактност. Определен е индексът на d -пълнота чрез параметъра $N(X)$ и е намерена праговата точка на $WTLE(k)$ оценките за линейната логистична регресия с обобщена свързваща функция (Proposition 3.3). Резултатът е обобщение на предишни резултати

на Vandev and Neykov (1998) и Müller and Neykov (2003). Резултатите от тази глава са публикувани в статиите [3], [9] и [10].

В Четвърта глава се разглежда оценяване на параметрите на регресионни модели със зависима променлива, която има разпределение от класа на разпределенията на екстремалните стойности (GEV). Основният резултат е намирането на праговата точка на $TLE(k)$ оценките за разпределението на Гумбел, което е от класа GEV (Proposition 4.1, доказана в приложението). Приложени са Монте-Карло методи за изследване на максимално правдоподобните оценки (МПО) и $TLE(k)$ оценките с различни проценти на тримиране. Използвани са данни с и без наличие на несъгласувани наблюдения в зависимата променлива и в предикторни променливи. Чрез симулации е установено, че МПО са ненадеждни при наличие на несъгласувани наблюдения в данните, докато $TLE(k)$ оценките са подходящи при процент на тримиране по-висок от процента на замърсяване на данните с несъгласувани наблюдения. Резултатите от тази глава са публикувани в [4].

В Пета глава е предложен подход, основан на тримираните правдоподобни оценки за робастно оценяване на параметрите на крайна смес от вероятностни разпределения. В основното твърдение в тази глава, Proposition 5.1, е определена долна граница за праговата точка на $WTLE(k)$ оценките за крайна смес от разпределения с техниката на d-пълнота. Доказателството е включено в Приложението. Предложен е адаптивен подход за определяне на параметъра на тримиране k , основан на тримирания информационен критерий на Бейс. За приближено пресмятане на неизвестните параметри на смесите от разпределения е създаден адаптиран FAST-TLE алгоритъм, подходящ за такива разпределения. Предимствата на този подход в сравнение с МПО е илюстриран чрез симулационни експерименти. Предложена е процедура за робастно клъстериране на данни, основано на тримираното класификационно правило. Изследвано е поведението на МПО и $WTLE(k)$ оценките с Монте Карло симулационни експерименти. Изследвани са данни с различен процент на тримиране и различен брой компоненти на сместа, направени са съответни изводи. Резултатите от тази глава са публикувани в статиите [5] и [11].

Глава 6 третира оценяването на неизвестните параметри на обобщени линейни модели с разпределения от дисперсионната фамилия от разпределения, частен случай на която са разпределенията от линейната експоненциалната фамилия. При съвместното оценяване на неизвестни средна стойност и дисперсия на разпределение обикновено се използват обобщени квази-правдоподобни оценки. Тук е въведена робастна версия на такива оценки $ETQL(k)$, основани на тримиране на функцията на квази-правдоподобие, който е робастен спрямо наличието на несъгласувани наблюдения в данните. Основният резултат е Theorem 6.1, която характеризира праговата точка на $ETQL(k)$ оценките. За приближено пресмятане на оценките е направено обобщение на FAST-TLE алгоритъма. Предложен е и е изследван алгоритъм за приближено пресмятане на $wGTE$ оценките, който е обобщение на FAST-TLE алгоритъма, предложен

в Глава 2. Разгледани са няколко примера, с данни от Монте Карло симулации, които илюстрират предимствата на $ETQL(k)$ оценките, в сравнение с МПО оценките и обобщените квази-правдоподобни. Резултатите от тази глава са публикувани в [6].

В Глава 7 се разглежда линейна квантилна регресия чрез тримиране. Линейна квантилна регресия е широко използван статистически модел. Известно е, че оценките, получени чрез него, са много чувствителни на наличието на несъгласувани наблюдения в пространството на предикторните променливи. В литературата са известни различни подходи за редуциране на влиянието на тези наблюдения, но като цяло всеки от тях има недостатъци. За подобряване на качествата на оценки тук е въведен клас квантилни регресионни оценки $LTQR(k)$, основани на тримиране, който е робастен спрямо наличието на несъгласувани наблюдения в предикторите. В Теорема 7.1 е доказана една горна граница за праговата точка на $LTQR(k)$, В Теорема 7.2 е доказана състоятелността на оценките при определени условия. Направен е сравнителен анализ с други робастни регресионни квантилни оценки. Подходът е илюстриран с подходящи примери от Монте Карло експерименти с различни проценти на тримиране. Резултатите от тази глава са публикувани в [7].

