

РЕЦЕНЗИЯ

На дисертационния труд „Робастно статистическо моделиране чрез тримиране“
на доц. д-р Нейко Матеев Нейков, НИМХ-БАН,

за получаване на научна степен „доктор на науките“ в област на висше образование

4. Природни науки, математика и информатика, професионално направление

4.5. Математика, научна специалност 01.01.10 „Теория на вероятностите и математическа статистика“

Рецензент: Огнян Иванов Кунчев, Професор, дмн, ИМИ-БАН

Нейко М. Нейков е завършил с Диплом по математика, притежава квалификацията „магистър по математика“ по специалността „Теория на вероятностите и математическа статистика“ от Факултета по математика и механика на СУ „Кл. Охридски“ през 1976 г. с много добър успех. През периода 1993-1995г. е бил свободен аспирант към НИМХ-БАН, с научен ръководител/консултант доц. д-р Д. Въндев. През учебната 1976-1977г. е работил като хоноруван асистент по математика във ВИХП, гр. Пловдив. През периода 1978-1988г. е работил в ЕЦ по Биология, БАН, като научен сътрудник в лабораторията по Математическо моделиране и анализ на данни. От 1988г. до 1994г. е работил като научен сътрудник в направление Информатика на НИМХ, БАН. През периода 1994-2000г. е н.с. в направление Краткосрочни и средносрочни хидро-метеорологични прогнози, в НИМХ, БАН. От 2000г. до настоящия момент работи в Департамент прогнози на времето, НИМХ, БАН, като доцент. През 1996г. година защитава дисертация на тема „Робастни методи в многомерния статистически анализ“ пред СНС при ВАК, по Математика, Информатика и Механика. През 2000г. е избран за ст.н.с. II ст. по специалността „Статистически методи в метеорологията и хидрологията“ от СНС при ВАК, по Математика, Информатика и Механика.

Областите му на научна квалификация са в класически и робастни статистически методи, алгоритми и програмни системи за анализ на данни, класификация и дискриминация, статистическо моделиране в метеорологията, хидрологията и опазване на околната среда по създаване на пространствено-времеви стохастични модели на дневните суми на валежите, физико-статистически модели на часовите концентрации на азотен диоксид, оценка на риска от засушаване и наводнения.

Нейко М. Нейков е бил ръководител на група „Прогностични методи и системи“, 2003-2014г. и „Оценка на прогнозите“, 2015г. до днес, Департамент Прогнози на времето. Чете лекции по Обобщени линейни модели и модели на екстремални събития на студентите от магистърска програма на катедра Вероятности операционни изследвания и статистика (ВОИС), ФМИ, СУ Св. „Кл. Охридски“ от 2002 насам. Чел е лекции по Обобщени линейни модели в съвместната магистърска програма по Приложна статистика на ИМИ-БАН с НБУ. Бил е ръководител на повече от 10 дипломанта/магистри от ФМИ.

Член е на НС на НИМХ, БАН от 2008г. насам, на Българското Статистическо дружество, на Международната асоциация по изчислителна статистика (IASC), член на борда на директорите на тази организация за Европа през периода 2008-2012г., а през 2014г. е избран за член на Международния статистически институт (ISI).

Участвал е в 20 международни проекта на FP7, DFG, COST, Техническият Университет на Виена, Гьотингенския Университет, Техническият Университет на Делфт, Техническият на Лисабон. Чел е интензивни курсове по робастна статистика и статистическо моделиране на студентите от докторските програми по статистика от Университетите на Неапол и Торино. Над 30 пъти е изнасял доклади на международни научни конференции, като 12 пъти е бил с доклад по покана.

В списъка от публикации по дисертацията са представени 11 работи, публикувани в известни международни и наши списания, като J. of Statistical Planning and Inference,

Computational Statistics and Data Analysis, Statistical Papers, Доклади на БАН, Плиска и трудове на конференции по Робастна статистика. Тези публикации са направени след получаването на званието доцент през 2001г.

Като резюме, бих казал, че горните биографични данни за автора дават ясна представа за обстоятелствата по написване на настоящия труд, който е кулминация на неговите многогодишни усилия, и е плод на големия му професионален опит, натрупан в областта на Приложната статистика и по общо, в областта на Анализ на данни.

