

РЕЦЕНЗИЯ

за дисертацията на Нина Руменова Даскалова
на тема:

**”Алгоритми от тип ЕМ за статистическо оценяване в
разклоняващи се стохастични процеси”**

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 01.01.10 ”Теория на вероятностите
и математическа статистика”.

Рецензент: доц. д-р Марусия Никифорова Божкова,
р-л катедра ”Вероятности, операционни изследвания и статистика”
(ВОИС), ФМИ, СУ ”Св. Климент Охридски”
и секция ”Вероятности и статистика” на ИМИ–БАН

Представям рецензията си по тази защита като член на Научното
жури, определено със Заповед No 150/ 15. 05. 2012 г. на Директора на
ИМИ–БАН съгласно Решение на НС на ИМИ–БАН (Протокол No 6/ 11.
05. 2012 г.). Рецензията е изготвена според изискванията на:

- Закона за развитието на академичния състав в Република Бълга-
рия (ЗРАСРБ),
- Правилника за прилагане на ЗРАСРБ,
- Правилника за условията и реда за придобиване на научни степени
и заемане на научни длъжности в БАН,
- Правилника за условията и реда за придобиване на научни степени
и за заемане на академични длъжности в Института по математика
и информатика на БАН.

Биографична справка. Нина Руменова Даскалова е родена на 20. 04.
1978 г.

Завършва висше образование, специалност ”Информатика” през 2002
г., а по-късно специализира ”Приложна статистика” към катедра ВОИС
и завършва със защита на дипломна работа на тема: ”Статистически
анализ на взаимодействието на линкерни хистони”.

Задочен докторант е в ИМИ–БАН към секция ”Вероятности и статис-
тика” по специалността 01.01.10 Теория на вероятностите и математи-
ческа статистика от 01.01.2006 до 30.06.2010 с научен ръководител проф.
д-мн Николай Янев. В периода 15.03.2004 до 14.03.2008 работи като ма-
тематик към секция ”Вероятности и статистика” на ИМИ–БАН, а от

15.03.2008 постъпва на работа като асистент по Теория на вероятностите и математическа статистика към СУ, ФМИ, катедра "Вероятности, операционни изследвания и статистика" (ВОИС), където работи и в момента.

Общо описание на дисертацията. Дисертацията се състои от увод, четири глави, заключение, списък с литература, съдържащ 159 заглавия и програмни приложения. Общият обем на дисертацията е 89 страници.

Анализ на научните и научно-приложните постижения в дисертационния труд.

С цел да се построят максимално-правдоподобни оценки (МПО) за вероятностите на разпределението на потомството по непълни данни за многотипов разклоняващ се процес (МТРП), в дисертацията се решават следните задачи:

а) да се намери подходящо представяне на МТРП чрез стохастични контекстно-свободни граматики (СКСГ) и да се приложи съществуващ за тях алгоритъм от тип ЕМ за МТРП;

б) да се дефинира изцяло нов ЕМ алгоритъм, конкретно за разклоняващи се процеси;

в) двата подхода да се реализират софтуерно, да се изследва тяхното поведение и емпиричните свойства на получените оценки.

Като напълно завършващ изследването е разгледан пример с реални данни за моделиране на клетъчна пролиферация, в който е уместно прилагането на разработения метод.

Уместно е да отбележим на това място, че разклоняващите се стохастични процеси се свързват напълно естествено, на първо място с редица приложения в биологията, медицината, биологичното разнообразие на видовете, но всъщност развитието на теорията от далечната 1845 г. досега показва, че те са също подходящи модели на явления във физиката, химията, демографията, икономиката, информатиката, финансовите активи и т.п. Изобщо, съвременната теория на разклоняващите се процеси може да се прилага при всяка идеализирана популация, чиито елементи генерират/пораждат нови множества от елементи.

Тези елементи (или както е прието да се наричат в литературата, частици или индивиди) могат да имат най-разнообразна природа, например, елементарни частици, атоми, молекули, микроорганизми, клетки, живи организми, информация, парични средства и т.н., които в съответствие с някакви вероятностни закони се превръщат в обекти от същия вид.

Най-простият и най-добре изучен модел на РП е с един тип частици и дискретно време, или както е прието да се нарича процес на Биенеме-Галтон-Уотсън (БГУ), по имената на френския математик Би-

енеме, който първи разглежда въпроса за израждане на фамилиите и на двамата английски математици Галтон и Уотсън, които изследват проблема за съдбата на аристократичните фамилии в Англия.

Всъщност терминът "разклоняващи се процеси" е предложен от А. Н. Колмогоров през 40-те години на XX-тия век, и оттогава датират първите работи по РП в съавторство с неговите ученици Севастьянов, Дмитриев, Яглом и др.

