РЕЦЕНЗИЯ

за дисертацията на Боно Стойчев Нончев
на тема:
"MDL принцип за избор на модел при анализ на данни (Model Selection for Data Analysis Based on the MDL Principle)"
представена за присъждане на образователна и научна степен "Доктор",
Област на висше образование: 4. Природни науки, математика и информатика, в професионално направление
4.5 Математика, научна специалност "Теория на вероятностите и математическа статистика".
Рецензент: проф. д-р Марусия Никифорова Божкова – ФМИ, СУ "Св. Климент Охридски"

Представям рецензията си по тази защита като член на Научното жури, определено със Заповед No РД 38-388/5.06.2015 г. на Ректора на СУ "Св. Климент Охридски" съгласно Решение на ФС (Протокол No 5, 26.05. 2015 г.). Рецензията е изготвена според изискванията на:

- Закона за развитието на академичния състав в Република България (ЗРАСРБ),
- Правилника за прилагане на ЗРАСРБ,
- Правилника за условията и реда за придобиване на научни степени и заемане на научни длъжности в СУ "Св. Климент Охридски",
- Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности във Факултета по математика и информатика при СУ "Св. Климент Охридски".

Обща информация за докторанта

Боно Нончев е задочен докторант във ФМИ–СУ към катедра "Вероятности, операционни изследвания и статистика" (ВОИС) от м. юли 2011 г. до м. юли 2014 г., отчислен с право на защита. Завършва висше образование – бакалавър, специалност "Приложна Математика" през 2008 г., а по-късно през 2010 г. завършва магистратура по същата специалност, специализация "Вероятности и статистика", със защита на дипломна работа на тема: "Modeling and managing the risk utilizing the storage facility", с научен ръководител доц. д-р Пламен Матеев.

Междувременно от май 2006 г. до настоящия момент е работил последователно в Лакорда АД, София по поддръжка и адаптация на CLucene – машина с приложение в правно–административната система за бързо търсене на документи; в Centre for Human Language, Technology and

Bioinformatics, University of Beira Interior, Ковиля, Португалия и е отговарял за преработка и имплементация на WISE (нов алгоритъм за метатърсене); в ТайръсБлу ЕООД, София като ръководител на екип програмисти на Java, обслужващ компанията Jetix Inc., Лондон; в Ладоре ООД, София по разработка на концептуални приложения за мобилни телефони. От април 2010 работи във ФинАналитика ООД като математик върху изследване на статистически модели и анализи, тестване на модели и имплементация.

Във ФМИ към катедра ВОИС води упражнения по Статистика и емпирични методи (СЕМ) – задължителна дисциплина за бакалаври спец. ИС и Кодиране и теория на информацията – избираема дисциплина.

Анализ на съдържанието, резултатите и приносите на дисертационния труд

Дисертационният труд е в областта на статистическото оценяване на разпределение на дадена извадка като е свързан с теория на информацията, доколкото изследва в каква степен извадката съдържа информация за неизвестното разпределение, от което тя е генерирана. В тази връзка по–точно може да се каже, че дисертационният труд е свързан с идентифициране на разпределението на извадка – задача от особена важност, тъй като предхожда всички по–нататъшни статистически процедури за оценяване.

Тази задача възниква в редица важни приложни задачи, една от които е например прогнозиране бъдещото поведение на финансовите пазари. Също така тя се явява от ключово значение в областта на биостатистическите приложения на класове стохастични процеси като модели на динамиката на клетъчни популации, където от съществено значение е идентификацията на разпределението, от което е направена извадката. Всичко това налага извода, че задачата, на която е посветена дисертацията е всеобхватна и многообразна и присъства в съвременните научни изследвания във връзка със статистическо моделиране в различни области на човешката практика. Ще отбележим, че проблемите в дисертационния труд са били обект на научния интерес на Колмогоров в далечната 1963 г. и са вдъхновяващ пример за научна работа до наши дни.

Дисертацията съдържа 165 стр. текст на английски език, структуриран в пет глави включително увод и заключение и списък с литература, състоящ се от 69 заглавия, от които приблизително 70% са от последните 5–10 години, което показва, че дисертантът има задълбочени познания върху съвременното състояние на проблематиката. Добавени са три приложения, които освен че дават необходимите за разбирането на резултатите дефиниции на основни и специфични понятия от теория на информацията, изчисляване на бета функцията на Дирак и засягат един

сравнително нов оптимизационен метод и програмната му реализация, могат да служат като помагало и за образователни цели. Приложната част е поместена в обем от 37 стр. от общия обем на дисертацията.

Преглед на резултатите по глави.

В дисертацията се решават три основни задачи:

(I) Теоретично разработване на метод, основан на MDL (Minimum Description Lenght) принципа (по–нататък ще използвм английската абревиатура) за определяне на неизвестно разпределение, от което е генерирана конкретна извадка чрез въвеждане на новото понятие *сложност на разпределение*:

- за некорелирани извадки и фамилии разпределения, характеризирани с параметри за мащаб и локация;
- за некорелирани извадки и сферични фамилии разпределения, характеризирани с параметри за мащаб и локация;
- за независими извадки и фамилии разпределения, характеризирани с параметри за мащаб и локация;
- за независими извадки и фамилии разпределения, характеризирани с параметри за мащаб и локация и за форма на разпределението.

(II) Разработване на алгоритми за пресмятане на теоретично изследваните методи за оценка на сложността на разпределение при фамилии с мащаб и локация за некорелирани и независими извадки.

(III) Приложение на получените резултати за моделиране на симулационни данни и реални данни от областта на финансовите пазари, последните от които по своята същност са и мотивацията за разработването на горе–споменатите методи и алгоритми.

Преминавам към съдържателен анализ на научните и научно–приложни постижения в дисертацията.

Уводната **Глава 1** има за цел да направи преглед на известните методи и подходи за разпознаване на неизвестно разпределение на извадка между определен брой потенциални кандидати за това като в отделните параграфи последователно се дефинират основните понятия и критерии на класическото статистическо (фрикуентистко) оценяване; метода на максималното правдоподобие, Бейсовия подход, концепцията за достатъчните статистики на Фишер и тяхната характеризация в термините на Теоремата за факторизаци на Фишер–Неймън, както и достатъчността в

Гаусовата фамилия от разпределения. Специално внимание е отделено на информационните подходи и мерки за оценяване на разстоянието между две разпределения – дефинирани са основните понятия от теория на информоцията - ентропия и дивергенция на Кулбак– Лайблер, свързани с оптималните кодове в теорията на кодирането. В последния параграф се разглеждат двете концепции за обяснителната сила (goodness of fit) и обобщимостта (generalizability) при избор на модел и това е илюстрирано с подходящ пример.

Глава 2 е посветена специално на принципа MDL и е направен преглед на класическите резултати в областта, както и на връзките с теория на информацията, теорията на сложността по Колмогоров и статистиката. MDL принципът е разработен основно от Рисанен в серия от статии, първата от които е от 1978 г. и е съдържание на монографията на Grunwald, 2007, посветена на класическите резултати в областта. Тази глава, освен със задълбочения прочит на 40 статии от периода 2007-2014 година, формулира точно принципа на MDL и обосновава необходимостта от въвеждането на новото понятие *сложност на разпределение* (на английски distribution complexity (DC)), което е оригинална идея на докторанта. Съществуващите модификации на MDL принципа се основават на идеята, че на всяко разпределение съответства оптимален код (т.е. който постига очаквана дължина равна на ентропията) и обратното на всеки код съответства разпределение, при което кодът е оптимален и в тази връзка един такъв конкретен код, който се използва за конструиране на еквивалентност между код и разпределение е кодът на Шанън. Основният проблем при този подход е избора на еднозначен критерий за избор на код на хипотезата, която се проверява. По-късно Рисанен през 1989 г. въвежда понятието стохастична сложност (stochastic complexity (SC)) на модела, който се предлага като описание на данните, в чиято дефиниция се появяява именно математическия проблем, свързан с разходимост на интеграла, използван за дефиниране на понятието сложност на модел. Тук трябва да се отбележи, че е мястото на разработената от докторанта нова концепция за елиминиране на проблема с разходимостта, което е безспорен принос към теорията на MDL принципа в частност и теория на информацията и статистиката като цяло.

Основните теоретични резултати, получени в дисертационния труд са свързани със създаване на приложим метод, базиран на MDL принципа за различаване на разпределения с тежки опашки от Гаусовите и са съдържание на **Глава 3**. Като мотивационна задача са разгледани две многомерни фамилии от разпределения, а именно многомерното Гаусово и многомерното t разпределение на Стюдънт, като наблюденията се предполага, че са некорелирани. При некорелираните разпределения

сложността на разпределението е получена в затворен вид. Съществения резултат е Теорема 3.11, в която на практика се доказва, че сложността на модела, описващ данните се състои от две компоненти, едната независеща от размера на извадката (а само от граничните условия на областта на изменение на параметрите) и другата, която е новодефинираната сложност на разпределение (DC), зависеща само от маргиналното разпределение (в основата на многомерното). Последното решава проблема с разходимостта. С не по-малка важност са Лема 3.16, Лема 3.18 и Теорема 3.19, в които се установява, че тъй като генератора на сферичните разпределения се отделя от интеграла в дефиницията на DC, то DC се оказва, че зависи само от размера на извадката и генератора на сферичните разпределения. По този начин се получава в Следствие 3.21, че сферичните разпределения не са различими в термините на дължината на описание, което от друга страна обаче само по себе си се явява една нова характеризация за тази фамилия разпределения. Като илюстрация на полезността на Теорема 3.19 е пресметната DC на Гаусовото, t разпределението на Стюдънт и разпределението на Лаплас.

В параграф 3.4 се въвежда DC в общия случай за независими фамилии с параматри за мащаб и локация. За разлика от сферичните разпределения тук няма пряка връзка на параметрите и извадковите им оценки. Това налага модификация на дефиницията на понятието DC и по този начин се доказва Теорема 3.27 – аналог на Теорема 3.11 за ограничената сложност на модела. Важно е да отбележим в този случай, че така пресметната сложност може да се използва за намиране на най-добрия модел. Това е направено в параграф 4.3 за t разпределението при различни степени на свобода.

В последният параграф 3.5 е дефинирана DC за фамилии разпределения, които имат освен параметри за локация и мащаб и още един параметър за форма. Конструира се разбивка на DC по подобен начин на предишните параграфи и всичко това е доказано в Лема 3.30, Лема 3.32 и Теорема 3.34.

Глава 4 е посветена на числено пресмятане на сложността на сферичните разпределения. Представени са таблици със сложността на t разпределението, което дава възможност за разпознаване на модел на извадка от t разпределение с конкретни степени на свобода и гаусово разпределение. Разгледани са два експеримента – един симулационен и един с реални данни. Научно–приложният принос на докторанта в тази глава се състои в разработване на алгоритъм за числено намиране на сложността на t разпределение и гаусовото. Тази глава придава на дисертационния труд завършеност и пълнота по отношение на теоретични изследвания, алгоритмична имплементация и приложение върху реални

данни. Бих искала за отбележа, че високо оценявам интерпретацията на резултатите в тази глава и причислявам това също към приносите на докторанта.

Последната **Глава 5** синтезирано насочва към главните и оригинални постижения в дисертацията, както и набелязва насоките за понататъшно разширяване на резултатите върху α устойчивите разпределения на Леви и асиметричното t разпределение, които представляват особен интерес от иконометрична гледна точка.

Трябва да отбележим, че освен тази част с изследователски характер и приноси към теорията на MDL принципа, дисертационният труд е допълнен с три приложения, които значително внасят яснота в разбирането на поставения проблем. Също така има индекс на понятията, таблица на означенията, което също улеснява четенето, както и англо-български речник на понятията.

Забележки Имам забележки във връзка с дефиницята на понятието DC на стр. 74, дефиниция 3.10, където би трябвало математическото очакване да бъде от произведението на двете бета функции на Дирак. Същото важи и за аналогичната дефиниция 3.26. На стр. 82, р. 2 трябва да бъде K вместо h.

Забелязват се и технически грешки, на които няма да се спирам и считам, че не намаляват яснотата на резултатите и не предизвикват съмнение относно верността им.

Литература Цитираните литературни източници показват, че докторанта е много добре запознат със съвременното състояние на научните изследвания в областта на теория на информацията и статистиката, свързани с MDL принципа за определяне на модел на данни. Заедно с това, от получените в дисертацията резултати се вижда, че с притежаваната математическа култура и професионални компютърни умения дисертанта оригинално използува тези знания за решаването на нови задачи.

Авторефератът на дисертацията е изготвен в съответствие с изискванията на Правилника за условията и реда за придобиване на научни степени и за заемане на научни длъжности във ФМИ–СУ и едновременно пълно и точно отразява съдържанието и приносите на дисертационния труд.

Считам, че заявените от дисертанта приноси действително са такива.

Публикациите свързани с дисертацията са 3 на брой (необходими са две), една от тях е в Pliska Studia Mathematica Bulgarica, една в сборник на XVIII–та Европейската среща на младите статистици и една в авторитетната поредица Communications in Computer and Information science на Springer, която е в процедура за включване в списъка на изданията с

импакт фактор. Всички публикации са самостоятелни.

Личните ми впечатления за дисертанта са от момента на постъпване в магистърската програма "Вероятности и статистика", от участията му в Международните конференция по Вероятности и Статистика (2010, 2012, 2014), от доклади на Пролетната научна сесия на ФМИ (2012, 2013, 2014), както и от преките ни служебни контакти от постъпването му в магистърската програма "Вероятности и статистика" през 2009 досега. Отнася се с подчертана задълбоченост и прецизност при решаване на поставените проблеми. Резултатите, свързани с дисертацията са докладвани на XVIII–та Европейската среща на младите статистици в Осйек, Хърватия, 2013 и на 22-та международна конференция по нелинейна динамика на елестронните системи, Албена, България, 2014.

Заключение.

Въз основа на всичко изложено до тук считам, че представеният дисертационен труд отговаря на всички изисквания на ЗРАСБ, ПЗРАСБ и Правилниците за придобиване на научни степени и за заемане на научви длъжности в СУ И ФМИ. Убедено **препоръчвам на уважаемото научно жури да присъди на автора му Боно Стойчев Нончев образователно–научната степен "доктор"** в областта на висше образование "Природни науки, математика и информатика", професионално направление " Математика".

Дата: 24. 08. 2015 г. Подпис:.....

проф. д-р Марусия Божкова