Обем на дисертацията:

Дисертацията съдържа 200 машинописни страници, включващи увод, 8 глави и са цитирани над 200 литературни източника (статии и книги). Дисертацията е написана на английски език, а авторефератът е на български. В изложението е използвана двойна номерация.

а) Съдържателен анализ на научните и научно-приложните постижения в дисертационния труд. Характеризиране на основните постижения.

Един от фундаменталните проблеми на Приложната Математическа Статистика, по-общо на Анализ на Големи масиви от данни (Machine Learning и Big Data), е идентифицирането и редуцирането на **нетипични** наблюдения, наричани на английски **outliers**, които на български е възприето в Математическата Статистика да се наричат **несъгласувани наблюдения**. Настоящият дисертационен труд е посветен на развитието на един нов подход за решаването на проблема с несъгласуваните наблюдения, който се състои във въвеждането на нови статистически робастни методи с висока прагова точка, основаващи се на тримиране (адаптивно игнориране/изключване на част от наблюденията). Основно постижение на тези методи е характеризиранието на праговата точка на предложените статистически оценки. Напълно в духа на Приложната статистика, е предложена и методология за определяне на праговите точки, така наречената техника на d -пълнота, която е иницирана още от Димитър Въндев през 1993 г.

По същество, настоящият дисертационен труд прави по-нататъшно развитие на тази техника и я прилага към нови класове от статистически оценки. Отново, в духа на Приложната статистика, са намерени ефективни алгоритми за приближено пресмятане на предложените класове статистически оценки.

Въведени са класове робастни оценки, като: S -оценките, основани на минориране - мажориране от k -ти нареден елемент на d -пълно множество от неотрицателни функции, тримирани квази-правдоподобни оценки, тримирани квантилни оценки, тримирани (правдоподобни) оценки с пенализация в задачи с висока размерност. Предложени са обобщения на техниката на d -пълнота на Въндев (1993), за да бъдат характеризирани праговите точки на някои от изброените класове оценки, основани на тримиране. Съответно, предложени са FAST-TLE алгоритъм за приближено пресмятане на тримирани правдоподобни оценки и FAST-GTE алгоритъм за приближено пресмятане на обобщени тримирани оценки.

Основните постижения на дисертацията могат да бъдат резюмирани накратко така:

1. създадена е унифицирана методология за робастно оценяване с висока прагова точка, за разкриване и редуциране на несъгласувани наблюдения в данните, на основата на тримиране на най-широко използваните класически методи за статистическо моделиране: (a) с обобщени линейни модели със случайна компонента от дисперсионната фамилия от разпределения, частен случай на която е линейната експоненциална фамилия от разпределения; (b) с множествена линейна квантилна регресия; (c) със смеси от разпределения от линейната експоненциална фамилия от разпределения; (d) с линейни регресионни модели с разпределения на екстремалните стойности на грешката; (e) в задачи с големи размерности (на съвременен език, Big Data, където броят на предикторите може да е много по-голям от броя на наблюденията) за селектиране на значими предиктори с методите на пенализацията

(регуляризация) за разпределения от линейната експоненциална фамилия от разпределения;

2. изследване поведението на изброените класове оценки и сравняването им с класическите оценки върху крайни извадки от данни с и без наличие на несъгласувани наблюдения в данните чрез методите на имитационното моделиране.

На всяко от горе-изброените постижения е посветена по една глава в дисертационния труд. Трябва да се отбележи, че получените оценки и методи носят не просто академичен характер и интерес, но са намерили и намират голям брой съществени приложения при анализ на реални данни. Това е отбелязано и по-долу, където е анализирана цитируемостта на основните публикации, използвани в дисертацията.