Доказателство за бурното развитие на теорията на РП и нейните многобройни приложения са редица монографии, сред които са класическите на Харис (1963), Севастьянов (1971), Мод (1971), Атрея и Ней (1972), Ягерс (1975), Асмусен и Херинг (1983), Гуторп (1991). Интересни съвременни приложения на разклоняващите се процеси в биологията и клетъчната пролиферация са съдържание на монографиите на Яковлев и Янев (1989), Кимел и Акселрод (2002), Хаку, Ягерс и Ватутин (2005). Българската школа в областта на РП, създадена от проф. Н. Янев, един от учениците на Севастьянов, има своето достойно място и се ползва с висок авторитет със своите изследвания в областта на регенериращите, регулируемите и индексирани РП, както и с приносите в статистиката на РП и приложенията и в клетъчната пролиферация, епидемиологията и други области.

От друга страна, съвременната статистика разчита на повишения капацитет на компютрите и на тази база развива методи, които в миналото са били пренебрегвани заради голямата си изчислителна сложност. Един от тези методи, който в последните три десетилетия се развива и доказва като статистическа практика е ЕМ (Expectation Maximization) алгоритъмът, наречен така от Демпстър, Леърд и Рубин (Dempster et al. [19]) през 1977 г. Това е широко приложим алгоритъм, който предлага итеративна процедура за намиране на максимално-правдоподобна оценка (МПО) за статистически модели с липсващи данни, в които такава оценка би била директно изчислима при наличието на тези допълнителни данни.

Уводът на дисертацията е посветен на мотивацията за разработване на ЕМ-алгоритми за многотипови разклоняващи се процеси (МТРП), целите и задачите, които се решават в дисертацията.

В Глава 1 е направен преглед на литературата и съвременното състояние на изследванията, съответно в областта на статистиката на РП от една страна и на ЕМ алгоритмите от друга. Прави силно впечатление голямата брой-159 научни публикации включени в обзора на резултатите и точното им съотнасяне към проблематиката на дисертацията. Също така в параграф 1.2 са приведени в необходимия обем класически понятия и резултати от теорията на разклоняващите се процеси (РП), стохастичните контекстно-свободни граматики (СКСГ) и ЕМ алгоритмите, които

се използват по-нататък в дисертацията.

Глава 2 е посветена на съответствието между МТРП и СКСГ. Връзката между МТРП и СКСГ е установена още с работите на Sankoff [22] и на Geman and Johnson [24], в които се използват резултати от теорията на РП и се прилагат в теоретичните изследвания на граматиките. В дисертацията си Н. Даскалова адресира обратния (в определен смисъл) въпрос: Дали е уместно и до колко да се използват разработените в лингвистичен контекст ЕМ-алгоритми за оценка параметрите на разпределението на потомството в подходящо дефиниран РП? В параграф 2.1 е представен т. нар. "inside-outside" алгоритъм, разработен от Lari and Young [149, 150] за оценяване параметрите на СКСГ, който стои в основата на идеята за построение на съответната СКСГ на даден РП. В следващия параграф 2.2 е извършено самото построение на СКСГ за определен клас МТРП и се оказва, че дори за МТРП с два типа, поради големия брой типове не-терминали в съответната граматика, сложността на алгоритъма нараства, така че затруднява пресмятанията. Това е и основната причина да се търси начин за директно разработване на ЕМ-алгоритъм за оценяване на индивидуалните вероятности в МТРП. Последният параграф от тази глава е посветен на два примера, които илюстрират горе-описания подход. Първият е за МТРП с три типа частици, където третият тип е терминален. Уместно е да отбележа, че поради особеностите на моделирането, за МТРП е въведено понятието терминален тип, което означава, че частиците/индивидите от такъв тип не се репродуцират, но и не умират в смисъл на израждане в термините на разклоняващи се процес. Вторият пример е симулационен експеримент, основан на известен модел за развитието на олигодендрцити.

Глава 3 съдържа основните резултати получени в дисертацията. В нея се описва МТРП с терминални типове, за който всъщност се оказва възможно намирането на ЕМ-оценка на индивидуалните вероятности само по наблюдение върху състоянието на процеса в конкретен момент. В параграф 3.1 е дефиниран МТРП с терминални типове и в Твърдение 3.1.1., по аналогия на еднотиповия случай е получена МПО за индивидуалните вероятности на МТРП по пълно наблюдение над цялото дърво, съответно на процеса. В параграф 3.2.1 е изведен общия вид на ЕМ-алгоритъма и в Твърдение 3.2.1. е изчислена аналитично М-стъпката. В параграф 3.2.2. е предложена рекурентна схема за изчисление на очакванията в Е-стъпката. В параграф 3.3 е извършена верификация на алгоритъма и е илюстрирано получаването на търсените оценки. Разглежда се МТРП с четири типа частици - 2 терминални и 2 не-терминални типа и конкретни репродукции. Пресметнати са външните и вътрешни вероятности и са получени оценките на ненулевите вероятности за възможните

преходи. Особено интересен е примера с бинарните дървета, при който оценките за индивидуалните вероятности се получават в явен вид, но за да се верифицира алгоритъма е създаден код на R и се оказва, че за всички проверени случаи, резултатите от алгоритъма съвпадат с вече получените директно. Накрая в параграф 3.3.3. чрез симулационен експеримент с различна големина на извадките и различни стойности на параметрите е изследвано поведението на оценките. Като цяло оценките се оказват състоятелни, а за критичен процес са по-нестабилни, отколкото за до- и надкритичен.

Глава 4 е посветена на един конкретен модел на клетъчна пролиферация, за който са налични експериментални данни. Същественото в случая е приложението на описания алгоритъм към РП с непрекъснато време, който се използва естествено за описание на проведения лабораторен експеримент. ЕМ-оценките в този случай се правят върху вложения процес на поколенията.

Забележки. Забележки по същество нямам.

За оформлението и изложението.

Има някои дребни печатни грешки, като например: на стр. 27, р. 2 е пропуснат знака за условие, на стр. 28, р. 7 би било по-ясно, ако някъде се уточни, че неравенството е за всяко θ и други по-незначителни. Тези грешки не са по същество, не намаляват яснотата на резултатите и не предизвикват съмнение относно верността им.

Независимо от тези забележки дисертацията е много добре структурирана, изложението е прецизно и ясно, приведени са в необходимия обем класически резултати от теорията на разклоняващите се процеси, както и от областта на ЕМ алгоритмите. Това дава възможност да се види мястото на направеното в дисертацията в сравнение с известните преди това резултати. Доказателствата са проведени изцяло и не предизвикват съмнение.

Литература. Цитираните литературни източници показват, че дисертантката е добре запозната с теорията и статистиката на разклоняващите се процеси и има отлични умения за програмиране на езика R, специално разработен за статистически приложения. Заедно с това, от получените в дисертацията резултати се вижда, че умело и оригинално използва тези знания за решаването на нови задачи.

Авторефератът на дисертацията е в обем от 22 стр. и отразява пълно и точно получените в дисертацията резултати, както и мястото и връзката им с резултатите на други автори.

Считам, че заявените от дисертантката приноси действително са такива.

Публикациите свързани с дисертацията са 3 на брой, една от тях е в Плиска: [26], една приета за печат в Доклади на БАН: [27] и една в Serdica Journal of Computing: [25]. Всички публикации са самостоятелни.

Личните ми впечатления за дисертанта са отлични. Те са главно от участията в XII-та, XIII-та, XIV-тата и XV-та Международната конференция по Вероятности и Статистика, както и в Първи и Втори Национален семинар по проблемите на образованието, Китен (2007 и 2009 г.), от доклади пред Националния семинар по "Вероятности и Статистика" с р-л проф. дмн Н. Янев, както и от преките ни контакти и взаимоотношения в катедра ВОИС. Нина Даскалова показва сериозно и отговорно отношение към преподавателската си работа, съчетани с прецизност и възискателност, постоянно търсене и предлагане на нови методи за подобряване качеството на образователния процес. Тя е един от съорганизаторите на Статистическа лаборатория към катедрата. Също така се включва активно по различните проекти – 4 на брой, като показва инициативност, бързина и качество в изпълнението на поставените и задачи. Ще спомена, че тя е разработила нови или актуализирала, вече съществуващи упражнения към задължителните дисциплини: по Теория на вероятностите и Математическа статистика (специалност Информатика), Вероятности и статистика (Компютърни науки), Статистика и емпирични методи (семинарни и практикум, за Софтуерно инженерство и Информационни системи), както е и разработила нова дисциплина Статистика с R (Статистика), Статистика с R – практикум / Линейни модели с R (МП Вероятности и статистика) и Приложна статистика (МП Био и медицинска информатика).

Заклучение.

Въз основа на всичко изложено до тук считам, че представеният дисертационен труд отговаря на всички изисквания на ЗРАСРБ, неговия Правилник и Правилниците за прилагане на ЗРАСРБ на БАН и ИМИ, затова убедено препоръчвам на почитаемия Научен Съвет на Института по математика и информатика да присъди на Нина Руменова Даскалова образователната и научна степен "Доктор" 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".

Дата: 17.08.2012г.

Марусия Божкова
Подпис: