РЕЦЕНЗИЯ

за дисертацията на Деница Парашкевова Григорова
на тема:
"ЕМ алгоритми за "probit" модели със случайни ефекти"
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: доц. 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 38-655/ 19. 12. 2013 г. на Ректора на СУ "Св. Климент Охридски" съгласно Решение на ФС (Протокол No 12/ 16. 12. 2013 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

Деница Григорова е редовен докторант във ФМИ–СУ към катедра "Вероятности, операционни изследвания и статистика" (ВОИС) от м. юни 2009 г. до м. юни 2012 г., като в периода февруари 2011 г. – юни 2011 г. специализира в областта на Статистиката в Шефилдския университет по програмата Еразмус. Завършва висше образование – бакалавър, специалност "Приложна Математика" през септември 2007 г., а по-късно през март 2009 г. завършва магистратура по същата специалност, специализация "Вероятности и статистика", със защита на дипломна работа на тема: "Оценяване на ефекта на генотипа и други фактори върху страничните ефекти от лъчетерапия при онкоболни", с научен ръководител доц. д-р Ралица Георгиева. Междувременно от май 2009 г. до септември 2012 г. работи последователно като математик и хоноруван асистент към катедра ВОИС, а от септември 2012 г. след спечелен конкурс е назначена

за асистент към катедра ВОИС. Води множество упражнения по: Вероятности и статистика, Статистика и емпирични методи (СЕМ), Увод в статистиката, Практикум по СЕМ в бакалавърските програми на ФМИ, както и упражнения по Линейни модели с R, Увод в Биостатистиката и Биостатистика за магистри, за които е разработила и обновила учебни помощни материали.

Анализ на съдържанието, резултатите и приносите на дисертационния труд

Дисертационният труд е в областта на моделирането и е посветен на разработването на нови стохастични модели и алгоритми, свързани с данни от дългосрочни наблюдения във времето, мотивиция за което са реални приложения в медицината.

Темата на дисертационния труд безспорно е много актуална, тъй като статистическите методи играят все по–съществена роля в редица области на научната и научно-приложната практика като утвърдена методология за анализ на данни от проучвания с цел прогнозиране, особено съчетани с изчислителната мощност на съвременните компютри.

Дисертацията се състои от 106 страници, включващи увод, четири глави, заключение и списък с литература, съдържащ 72 заглавия, от които приблизително половината са от последните 5–10 години, което показва, че дисертантката е запозната със съвременното състояние на проблематиката. Добавени са три приложения с R кода на ECM алгоритми за разработените 3 нови модели в дисертацията.

3. Преглед на резултатите по глави.

Нека още в самото начало да поясним утвърдилото се на български език съкращение EM и ECM, което ще се употребява многократно по–нататък. Това са стохастични алгоритми, чието наименование произтича от английското наименование Expectation/Maximization и Expectation/ Conditional Maximization, съответно. Самите алгоритми представляват итеративни процедури, в които на всяка итерация има две стъпки – намиране на очакването (E–стъпка) и максимизиране, при което се пресмятат новите стойности на параметрите (M–стъпка) и съответно нейната модификация CM– стъпка, при която максимизирането се извършва постъпково по всеки от параметрите на модела като се фиксират останалите. Основната цел на разработените в дисертацията ECM алгоритми е намиране на максимално–правдоподобни оценки (МПО) на параметри в модели, които зависят от **ненаблюдаеми данни**.

Най-общо, в тази връзка трябва да се каже, че съвременната статистика разчита на повишения капацитет на компютрите и на тази база развива методи, които в миналото са били пренебрегвани поради голямата си изчислителна сложност. Един от тези методи, който в послед-

ните три десетилетия се развива и доказва като статистическа практика е EM (Expectation/ Maximization) алгоритъмът, наречен така от Демпстър, Леърд и Рубин през 1977 г. Това е широко приложим алгоритъм, който предлага итеративна процедура за намиране на МПО за статистически модели с липсващи данни, в които такава оценка би била директно изчислима при наличието на тези допълнителни данни.

В дисертацията се решават две основни задачи:

(I) Разработване на ECM алгоритми за оценяването на неизвестни параметри на три корелирани "probit" модела:

- на една наредена категорна променлива с повторни наблюдения;
- на корелиран "probit" модел на няколко категорни променливи;
- корелиран модел на две променливи, проследени неколкократно във времето от различен тип – нормална и наредена категорна.

  Като неразделна част от разработването на софтуера за горе– посочените процедури е получена оценка на грешката, валидиране на разработените алгоритми и експериментално потвърждаване чрез симулационно моделиране.

(II) Приложение на получените резултати за моделиране на реални данни от областта на медицината, което по своята същност е и мотивацията за разработването на горе–споменатите алгоритми.

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

Глава 2 има обзорен характер върху развитието на моделите за данни с повторни наблюдения като се разглеждат по–подробно моделите със случайни ефекти, като последователно в параграф 2.1 са представени линейните модели с фиксирани ефекти, т.е. когато откликът за всеки индивид има нормално разпределение и очакването му е линейна функция на вектора на предикторите, които считаме за фискирани (неслучайни), а в параграф 2.2 се разглеждат модели със смесени ефекти, т.е. отклика се моделира като към фиксираните ефекти са добавени и случайни ефекти, като по този начин на практика се отчита наличието на зависимост на наблюденията.

В следващия параграф на същата глава са изложени обобщените линейни модели (ОЛМ), съответно с фиксирани и смесени ефекти. Те са обобщение на линейните модели, в което за отклика се предполага, че е реализация на случайна величина (сл. в.), чието разпределение е от по–общ клас, а именно от класа на експоненциалната фамилия разпределения, включваща, например, известните нормално, биномно, гама и Поасоново разпределения.

Изложението в тази глава ясно показва разликата на развитите от дисертантката модели от широкия клас ОЛМ и техни разширения със смесени ефекти, за които съществува развит софтуер.

По–специално, **Глава 3** е посветена на оценяване на неизвестните параметри на корелиран "probit" модел със случайни ефекти за една наредена категорна променлива от дългосрочни проучвания или повторни наблюдения (на англ. longitudinal data). Нека поясним, че т. нар. "'probit" модел е вид регресия, в който зависимата променлива, наречена още отклик, може да приема само две стойности. Накратко, целта на модела е да се оцени вероятността дадено наблюдение със специфични характеристики да бъде класифицирано към определена категория. Освен това, ако като критерий се използва изчислената вероятност да е по–голяма от 1/2 и тя се третира като класификатор на наблюдението в прогнозирана категория, то "probit" моделът може да се разглежда като вид бинарен модел на класификация. "Probit" моделът е популярен като разновидност на моделите с ординален или бинарен отклик. Като такъв, той се използва при подобни проблеми, при които се прилага и логистичната регресия. "Probit" моделите са въведени от Chester Bliss през 1934 г., а един бърз метод за МПО за тях е предложен от Роналд Фишер през 1935 година като приложение към работата Bliss.

В разработения модел се предполага, че има наблюдения над наредена категорна променлива и има латентна (скрита) **нормално** разпределена променлива, която поражда наблюдаваната величина. Параграф 3.2 е посветен на разширение на стохастичния ECM алгоритъм, при което дисертантката умело комбинира идеята на Ruud (1997) за използване на разстоянията между съседните неизвестни прагове, вместо самите прагове с тази на Kawakatsu and Largey (1997), които развиват EM алгоритъм за многомерна нормална и едномерна ординална величина. По този начин става възможно намирането на съвместното нормално разпределение за трансформираните линейно скрити величини (вж. стр. 18, ред 3 отдолу), условно по наблюдаваната категорна величина. По–нататък с използване на явния вид на лог–правдоподобието на пълния набор от данни са получени оценките на ковариационната матрица на случайните ефекти, регресионните коефициенти за фиксираните ефекти и уравненията за разликите между съседните прагове (параграф 2.2.2). Подходът, който се следва по–нататък е определен от ECM алгоритъма, поради което се доказва, че условните очаквания, необходими в E–стъпката зависят само от първите два момента на намереното орязано многомерно нормално разпределение, условно по наблюдаваната величина. Това дава възможност да се приложи Монте Карло метод за генериране на случайни числа от въпросното разпределение. В M–стъпката специфичното

е, че всеки параметър се максимизира индивидуално при условие, че останалите са фиксирани. За оценка на стандартните грешки е използван "bootstrap" метода. В параграф 2.3 е представено валидиране на разработения алгоритъм, резултатите от което са представени в Таблица 3.1 и може да се види, че средните на оценките на параметрите са равни на истинските стойности с точност до стотни. В следващия параграф разработеният "probit" модел е приложен към реални данни от Здравно и Пенсионно проучване в САЩ, което показва, че има силна корелация между наблюденията над един индивид, както и че има различия в самоценките на здравето на различните индивиди.

В **Глава 4** е разгледан съвместен "probit" модел за p наредени категорни променливи. За да бъде отчетена корелацията между категорните величини е предложен модел със случайни ефекти за латентните нормални величини, които се предполага, че ги генерират. Резултатите в тази глава са естествено обобщение на тези от предишната. Отново се оказва възможно чрез въвеждане на разликите между съседните прагове намирането на лог–правдоподобието на пълния набор от данни в явен вид и оттам получаване на затворени форми за оценките на неизвестните параметри. С помощта на явния вид на корелационната матрица на случайните ефекти, системата уравнения за регресионните параметри на фиксираните ефекти и квадратните уравнения за разликите между последователните прагове, оценките се обновяват на всяка стъпка от алгоритъма, при което в изчисляването на новите оценки участват условните очаквания на горе–описаните величини, при условие наблюдаваните данни. В параграф 3.2.3 е доказано, че тези условни очаквания зависят само от първите два момента на p–мерно орязано нормално разпределение. За реализиране на алгоритъма отново е използвана безплатната софтуерна среда за статистически изчисления R като е използван подхода на Manjunath and Wilhelm (2009) с аналитични форми, т.к. е достатъчно бърз и осигурява детерминираност на всяка стъпка на алгоритъма за разлика от Монте Карло метода използван в алгоритъма от предишната глава. За валидиране на разработения алгоритъм в тази глава са симулирани стойности за две категорни величини с по едно наблюдение и случайни ефекти. Направени са две симулационни изследвания като при това с по–малкия брой на обектите в извадката 100 получените оценки се оказват изместени, което има своето обяснение, че МПО са асимптотично неизместени и всъщност се съгласува с резултатите по втората извадка. Също така се потвърждава намаляване на стандартната грешка с увеличаване броя на наблюденията. Този модел е приложен върху реални данни от изследвания у нас в периода 2006–2008 година на зависимостите между страничните реакции и генетичните характеристики при жени с

рак на матката, лекувани чрез лъчетерапия. Предложеният корелиран "probit" модел за кожните и урогениталните реакции дава възможност да се направи ново откритие за положителна зависимост между двата типа странични реакции.

В **Глава 5** е разработен корелиран "probit" модел за една наредена категорна и една нормална променливи от дългосрочни проучвания. Предполагаме, че зад категорната сл.в. има нормална, която я генерира. Резултатите в тази глава представляват обобщение на получените в предходните две в следния смисъл:

- сл.в. са многомерни, което е отражение на дългосрочните наблюдения;
- отчита се зависимостта между наблюденията чрез случайните ефекти;
- отчитат се особеностите на отделните индивиди.

Един от приносите на разработената в тази глава методология е, че тя има предимства пред останалите такива за намиране на МПО поради това, че получените оценки са асимптотично неизместени и изчисленията не нарастват експоненциално с нарастване размерността на случайните ефекти.

Забележки Забележки по същество нямам.

За оформлението и изложението

Би било по–добре да има номерация на повтарящите се формули, напр. стр. 22, р. 2 отдолу, р. 12 отдолу. Навсякъде терминът "се схожда" е добре да бъде заменен с "е сходящ" или "клони".

Тези грешки не са по същество, не намаляват яснотата на резултатите и не предизвикват съмнение относно верността им.

Литература Цитираните литературни източници показват, че дисертантката е добре запозната със статистическите методи за оценяване на параметри при дългосрочни наблюдения, както и с алгоритмите за тяхата имплементация. Заедно с това, от получените в дисертацията резултати се вижда, че умело и 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 дисертантката приноси действително са такива.

Публикациите свързани с дисертацията са 3 на брой, една от тях е в Pliska Studia Mathematica Bulgarica, една в Serdica Journal of Computing и една в сборник на Докторантската конференция MIE'2013. Всички публикации са в съавторство с научния ръководител. Доказателство за високото научно ниво на получените в дисертацията резултати е присъдената на дисертантката награда за учен на 34-та международна конференция по клинична биостатистика, 2013.

От разговорите със съавторката имам основание да смятам, че приносите на дисертантката са не по–малки от приносите на другия съавтор.

Личните ми впечатления за дисертантката са положителни. Те са главно от участията ѝ в Международните конференция по Вероятности и Статистика (2010, 2012), Пролетната конференция на СМБ и XVI–та Европейската среща на младите статистици, от доклади пред Националния семинар по Стохастика към ИМИ–БАН, както и от преките ни служебни контакти от постъпването ѝ в магистърската програма "Вероятности и статистика" през 2009 досега. Отнася се с подчертана задълбоченост и прецизност при решаване на поставените проблеми.

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 Деница Парашкевова Григорова образователно–научната степен "доктор"** в областта на висше образование "Природни науки, математика и информатика", професионално направление " Математика".

Дата: 24. 03. 2014 г. Марусия Божкова

Подпис:.....