Нека направим и кратко изброяване на основните резултати по глави:

В Глава 1 са въведени основните понятия и дефиниции. Дефиниран е клас S -статистически оценки, основани на минориране - мажориране от k -тия елемент на d -пълно множество от функции и с предложената разширена техника на d -пълнота е характеризирана праговата им точка. Намерена е по-ниска граница за праговата точка на $wGTE(k)$ оценките без допълнителни условия за обема на извадката и параметъра на тримиране k . Тези резултати представляват обобщение на Теорема 1 от публикацията на Vandev and Neykov (1998). Определена е връзката на индекса на d -пълнота с параметъра $N(X)$, характеризиращ максималния брой повторени наблюдения в пространството на предикторите X в обобщените линейни модели с разпределение от линейната експоненциална фамилия. Като следствие на това, са характеризирани праговите точки на $WTLE(k)$ оценките за класическия линеен множествен регресионен модел с нормално разпределение на грешката, линейната бинарна (логистична) регресия и лог-линейната (Поасонова) регресия, а също така на линеен регресионен модел с разпределение на Лаплас от ред q с дисперсионен параметър, и на линейните регресионни S_{α} оценки. Значително са опростени доказателствата на основните резултати на Vandev-Neykov (1998). В Забележка на стр. 18 от Автореферата е изказано твърдение за валидност на резултатите за произволно d -пълно множество от положителни функции, а не само за плътности на разпределения, което считам за вярно. Резултатите от тази глава са публикувани в Neykov and Muller (2003), [1].

В Глава 2 е предложен локално сходящ FAST-TLE алгоритъм за приближено пресмятане на неизвестните параметри чрез $TLE(k)$ оценките. Дадени са няколко примера, чрез които е илюстрирано определянето на индексите на d -пълнота и $N(X)$.

В Глава 3 е предложена обобщена техника на d -пълнота за изследване на праговата точка на $wGTE(k)$ и $WTLE(k)$ оценките, поради ограничителното условие за субкомпактност на функция в ситуации на комплексни вероятностни модели. Предложения 3.1 и 3.2 представляват необходими условия за съществуване на решение на $wGTE(k)$ оптимизационната задача и същевременно дават долна граница на праговата точка за множество от непрекъснати функции, удовлетворяващи условието A_2 . Тези резултати представляват обобщение на Теорема 1 на Vandev-Neykov (1998) и Теорема 2 на Muller-Neykov (2003). Определен е индекса на d -пълнота, чрез параметъра $N(X)$, вследствие на което с Предложение 3.3 е характеризирана праговата точка на $WTLE(k)$ оценките за линейната логистична (бинарна) регресия с обобщена свързваща функция, което представлява обобщение на резултата за линейната логистична (бинарна) регресия, разгледан от Müller-Neykov (2003), [1]. Резултатите от тази глава са публикувани в Dimova-Neykov (2004), [3].

Глава 4 е посветена на $TLE(k)$ оценяване на параметрите на обобщеното разпределение на екстремалните стойности (GEV) с обобщени линейни предиктори в параметрите на положението (локацията) и мащаба (скалата). С Монте-Карло методите е изследвано поведението на максимално правдоподобните и $TLE(k)$ оценки по данни с и без наличие на несъгласувани наблюдения в зависимата и предикторни променливи с различни проценти на тримиране. Резултатите от симулационните експерименти убедително показват, че МП оценки са ненадеждни при наличие на несъгласувани наблюдения в данните, за разлика от

TLE(k) оценките при процент на тримиране по-висок от процента на замърсяване на данните с несъгласувни наблюдения. Използван е FAST-TLE алгоритъмът за приближено пресмятане на неизвестните параметри. Праговата точка на TLE(k) оценките за разпределението на Гумбел, частен случай на GEV разпределението, е характеризирана с Предложение 4.1, което е дадено в Апендикса към тази глава. Резултатите от тази глава са публикувани в Neykov et al. (2005), [4]

По-специално внимание бих обърнал на Глава 5, където са изложени резултатите от статия под номер [5], Neykov et al. (2007), получила широк отзвук сред специалистите. В тази глава са разгледани тримираните максимално правдоподобни оценки (WTLE(k)) за робастно оценяване на параметрите на смес от вероятностни разпределения, като е направено приложение на основните резултати на дисертацията към много важната от практическа гледна точка задача за **идентифициране на структури в нееднородни данни и робастно клъстериране при наличие на голям процент на несъгласувани наблюдения в данните**. В апендикс към тази глава е характеризирана праговата точка на WTLE(k) оценките за крайна смес от разпределения с техниката за d-пълнота. Предложен е адаптивен подход за определяна на параметъра на тримиране k, основан на тримирания информационен критерий на Бейс. За пресмятане на неизвестните параметри на смесите от разпределения е използван споменатият по-горе авторов FAST-TLE алгоритъм, с подходяща адаптация. Използвайки тези оценки, авторът предлага оригинална процедура за **клъстериране на данни**. Този метод е подложен на изчерпателна серия от експериментални проверки върху симулирани с Монте Карло данни, като са варирани броят на компонентите на сместа от разпределения, процентът на замърсяване с несъгласувани наблюдения и процентът на тримиране на данните. Експериментите показват превъзходството на предложените от автора оценки над метода на Максимално правдоподобните оценки. Не случайно този метод на автора за робастно клъстериране е забелязан и използван за анализ на реални данни от други автори.