В Глава 8 се разглежда робастен избор на предиктори в регресионен тип задачи с големи размерности чрез тримирани правдоподобни оценки с пенализация. В тези регресионен тип задачи броят на предикторите е по-голям от броя на наблюденията на извадката. Проблемът се усложнява допълнително, ако в предиктори има несъгласувани наблюдения. В обобщените линейни модели с висока размерност в пространството на предикторните променливи обикновено се използват максимално правдоподобните оценки с пенализация. Тук са въведени обобщените тримирани оценки $wGTE(k)$ с пенализация с цел намаляване на влиянието на несъгласуваните наблюдения в извадката и едновременно за редуциране на броя на предикторните променливи. Частен случай на тези оценки са $TLE(k)$ оценките с пенализация. Предложени са робастни двустъпкови скрининг процедури TSIS и TISIS за избор на предикторни променливи, основани на $TLE(k)$ оценките и $TLE_j(k)$ оценките с пенализация. Горни граници за праговите точки на тези оценки са определени чрез параметъра $N(X)$ за d-пълнота и се основа на резултатите за обобщените линейни модели от Глава 2 и Глава 6. За приближено намиране на оценките в тези процедури е използван FAST-TLE алгоритъма. Изследвано е поведението на МПО с пенализация и $TLE(k)$ оценките с пенализация върху данни от Монте Карло симулационни експерименти с различни проценти на тримиране и различен процент на несъгласувани наблюдения в данните. Резултатите от тази глава са публикувани в [8].

Като обобщение на съдържателния анализ на дисертационния труд, може уверено да се твърди, че Нейко Нейков е получил редица важни научни и научно-приложни резултати в една съвременна област на приложната статистика. Неговите приноси са съществени за развитието на робастната статистика и са получили признание с многобройни цитирания на статиите му от учени от областта.

Дисертацията е написана въз основа на 11 научни статии, публикувани от 2003 г. насетне. Осем от публикациите са в рецензирани научни списания, като пет от тях са в списания с импакт-фактор (J. Statist. Plann. and Inference - 1; Comput. Statist. and Data Anal. - 3; Stat. Papers - 1).

Авторът е представил списък от 214 цитирания на свои статии. Ще отбележа специално две от работите с по-голям брой цитати:

- Статия [5], съвместна с P. Filzmoser, R. Dimova, и P. Neytchev, от 2007 г. е цитирана поне 100 пъти от други автори.
- Статия [1], съвместна с Ch. Muller, от 2003 г. е цитирана 45 пъти.

Всички статии, на които се основава дисертацията, са в съавторство. Смятам, че в повечето от съвместните статии доц. Нейков има водеща роля, а в останалите участието му е равностойно на другите автори.

Нямам съществени критични бележки, които да будят съмнение в приносите на дисертационния труд. Прави впечатление, обаче, че структурата на дисертацията не е достатъчно хомогенен цялостен текст, а по-скоро компилация на отделни статии. Така много от общите дефиниции се повтарят в различни глави. Например дефиницията за d-пълнота, освен в увода, се среща във всички други глави. Смятам, че структурата би се подобрила, ако осемте глави се групират в по-малко на брой със сходни задачи, например Глава 1 и Глава 2.

Авторефератът представлява съкратена версия на дисертацията на български език и е в обем от 90 стр. В него правилно са обобщени основните приноси на дисертацията.

Заклучение. Направеният по-горе анализ показва, че дисертационният труд напълно удовлетворява съвкупността от критерии и показатели за придобиването на научна степен „доктор на науките“ съгласно ЗРАСРБ, неговия Правилник и Правилниците за прилагане на ЗРАСРБ на БАН и ИМИ. Ето защо убедено предлагам на Научното жури да присъди на неговия автор доц. д-р Нейко Матеев Нейков научната степен „доктор на науките“ в област на висше образование 4. Природни науки, математика и информатика, професионално направление 4.5. Математика, научна специалност „Теория на вероятностите и математическа статистика“.

София, 26 август 2016 г.

Подпис:

Евгения Стоименова