Глава 6 е посветена на робастното оценяване на неизвестните параметри на обобщени линейни модели с разпределения от дисперсионната фамилия от разпределения. Предложен е клас ETQL(k) оценки, основани на тримиране на функцията на квази-правдоподобие, който е робастен спрямо наличието на несъгласувани наблюдения в данните. Теорема 6.1 характеризира праговата точка на ETQL(k) оценките чрез по-големия от индексите на d-пълнота на обобщените линейни модели на очакването и дисперсионния параметър, до които се редуцира максимизирането на ETQL(k) целевата функция. За разпределенията от тази фамилия с дисперсионния параметър константа, Теорема 6.1 представлява обобщение на Теорема 1.3, 1.4 и 1.5 на Neykov-Muller (2003), [1] за линеен множествен регресионен модел с постоянна дисперсия, линейната бинарна (логистична) и лог-линеен (Посонов) регресионен модел. Предложен е FAST-wGTE алгоритъм за приближено пресмятане на wGTE(k) оценките. Предимствата на ETQL(k) оценките, в сравнение с квази-правдоподобните оценки, е илюстрирано чрез примери и разширено симулационно изследване с методите Монте Карло. Резултатите от тази глава са публикувани в Neykov et al. (2012), [6]. Ще отбележа, че в тази статия е характеризирана праговата точка на ETQL(k) оценките, но без формулирането на Теорема 6.1.

Предмет на изследване в Глава 7 е робастифицирането на квантилната линейна множествена регресия. Дефиниран е клас квантилни регресионни оценки LTQR(k), основани на тримиране, който е робастен спрямо наличието на несъгласувани наблюдения в пространството на предикторите. Праговата точка и състоятелност на тези оценки са характеризирани с Теорема 7.1 и 7.2. За приближено пресмятане на LTQR(k) оценките на неизвестните параметри е използвана адаптация на FAST-wGTE алгоритъма. Проведен е сравнителен анализ между LTQR(k) оценките и оценките на класическата квантилна регресия чрез примери и разширено симулационно изследване с методите Монте Карло. Резултатите от тази глава са публикувани в Neykov et al. (2012), [7].

Глава 8 е посветена на задачи от регресионен тип с висока размерност, където броят на предикторите може да е много по-голям от броя на наблюденията в извадката, чрез тримиране и пенализация. Дефиниран е клас wGTE(k) оценки с пенализация за

преодоляването на проблема с несъгласуваните наблюдения в извадката и редуциране размерността на предикторите. Предложени са робастни двустъпкови скрининг процедури TSIS и TISIS за селектиране на значими предикторни променливи, основани на TLE(k) оценките и TLE(k) оценките с пенализация, частен случай wGTE(k) оценки с пенализация. Характеризирани са праговите точки на тези процедури чрез параметъра на индексите на дълготайността и $N(X)$, на основата на резултатите за обобщените линейни модели в Müller-Neykov (2003), [1] и Neykov et al. (2012), [6]. За приближено намиране на оценките в тези процедури е използван FAST-TLE алгоритъма. С методите Монте Карло, в разширен симулационен експеримент при различни проценти на тримиране и несъгласувани наблюдения в данните, е изследвано поведението на максимално правдоподобните оценки с пенализация и TLE(k) оценките с пенализация. Резултатите от тази глава са публикувани в Neykov et. al (2014), [8].

б) Общо описание на публикациите, които отразяват дисертацията – монографии, статии, свидетелства и патенти, класифицирани по тематика или друг признак и редуциране поради съвпадение или припокриване.

Настоящият дисертационен труд се основава на публикациите с номера 1,2,4,5,6,7 и 8. Трета глава е основана на публикации 3,9 и 10. Пет от статиите са публикувани в реномирани списания, споменати по-горе, със значителен Импакт фактор. Останалите, с номера 2,3 и 11, са публикувани в реномирани издания, като Development in Robust Statistics, Springer Verlag; Theory and application of recent robust methods, Birkhauser, Proc. in Computational Statistics, Physica-Verlag. Две от статиите са публикувани в списание Плиска и Доклади на БАН.

в) Отражение на резултатите на дисертацията в трудовете на други автори. Числови показатели - цитати (без автоцитатите), импакт-фактор и др.

Съществено е да отбележим високото цитиране на статиите, послужили за написването на дисертационния труд, като това би било също така оценка за научното ниво на самия труд.

Статията с номер 1 е цитирана 45 пъти от други автори, в статии публикувани в реномирани списания като Statistical and Probability Letters, Computational Statistics and Data Analysis, Statistical Planning and Inference, Encyclopedia of Actuarial Sciences, Econometric theory, Statistical sciences, J. of American Statistical Association, Medical Image Analysis, Neuroradiology, Information Sciences, Advances in Data Analysis and Classification, Neural Networks, Technometrics.

Статията с номер 2 е цитирана 28 пъти от други автори, в статии публикувани в реномирани списания като Statistical Data Analysis, Intelligent Data Analysis, Communications in Statistics - Simulation and Computation, IEEE Trans. on Pattern Analysis and Machine Intelligence.

Статията с номер 3 е цитирана 11 пъти от други автори в реномирани списания по разпознаване на образи в Биомедицинските изследвания.

Статията с номер 5 е цитирана 100 пъти от други автори в реномирани списания по разпознаване на образи в Биомедицината, IEEE Trans. On Medical Imaging, Statistics and Computing, Pattern Recognition, etc.

Статията с номер 6 е цитирана шест пъти, статия 7 - четири пъти, напр. в European J. of Operations Research, 8 – четири пъти, 9 – два пъти, 11 – единадесет пъти.

Големият брой цитирания на статия 5 вероятно е поради определено големия интерес към разработения робастен метод за разпознаване на образи в условията на неопределеност при наличие на висок процент шум, идентификация на множествена склероза и тумори при анализ на изображения от Ядрено-магнитен резонанс.

г) При колективните публикации приносът на кандидата може да се характеризира като равностоен:

В публикациите, на които се основава дисертационния труд, 5 са с един съавтор, 3 са с двама съавтори и 3 са с трима съавтори. Оценявам за равностойно участието на доц. Нейков и съавторите му, поради непредставени разпределителни протоколи за индивидуални

приноси. В заключение, бих искал да подчертая неговата водеща роля, въпреки този формален извод, отчитайки резултатите в предишните му публикации по робастна статистика, основана на тримиране, които не са включени в настоящата дисертация.

д) Критични бележки и препоръки на рецензента.

Рецензентът би предложил настоящият текст да бъде изгладен и редактиран, за да бъдат отстранени повторенията в различните глави (на дефиниции и означения). Отчитайки големия интерес от страна на специалистите в областта на Приложна Статистика и анализ на данни с високи размерности, считам, че една подходяща редакция на този текст ще намери добър прием в подобаваща серия по Приложна Статистика, например Lecture Notes in Statistics на Springer Verlag, или в Springer Series in Statistics.

е) Качества на автореферата, вкл. доколко правилно отразява приносите на дисертацията.

Авторефератът е направен много обстойно на 90 страници, написан е на български, и в него са отразени всички резултати от дисертационния труд.

Казаното дотук ми позволява да твърдя, че доц. д-р Нейко М. Нейков е написал един дисертационен труд с високо научно съдържание в областта на Приложната статистика, който е на световно ниво съгласно публикациите, които са публикувани в международни реномирани издания, а също са получили голямо признание, за което говорят множеството цитирания. Считам, че неговите приноси в областта на робастното статистическо оценяване напълно заслужават най-висока оценка.

Всичко това ми дава право да препоръчам на Уважаемото Жури да присъди на доц. д-р Нейко М. Нейков научното звание „доктор на науките“ в областта на висшето образование: 4. Природни науки, математика и информатика; професионално направление: 4.5. Математика; научна специалност: 01.01.10 „Теория на вероятностите и математическата статистика“

30.08.2016 г.

София

Рецензент: