БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
СЕКЦИЯ „ВЕРОЯТНОСТИ И СТАТИСТИКА“

Нина Руменова Даскалова

АЛГОРИТМИ ОТ ТИП ЕМ ЗА СТАТИСТИЧЕСКО ОЦЕНЯВАНЕ В РАЗКЛОНЯВАЩИ СЕ СТОХАСТИЧНИ ПРОЦЕСИ

АВТОРЕФЕРАТ

на дисертация
за присъждане на образователна и научна степен
„Доктор“

Научен ръководител:
проф. дмн Николай Михайлов Янев

София, 2012

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
СЕКЦИЯ „ВЕРОЯТНОСТИ И СТАТИСТИКА“

Нина Руменова Даскалова

АЛГОРИТМИ ОТ ТИП ЕМ ЗА СТАТИСТИЧЕСКО ОЦЕНЯВАНЕ В РАЗКЛОНЯВАЩИ СЕ СТОХАСТИЧНИ ПРОЦЕСИ

АВТОРЕФЕРАТ

на дисертация
за присъждане на образователна и научна степен
„Доктор“

Професионално направление: *4.5 Математика*
Научна специалност: *Теория на вероятностите и математическа статистика (01.01.10)*

Научен ръководител:
проф. дмн Николай Михайлов Янев

София, 2012

Данни за дисертационния труд:
Обем на дисертацията: 89 стр.
Основен текст: 69 стр.
Литература: 159 заглавия.
Публикации по дисертацията: 3 броя.

Дисертационният труд е обсъден и препоръчан за започване на процедура по защита на разширено заседание на секция „Вероятности и статистика“ при Института по математика и информатика – БАН, проведено на 18.04.2012 г.

Материалите по защитата са на разположение на интересуващите се в секция „Вероятности и статистика“ и в библиотеките на ИМИ – БАН и на ФМИ – СУ.

Автор: *Нина Руменова Даскалова*
Заглавие: *Алгоритми от тип EM за статистическо оценяване в разклоняващи се стохастични процеси.*

Мотивация, цели и задачи на дисертацията

Разклоняващите се стохастични процеси моделират явления в различни области на научното познание, където се наблюдава еволюцията (размножаване или преобразуване) на различни обекти. Това могат да бъдат определени частици (атоми, молекули, фотони и др.) във физиката или химията; индивиди (хора, животни, растения, микроорганизми, клетки) в биологията, епидемиологията или демографията; ДНК или гени в молекулярната биология и генетиката, дори алгоритми и програми в компютърните науки. Добре развитият математически апарат е основа за разработването на нови приложения на разклоняващите се процеси във все по-широк кръг от области, като в същото време тези приложения създават предизвикателства за по-нататъшното развитие на теоретичните математически основи.

Първите документирани сведения за научен интерес в областта датират от средата на 19-ти век. За основополагаща се смята работата на Галтон и Уотсън (Galton and Watson [1]), в която се разглежда проблемът за изчезването на аристократичните фамилии. Този модел е известен под името процес на Биенеме-Галтон-Уотсън (БГУ), тъй като (макар и доста по-късно – в края на 20-ти век) е открито, че подобни проблеми са разглеждани и в работата на Биенеме (Bienaymé [2]) от 1845 г. Според модела на БГУ индивидите живеят единица време и в края на живота си оставят потомство, съобразно някакво вероятностно разпределение. Първите асимптотични резултати в тази област са получени от Колмогоров [3] през 1938 г., а за първи път терминът „разклоняващи се процеси“ е въведен в работата на А. Н. Колмогоров и Н. А. Дмитриев [4] от 1947 г. Белман и Харис (Bellman and Harris [5]) обобщават процеса на БГУ, като въвеждат случайно време на живот на индивидите, в края на който те дават потомство според (неизменящ се) случаен закон. В своя модел, Севастьянов [6] въвежда и зависимост на възпроизводството от времето на живот. Най-общия модел на разклоняващ се процес е въведен независимо от Кръмп, Мод и Ягерс (Crump and Mode [7, 8] и Jagers [9]), като в него индивидите имат случайно време на живот, в произволна зависимост със закона за възпроизвеждане, а потомството се поражда в рамките на точков процес през времето на живот на индивида. На базата на тези основни модели се разработват допълнителни възможности като например въвеждането на два пола, различни видове миграция и други източници на контрол в популацията.

Друга насока за обогатяване на моделите е разглеждането на възможността частиците на процеса да имат различни типове. За първи път многотипови процеси се срещат в работите на А. Н. Колмогоров и Н. А. Дмитриев [4] и А. Н. Колмогоров и Б. А. Севастьянов [10]. Процеси на Белман-Харис с повече от един тип частици са разгледани от Ней (Ney [11]) и Севастьянов [6]. Мод (Mode [12]) въвежда в общия процес и възможност за различни типове.

Подробно изложение на основните теоретични и приложни аспекти на разклоняващите се стохастични процеси могат да бъдат намерени в следните издания: Harris [13], Севастьянов [14], Mode [12], Athreya and Ney [15], Jagers [9], Asmussen and Hering [16], както и в излязлата наскоро на български език книга на М. Славчова-Божкова и Н. Янев [17]. Повече за историята на разклоняващите се процеси може да се види в есето на П. Ягерс (Jagers [18]).

С развитието на съвременните компютърни технологии стана възможно събирането и съхраняването на големи количества данни. Това обуславя и интензивното развитие на изчислителните методи за обработка на данни, в това число и статистически. 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 разви и доказа като статистическа практика е EM (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. Повече за теорията и приложенията на EM алгоритъма може да се види в книгата на McLachlan and Krishnan [20].

При вероятностното моделиране често се срещат случаи, когато наблюдаваните величини могат да се разглеждат като резултат от ненаблюдаем модел. Този модел може да бъде и разклоняващ се стохастичен процес. Пример за това са биологични експерименти, изследващи клетъчна пролиферация, в които деленето започва от клетка от даден вид и в определен момент се преброяват клетките, породени от нея. Тук размножаването и диференциацията на клетките се описват от многотипов разклоняващ се

процес (МТРП), който не може да бъде наблюдаван в неговата цялост, а само като брой на обектите в даден момент. Това е типичен случай за приложение на ЕМ алгоритъм.

Целта на дисертационния труд е да се въведе по подходящ начин използването на ЕМ алгоритъм за намиране на МПО за разпределението на потомството в МТРП при непълни наблюдения. Да се изследва връзката (вж. Sankoff [21], Miller and O'Sullivan [22], Geman and Johnson [23]) на МТРП с други дискретни вероятности структури, наречени стохастични контекстно-свободни граматики (СКСГ), за които съществува подобен алгоритъм. Задачата е да се разработят два подхода към поставения проблем – единият използва представяне на МТРП чрез СКСГ и прилагане на съществуващия алгоритъм, а другият дефинира изцяло нов ЕМ алгоритъм, конкретно за разклоняващи се процеси. И двата подхода да се реализират софтуерно и да се изследват тяхното поведение и емпиричните свойства на получените оценки. Да се разгледа пример за моделиране на клетъчната пролиферация, в който е уместно прилагането на разработения метод.

Структура и съдържание на дисертацията

• В Глава 1 е направен кратък обзор на литературата в областта на разклоняващите се процеси (РП) и ЕМ-алгоритъма, който има за цел да покаже актуалното състояние на изследваната тематика и да мотивира основната задача на дисертацията, а именно въвеждането на ЕМ-алгоритъм за непараметрична оценка на индивидуалното разпределение в определен клас РП.

Накратко са изложени и основните дефиниции и теоретични резултати, които са използвани в дисертацията. В следващото изложение е спазвана номерацията на твърденията в текста на дисертацията, като е добавена буквата Д пред номера.

Разклоняващият се процес на Биенеме-Галтон-Уотсън (БГУ) може да се дефинира по следния конструктивен начин:

Определение Д. 1.2.1 *Нека на вероятностното пространство* $(\Omega, \mathcal{A}, P)$ *са дефинирани целочислени неотрицателни случайни величини (сл.в.)* $\xi_i(t)$, $i = 1, 2, \ldots$, $t = 0, 1, 2, \ldots$, *които, освен това, са независими и еднакво*

разпределени (н.е.р.)

$$P(\xi_i(t)=k)=p_k, \quad \sum_{k=0}^{\infty} p_k = 1, \quad F(s)=\sum_{k=0}^{\infty} p_k s^k. \tag{1}$$

Тогава процес на БГУ се дефинира чрез:

$$Z_{t+1} = \begin{cases} \sum\limits_{i=1}^{Z_t} \xi_i(t+1) & , \ Z_t > 0, \\ 0 & , \ Z_t = 0, \end{cases} \quad t=0,1,2,\ldots, \quad Z_0 = 1. \tag{2}$$

Ако параметърът t се интерпретира като номер на поколение, а състоянието на процеса като брой на частиците, то тогава $\xi_i(t+1)$ означава броя на преките потомци в $t+1$ поколение на $i-$та частица, съществуваща в $t-$то поколение, а Z_t е общия брой на частиците в $t-$то поколение. В такъв случай, основното свойство (2) изразява независимостта на еволюцията на частиците от общия брой частици, съществуващи в дадено поколение. То показва, че състоянието нула е поглъщащо за разклоняващия се процес, т. е. ако процесът веднъж попадне в нулата, то той остава там завинаги.

Вероятностите $p_k = P(Z_1 = k)$ се наричат *индивидуални вероятности*, а $F(s) = Es^{Z_1}$ – *индивидуална пораждаща функция*. Оказва се, че индивидуалните характеристики напълно определят всички останали характеристики на процеса Z_t и в частност *преходните вероятности*:

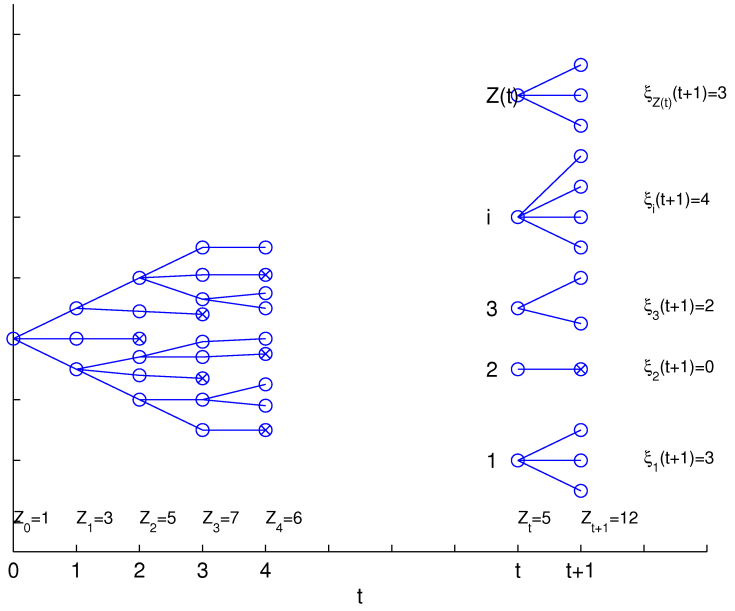
$$P_n(t) = P(Z_{t+\tau} = n | Z_\tau = 1) = P(Z_t = n)$$

и съответните пораждащи функции

$$F(t;s) = \sum_{n=0}^{\infty} P_n(t) s^n = Es^{Z_t}.$$

При фиксирано $\omega \in \Omega$ функцията $Z_t(\omega)$, $t \geq 0$ се нарича *траектория (реализация) на процеса*. Траекториите на разклоняващия се процес могат да бъдат представени като дърво (оттам идва и терминологията „разклоняващ се“ процес). Разклоняващият се процес $\{Z_t(\omega), \omega \in \Omega, t \geq 0\}$ можем да си представим като множеството от всички възможни дървета (фиг. 1), които се генерират от зададеното разпределение (1).

От гледна точка на теорията на марковските процеси не е съществено дали процесът е с един или с няколко типа частици, тъй като и в двата



Фигура 1: Фамилно дърво на разклоняващ се процес.

случая процесът има изброимо множество на състоянията. Разликата е, че докато в първия случай имаме естествено номериране на състоянията във фазовото пространство $\mathbf{N} = \{0, 1, 2, \dots\}$, то във втория по-общ случай ще означаваме състоянията с вектори $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbf{N}^n$ и това ще означава, че в популацията имаме α_1 частици от тип T_1, α_2 частици от тип T_2, $\dots \alpha_n$ частици от тип T_n. Разклоняващият се процес тогава ще дефинираме чрез преходните вероятности $P^{(i)}_{\boldsymbol{\alpha}}(t) = P^{(i)}_{(\alpha_1,\alpha_2,\dots,\alpha_n)}(t)$ за това, че една частица от тип T_i за време t поражда съвкупност от частици, съответна на вектора $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbf{N}^n$.

Може да се даде конструктивна дефиниция на процесите, както в едномерния случай.

Определение Д. 1.2.4 *Ако* $\mathbf{Z}(0) = \boldsymbol{\delta}^m$, $m = 1, 2, \dots, n$, *то за* $t = 0, 1, 2, \dots,$

$$Z_k(t+1) = \sum_{i=1}^{n} \sum_{j=1}^{Z_i(t)} \xi_k^{(i)}(t; j), k = 1, \dots, n, j = 1, 2, \dots$$

където $\xi_k^{(i)}(t, j)$, $k, i = 1, \dots, n$, $j = 1, 2, \dots$ *са независими случайни величини над едно и също вероятностно пространство* $(\Omega, \mathrm{A}, \mathbf{P})$.

Когато $i = 1, 2, \ldots$ е фиксирано, векторите $\boldsymbol{\xi}^{(i)}(t; j)$ са еднакво разпределени за $j = 1, 2, \ldots$, $t = 0, 1, 2, \ldots$. Ако интерпретираме параметъра t като номер на поколение, то $\xi_k^{(i)}(t; j)$ означава броя частици от тип k, породени за едно поколение от j-та частица от тип i, съществуваща в t-то поколение.

Да предположим, че можем да наблюдаваме траекторията на един процес на БГУ като едно дърво до даден момент t (вж. Фиг. 1). Това всъщност е най-пълната информация, която можем да имаме за един разклоняващ се процес. От това дърво лесно могат да се определят статистиките $Z_j(k)$, където $Z_j(k)$ означава броя на частиците в j-тото поколение, които имат точно k наследника в следващото поколение, $j = 0, 1, 2, \ldots, t$; $k = 0, 1, 2, \ldots$. Тогава за функцията на правдоподобие получаваме:

$$L_t(p_0, p_1, p_2, \ldots) = \prod_{k=0}^{\infty} p_k^{\sum_{j=0}^{t-1} Z_j(k)}, \tag{3}$$

където $p = \{p_0, p_1, p_2, \ldots\}$ е индивидуалното разпределение и сме взели предвид независимостта на еволюциите на частиците.

Максимално-правдоподобната оценка (МПО) за параметъра p лесно може да бъде намерена от (3) като се използва метода на Лагранж.

Твърдение Д. 1.2.2 *МПО за неизвестните параметри* p_k, $k = 0, 1, 2, \ldots$. *е:*

$$\hat{p}_k(t) = \frac{U_t(k)}{U_t}, \quad k = 0, 1, 2, \ldots \tag{4}$$

Разглеждането на различните техники за оценяване, използващи EM-подход за решаване на широк кръг от статистически задачи, дава основание да се предположи, че такъв метод може да бъде полезен и в ситуации, където е уместно моделиране с РП. В реални приложения на такива модели често се случва да не е възможно да се получи пълно наблюдение над траекторията (дървото) на процеса, а само частично, на състоянието на процеса в определен момент. Точно в такива случаи е необходимо създаването на нови статистически методи и процедури, които да дадат възможност за оценка при наличните експериментални данни.

Идеята зад EM алгоритъма е, че за намиране на МПО за непълния модел се използва функцията на правдоподобие на пълния модел, която в много случаи има добри математически свойства. Нека параметрите на модела означим с вектора θ, x е векторът на наблюдението, а Y са някакви „скрити“

данни, които определят вероятностното разпределение на x. Тогава, ако плътността на „пълното“ наблюдение означим с $f(x, y|\theta)$, то плътността на „непълното“ наблюдение е маргиналната плътност $g(x|\theta) = \int f(x, y|\theta)dy$. Записваме правдоподобията и условната плътност на Y при условие x и θ по следния начин:

$$L(\theta|x, y) = f(x, y|\theta), \quad L(\theta|x) = g(x|\theta), \quad k(y|\theta, x) = \frac{f(x, y\theta)}{g(x|\theta)}.$$

Целта е да се максимизира логаритъма от правдоподобието

$$\log L(\theta|x) = \log L(\theta|x, y) - \log k(y|\theta, x). \tag{5}$$

Нека означим с $E_{\theta'}$ условното очакване на Y при условие че $X = x$ и θ' е текущата стойност на параметъра, т.е. спрямо условната плътност $k(y|\theta', x)$. Тогава, вземайки очакваните стойности от двете страни на равенство (5), ще получим следния резултат:

$$\log L(\theta|x) = E_{\theta'}[\log L(\theta|x, Y)] - E_{\theta'}[\log k(Y|\theta, x)]. \tag{6}$$

Означавайки

$$Q(\theta|\theta') = E_{\theta'}[\log L(\theta|x, Y)] \quad \text{и} \quad H(\theta|\theta') = E_{\theta'}[\log k(Y|\theta, x)],$$

равенство (6) придобива следния вид:

$$\log L(\theta|x) = Q(\theta|\theta') - H(\theta|\theta'). \tag{7}$$

Нека означим с $\theta^{(i)}$ оценката на параметъра, получена на i-тата итерация. EM алгоритъмът обикновено се дефинира така:

Определение Д. 1.2.9 *EM алгоритъмът се състои от две стъпки:*
E-стъпка: Определяне на функцията $Q(\theta|\theta^{(i)}) = E_{\theta(i)}[\log L(\theta|x, Y)]$.
M-стъпка: Избиране на $\theta^{(i+1)}$ *като стойността, която максимизира* $Q(\theta|\theta^{(i)})$ *по отношение на* θ.
E-стъпката и M-стъпката се повтарят алтернативно докато разликата $L(\theta^{(i+1)}|x) - L(\theta^{(i)}|x)$ *стане по-малка от предварително избран праг.*

Демпстър, Леърд и Рубин показват, че на всяка стъпка непълното правдоподобие може само да нараства:

Твърдение Д. 1.2.3 *При горните означения, за редицата* $\{\theta^{(i)}\}$ *е изпълнено:*

$$\log L(\theta^{(i+1)}|x) \geq \log L(\theta^{(i)}|x).$$

Сходимостта на редицата от ЕМ оценките $\{\theta^{(i)}\}$ зависи от вида на функциите на правдоподобие $L(\theta|x)$ и на очакваното правдоподобие $Q(\theta|\theta^{(i)})$. Следното условие, дефинирано от Wu [24], гарантира сходимост към стационарна точка. Представяме го, както е цитирано в Casella and Berger [25].

Теорема Д. 1.2.5 *Ако очакваното логаритмувано правдоподобие на пълния модел* $E_{\theta'}[logL(\theta|x,Y)]$ *е непрекъсната функция едновременно по* θ *и по* θ'*, то всички гранични точки на ЕМ редицата* $\{\theta^{(i)}\}$ *са стационарни точки за* $L(\theta|x)$*, и* $L(\theta^{(i)}|x)$ *клони монотонно към* $L(\hat{\theta}|x)$ *за някоя стационарна точка* $\hat{\theta}$*.*

Като следствие от теоремата имаме, че когато условието е изпълнено и правдоподобието е унимодално, алгоритъмът схожда към глобален максимум. В случаите на мултимодално правдоподобие обаче, това не е гарантирано и е необходимо допълнително изследване на поведението на алгоритъма с различни начални стойности, за да се избегнат локални екстремуми.

Терминът „алгоритъм“ в статията на Демпстър, Леърд и Рубин е бил критикуван, защото всъщност не се предлага конкретна последователност от действия за реализацията Е- и М-стъпките. В този смисъл, понятието „ЕМ алгоритъм“ по-скоро се разбира като рамка за дефиниране на широк кръг от конкретни алгоритми и по-уместно е да се използва терминът „алгоритъм от тип ЕМ“.

Формалните граматики са добре развит инструмент за моделиране на символни низове, използван в компютърната/изчислителната лингвистика. В т. нар. „йерархия на Чомски“ граматиките са подредени по сложност, като колкото по-високо се намират, толкова по-богат е езикът, който пораждат, но се усложнява и математическият формализъм, необходим за описанието им. Контекстно-свободните граматики (КСГ) се намират някъде по средата в тази йерархия. Те са сравнително прости за дефиниране и могат да бъдат достатъчно добре изследвани като математически модел. От друга страна са достатъчно сложни, за да дефинират както формални езици като компютърните, така и естествени езици, намирайки приложение включително и в геномиката за моделиране на „езика“ на ДНК например. За определени

модели е възникнала идеята, че към КСГ може да се приложи вероятностен подход и така се появяват стохастичните (или вероятностни) граматики (вж. например Sankoff [26] за използването им в лингвистиката).

Определение Д. 1.2.6 *Контекстно-свободна граматика (КСГ) наричаме четворката* $\Gamma = \langle \mathcal{N}, \mathcal{T}, S, \mathcal{R} \rangle$, *за която имаме множество от символи* $\{\mathcal{N} \cup \mathcal{T}\}$, $S \in \mathcal{N}$ *и множество от правила* $\mathcal{R}$. *Символите биват два вида – абстрактни не-терминали* ($\mathcal{N} = \{X, Y, Z, \ldots\}$), S *е специален „начален“ не-терминал, и терминали* ($\mathcal{T} = \{a, b, c, \ldots\}$). *Правилата* $\mathcal{R}$ *имат вида* $\alpha \to \beta$, *където* $\alpha \in \mathcal{N}$ *се състои от един не-терминал, а* $\beta \in (\mathcal{N} \cup \mathcal{T})^+$ *може да съдържа произволна поредица от не-терминали и терминали.*

Можем да интерпретираме терминалните символи като думите в едно изречение, т.е. това са реално наблюдаваните обекти, докато не-терминалите отговарят на граматичните категории и в този смисъл изразяват структурата на изречението. Правилата показват възможните трансформации между категориите и думите и осигуряват граматичната „правилност“ на изречението. Всяко изречение, получено чрез правилата на дадена граматика се нарича „извод“ на тази граматика. Изводът на една КСГ може да се представи чрез граф-дърво с корен S. В такъв граф всеки връх, който няма наследници, ще наричаме „лист“.

Определение Д. 1.2.7 *За дадена КСГ* $\Gamma = \langle \mathcal{N}, \mathcal{T}, S, \mathcal{P} \rangle$, *се определя дърво на извод* $D(V, E)$ *с корен* $r \in V$ *и функция* $f : V \to (\mathcal{N} \cup \mathcal{T})$, *съпоставяща на всеки връх символ от граматиката, така че:*
– *за всеки връх* v *на* D, *който не е лист,* $f(v) \in \mathcal{N}$, *като* $f(r) = S$;
– *за всеки лист* l *на* D, $f(l) \in \mathcal{T}$;
– *за всеки връх* v *(който не е лист) с наследници* $v_{i_1}, v_{i_2}, \ldots v_{i_k}$, *е в сила:* $f(v) \to f(v_{i_1}) f(v_{i_2}) \ldots f(v_{i_k}) \in \mathcal{P}$.

Стохастичната контекстно-свободна граматика (СКСГ) се получава, когато добавим вероятностно разпределение към КСГ (вж. Sankoff [26]). За всеки не-терминал X, нека $r_1, r_2 \ldots r_k$ са правилата на граматиката, в които X участва от лявата страна. На правилото r_i съпоставяме вероятност p_i, така че $p_1 + p_2 + \ldots + p_k = 1$ и така за всеки нетерминален символ. По този начин се дефинира вероятност на всяко дърво на извод, а именно произведението от вероятностите на всички правила, използвани за конструирането на това дърво. Това се основава на независимостта на избора на

всяко правило, използвано за всяко участие на всеки не-терминал. Вероятността на дадено изречение от езика, породен от граматиката, е сумата от вероятностите на всички негови дървета на извод. Така, можем да въведем следното определение:

Определение Д. 1.2.8 *Стохастична контекстно-свободна граматика (СКСГ) наричаме съвкупността* $\Gamma = \langle \mathcal{N}, \mathcal{T}, S, \mathcal{R}, \mathbf{P} \rangle$, *където първите четири елемента дефинират КСГ, а* $\mathbf{P}$ *е вероятностно разпределение върху правилата на граматиката, такова че на всяко правило* $r \in \mathcal{R}$ *съпоставяме вероятност* $p(r)$, *и сумата от вероятностите на всички правила, за които нетерминалът* X *участва от лявата им страна е равна на* 1*:* $\sum_{r:X\to\beta} p(r) = 1, \ \forall X.$

Изводът на дадено изречение се явява стохастичен процес, който отговаря на дефиницията на многотипов РП на БГУ. В него типовете частици са множествата на терминалите и не-терминалите $\{\mathcal{N} \cup \mathcal{T}\}$, а разпределението на потомството се задава от вероятностите $\{p(r)\}$. СКСГ задава определени ограничения върху съответстващия и МТРП. Едното от тях е, че изводът от граматиката винаги се състои само от терминални символи, които не търпят по-нататъшна еволюция. Всяко фамилно дърво (дървото на извод на граматиката) е крайно, а това определя РП като докритичен. Друго ограничение е, че изводът на граматиката най-често се разглежда като наредена последователност от символи, докато в теорията на РП се взема под внимание само съвкупността (множеството) от частици, без оглед на наредбата.

Многообразието от възможни правила в КСГ-и затруднява работата с тях на абстрактно ниво. За удобство се разглеждат определени трансформации на граматиките, наричани „нормални форми“. Казваме, че една КСГ е в нормална форма на Чомски (НФЧ), ако всички правила имат вида $X \to YZ$ или $X \to a$, където X, Y, Z са не-терминали, а a е терминален символ. Всяка КСГ може да бъде представена в НФЧ. Конкретният алгоритъм за това може да бъде намерен например в Манев [27] или Hopcroft et al. [28].За граматики в НФЧ съществува ЕМ алгоритъм, наречен „inside-outside“ (Lari and Young [29, 30]), чрез който се намира максимално-правдоподобна оценка на параметрите на граматиката, а именно вероятностите на отделните правила.

- Глава 2 разглежда връзката на многотиповите разклоняващи се про-

цеси (МТРП) със стохастичните контекстно-свободни граматики (СКСГ). Фактът, че СКСГ са по същността си МТРП, отдавна е забелязан и използван в теоретичните изследвания за граматиките (вж. напр. Sankoff [21], Geman and Johnson [23]). Тази връзка би могла да бъде полезна и в обратна посока, а именно, представяйки процеса чрез съответна граматика, да бъдат използвани техниките за оценяване параметрите на тази граматика, за да се оцени индивидуалното разпределение на процеса. Доколко това е възможно, в какви случаи е приложимо и какви ограничения възникват, е предмет на изложението в тази глава от дисертацията. Показано е как за даден МТРП може да бъде конструирана граматика, чиито параметри определят индивидуалните вероятности на процеса и е използван „inside-outside“ алгоритъма за оценяване в СКСГ, за да се получат оценки за МТРП. Изложените резултати са публикувани в статиите [31, 32].

• В Глава 3 са изложени основните резултати на дисертационния труд. Описан е модел на МТРП с терминални типове. Това е класът РП, в който е възможно да се направи ЕМ-оценка на индивидуалните вероятности, ако е наблюдавано само състоянието на процеса в даден момент, а самата траектория е „скрита“. Изведен е общият вид на ЕМ-алгоритъма в този случай и е изчислена аналитично М-стъпката. Предложена е рекурентна схема за изчисление на очакванията в Е-стъпката. Действието на предложения алгоритъм е демонстрирано с конкретен числен пример. Съществена част от резултатите е публикувана в [33].

Даден многотипов разклоняващ се процес (МТРП) може да се представи чрез вектор $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \ldots Z_d(t))$, където с $Z_k(t)$ означаваме броя обекти от тип T_k съществуващи в даден момент t, $k = 1, 2, \ldots d$. Индивидите от тип k имат поколение от различни типове съобразно d-мерно разпределение $p_k(x_1, x_2, \ldots, x_d)$ и всеки обект еволюира независимо от останалите. Ако $t = 0, 1, 2, \ldots$, то имаме процес на Биенеме-Галтон-Уотсън. Ако процесът е с непрекъснато време $t \in [0, \infty)$, ще дефинираме *вложен процес на поколенията* по следния начин (Athreya and Ney [15]).

Определение Д. 3.1.1 *Нека $\mathbf{Y}_n$ = броят обекти в n-тото поколение на $\mathbf{Z}(t)$. Ако вземем извадковото дърво π и го трансформираме в дървото π', което е идентично на π, но дължините на всичките му клони са равни на единица, то $\mathbf{Y}_n(\pi) = \mathbf{Z}_n(\pi')$, където $\mathbf{Z}_n$ е процес на Биенеме-Галтон-Уотсън. Ще наричаме $\mathbf{Y}_n$ вложен процес на поколенията на процеса $\mathbf{Z}(t)$.*

Дървото на процеса (за дискретно време) или това на вложения процес

(за непрекъснато време) ще бъде използвано, за оценяване разпределението на потомството.

Нека разгледаме МТРП, в който определени типове се определят като *терминални* и, веднъж създадени, обектите от такъв тип нито умират, нито се възпроизвеждат (вж. Sankoff [21]).

Определение Д. 3.1.2 *Нека за даден МТРП множеството от типове може да се раздели на два класа* $\mathbf{T} = \{T_1, T_2, \ldots, T_m\}$ *и* $\mathbf{T^T} = \{T_1^T, T_2^T, \ldots, T_{d-m}^T\}$*, така че обектите от тип* T_i *дава потомство от кой да е тип, а обектите от тип* T_j^T *не дава никакво потомство. Тогава типовете от първия клас ще наричаме не-терминални, а тези от втория – терминални типове. Процесът ще наричаме МТРП с терминални типове (МТРПТТ).*

Дефинирането на терминални типове възниква естествено при моделирането на различни видове популации. Така например в човешкия организъм има много видове клетки, които достигат определен етап в развитието си, когато спират да се делят или диференцират. С терминален тип можем да означим и „мъртвите“ обекти, като по този начин моделираме ситуация, в която те не изчезват, а се регистрират и присъстват като терминални в следващите поколения. В класическата теория на РП умрелите индивиди не участват в популацията – когато всички умрат, процесът се изражда, т.е. става нула. Това „изчезване“ на обектите след смъртта им затруднява статистическата оценка, защото по този начин се губи информация за цял клон от фамилното дърво. Това важи в най-силна степен точно когато разглеждаме дървовидната структура като „скрита“. В този случай, ако има такива изчезващи обекти, оценяването ще бъде невъзможно. Затова е необходимо да се използва модел, в който терминалните типове означават различни състояния на обектите, включително и тяхната смърт. Интересуваме се от оценка на вероятностите на разпределението на потомството. Следното твърдение е аналог на Твърдение (стр. 6) за многомерния случай.

Твърдение Д. 3.1.1 *Ако е наблюдавано цялото дърво* π*, то МПО на вероятностите на индивидуалното разпределение на МТРП е*

$$\hat{p}(T_v \to \mathcal{A}) = \frac{c(T_v \to \mathcal{A})}{c(T_v)}, \tag{8}$$

където $c(T_v)$ *означава броя на срещанията на обект от тип* T_v *в дървото* π, *а* $c(T_v \to \mathcal{A})$ *е броят пъти, в които обект от тип* T_v *дава потомство* $\mathcal{A}$ *в* π.

Не винаги обаче е възможно да се наблюдава цялото дърво, често имаме следната извадкова схема $\{\mathbf{Z}(0), \mathbf{Z}(t)\}$, за някое $t > 0$. Нека $\mathbf{Z}(0)$ се състои от един обект от някакъв тип. Предполагаме, че можем да наблюдаваме няколко независими реализации на процеса, започващи с идентични обекти и имащи един и същ закон за репродукция. Такава схема е често срещана в биологични експерименти. Ако t е дискретно, то $\mathbf{Z}(t)$ е броят обекти в t-тото поколение. За непрекъснато време това е поколение във вложения процес на Биенеме-Галтон-Уотсън. Тук предположението, че „мъртвите“ обекти не изчезват и могат да бъдат наблюдавани в следващите поколения като обекти от терминален тип, е съществено. Нека x е наблюдаваното множество от частици, π е не-наблюдаваната дървовидна структура, а с вектора $\mathbf{p}$ означим параметрите, които ще се оценяват – вероятностите на потомството $p(T_v \to \mathcal{A})$ (вероятността частица от тип T_v да породи множеството $\mathcal{A}$). Тогава правдоподобието на „пълното“ наблюдение ще бъде:

$$L(\mathbf{p}|\pi, x) = P(\pi, x|\mathbf{p}) = \prod_{\omega} \mathbf{p}(\omega)^{c(\omega;\pi,x)} = \prod_{v,\mathcal{A}:T_v \to \mathcal{A}} p(T_v \to \mathcal{A})^{c(T_v \to \mathcal{A};\pi,x)},$$

където c е брояща функция – $c(T_v \to \mathcal{A}; \pi, x)$ е броят пъти, когато частица от тип T_v поражда множеството от частици $\mathcal{A}$ в дървото π, ако е наблюдавано x. Правдоподобието на „непълното“ наблюдение съответно е равно на маргиналната вероятност $L(\mathbf{p}|x) = P(x|\mathbf{p}) = \sum_\pi P(\pi, x|\mathbf{p})$. Тя може да се намери аналитично, което е показано в следната

Теорема Д. 3.1.1 *М-стъпката на ЕМ-алгоритъма има следния вид:*

$$p^{(i+1)}(T_v \to \mathcal{A}) = \frac{E_{\mathbf{p}^{(i)}} c(T_v \to \mathcal{A})}{\sum_{\mathcal{A}} E_{\mathbf{p}^{(i)}} c(T_v \to \mathcal{A})} = \frac{E_{\mathbf{p}^{(i)}} c(T_v \to \mathcal{A})}{E_{\mathbf{p}^{(i)}} c(T_v)}, \tag{9}$$

където очакваният брой пъти частица от тип T_v *участва в дървото* π *е:*

$$E_{\mathbf{p}^{(i)}} c(T_v) = \sum_{\pi} P(\pi|x, \mathbf{p}^{(i)}) c(T_v; \pi, x),$$

а очакваният брой пъти частица от тип T_v *дава поколение* $\mathcal{A}$ *в дървото* π *е:*

$$E_{\mathbf{p}^{(i)}} c(T_v \to \mathcal{A}) = \sum_{\pi} P(\pi|x, \mathbf{p}^{(i)}) c(T_v \to \mathcal{A}; \pi, x).$$

В този случай M-стъпката се пресмята аналитично, така че изчислителните усилия ще бъдат върху E-стъпката на алгоритъма. Проблемът е, че в общия случай изброяването на всички възможни дървета π има експоненциална сложност, така че е необходим метод, който намира очаквания брой срещания на даден тип T_v или на дадена репродукция $T_v \to \mathcal{A}$, без да е необходимо да се обхождат всички възможни дървета. Това е възможно, благодарение на съществуващата рекурентна връзка между определени вероятности, която ще бъде описана по-долу.

Определение Д. 3.2.1 *Нека дефинираме* **вътрешната** *вероятност* $\alpha(\mathbf{I}, v)$ *на поддървото с връх* T_v *да произведе* $\mathbf{I} = \{i_1, i_2, \ldots, i_d\}$ *частици, където* i_k *е броят обекти от тип* k *(фиг. 2). От основното разклоняващо свойство на процеса имаме следната рекурентна връзка:*

$$\alpha(\mathbf{I}, v) = \sum_{\mathbf{w}} p(T_v \to \{T_{w_1}, \ldots, T_{w_k}\}) \sum_{\mathbf{I}_1 + \ldots + \mathbf{I}_k = \mathbf{I}} \alpha(\mathbf{I}_1, T_{w_1}) \ldots \alpha(\mathbf{I}_k, T_{w_k}),$$

където $\mathbf{w} = \{T_{w_1}, \ldots, T_{w_k}\}$ *са всички възможни множества от частици, които* T_v *може да породи.*

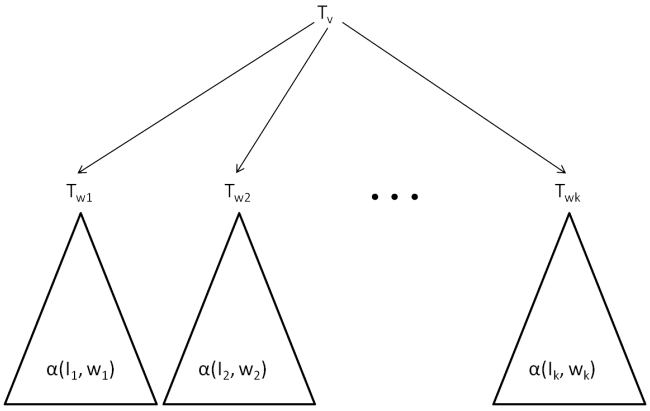
Определение Д. 3.2.2 Външната *вероятност* $\beta(\mathbf{I}, v)$ *дефинираме като вероятността на цялото дърво с изключение на поддърво с корен частица от тип* T_v*, пораждащо* $\mathbf{I} = \{i_1, i_2, \ldots, i_d\}$ *(фиг. 3). Рекурентната връзка тук е:*

$$\beta(\mathbf{I}, v) = \sum_{w} \sum_{\mathbf{v}} p(T_w \to \{T_v, T_{v_{(2)}}, \ldots, T_{v_{(m)}}\})$$
$$\times \sum_{\mathbf{J} \subset \mathbf{X} - \mathbf{I}} \beta(\mathbf{I} + \mathbf{J}, w) \sum_{\mathbf{J}_2 + \ldots + \mathbf{J}_m = \mathbf{J}} \alpha(\mathbf{J}_2, v_{(2)}) \ldots \alpha(\mathbf{J}_m, v_{(m)}),$$

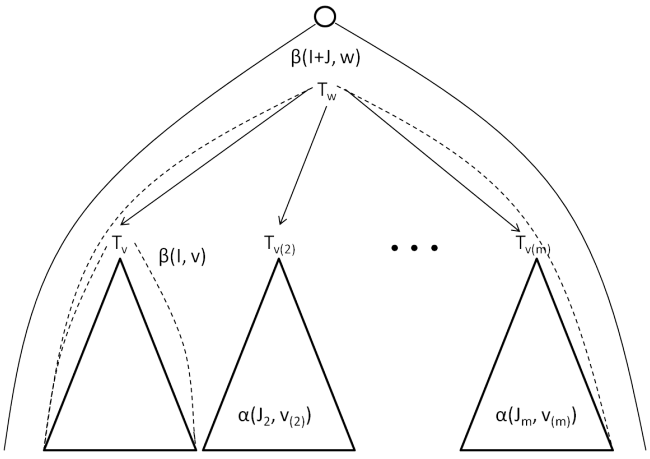
където $\{T_v, \mathbf{v}\} = \{T_v, T_{v_{(2)}}, \ldots, T_{v_{(m)}}\}$ *са всички възможни множества, които* T_w *може да породи и* $\mathbf{X} = \{x_1, x_2, \ldots, x_d\}$ *е наблюдаваното множество от частици.*

Така дефинираните вътрешни и външни вероятности, които могат да бъдат пресметнати рекурсивно, се оказват достатъчни са пресмятането на очакванията от E-стъпката на алгоритъма.

Теорема Д. 3.2.2 *Очакваните стойности на броя срещания на даден тип или дадена репродукция се изразяват чрез вътрешните и външните веро-*



Фигура 2: Рекурентната връзка за вътрешните вероятности.



Фигура 3: Рекурентната връзка за външните вероятности.

ятности по следния начин:

$$E_{\mathbf{p}^{(i)}} c(T_v) = \sum_{\mathbf{I}} \alpha(\mathbf{I}, v)\beta(\mathbf{I}, v), \tag{10}$$

$$E_{\mathbf{p}^{(i)}} c(T_v \to \{T_{w_1}, \ldots, T_{w_k}\}) \tag{11}$$
$$= \sum_{\mathbf{I}} \sum_{\mathbf{I}_1+\ldots+\mathbf{I}_k=\mathbf{I}} \beta(\mathbf{I}, v)\alpha(\mathbf{I}_1, w_1)\ldots\alpha(\mathbf{I}_k, w_k)p^{(i)}(T_v \to \{T_{w_1}, \ldots, T_{w_k}\}).$$

Така от теореми Д.3.1.1 и Д.3.2.2, окончателно получаваме, че итеративната преоценка на параметрите е:

$$p^{(i+1)}(T_v \to \{T_{w_1}, \ldots, T_{w_k}\})$$
$$= \frac{\sum_{\mathbf{I}} \sum_{\mathbf{I}_1+\ldots+\mathbf{I}_k=\mathbf{I}} \beta(\mathbf{I}, v)\alpha(\mathbf{I}_1, w_1)\ldots\alpha(\mathbf{I}_k, w_k)p^{(i)}(T_v \to \{T_{w_1}, \ldots, T_{w_k}\})}{\sum_{\mathbf{I}} \alpha(\mathbf{I}, v)\beta(\mathbf{I}, v)}.$$

Нека разгледаме процес с два типа частици T_1 и T_2, в който са позволени следните репродукции $T_1 \to T_2$ и $T_1 \to \{T_1, T_1\}$, съответно с вероятности p_2^1 и p_{11}^1. Такъв процес поражда бинарни дървета, такива че по листата им има 1 и 2, а вътрешните върхове са само 1-ци. Нека са наблюдавани m единици и n двойки, т.е. това са листата на дървото. Като се използва добре известния факт, че броят на вътрешните върхове в едно такова дърво е с едно по-малко от броя на листата, могат да бъдат получени оценките за индивидуалните вероятности директно (за една стъпка на алгоритъма) и те имат следния вид:

$$\hat{p}_2^1 = \frac{m}{2m+n-1}, \qquad \hat{p}_{11}^1 = \frac{m+n-1}{2m+n-1}. \tag{12}$$

Оказва се, че в този конкретен случай ЕМ-оценките, получени по наблюдението само върху листата на дървото съвпадат с МПО, които биха се получили, ако наблюдаваме кое да е пълно дърво на процеса. Този резултат би довел до опростяване на някои експерименти по наблюдението на такива процеси, тъй като няма да е необходимо пълното проследяване на процеса, за да се получат съответните МПО.

• Приложението на досега получените резултати за един модел на клетъчна пролиферация, за който са налични експериментални данни, е разгледано в Глава 4. Проведеният лабораторен експеримент е описан посредством модел на многотипов разклоняващ се процес с терминални типове.

Това е модел с непрекъснато време на Белман-Харис, за който са изведени уравнения за моментите. За получаване на ЕМ-оценка на индивидуалните вероятности е използван вложения процес на Биенеме-Галтон-Уотсън. Показани са резултатите за двата предложени модела.

• В Приложение 1 са дадени основните функции на R, използвани за получаване на изчислителните резултати в дисертацията, придружени от кратко описание. Приложение 2 съдържа една от извадките, получени чрез симулация и използвани в експеримента от Глава 3.

Апробация

Резултатите в дисертационния труд са докладвани на Националния семинар по теория на вероятностите и математическа статистика, на Пролетната научна сесия на ФМИ-СУ, както и на Workshop on Branching Processes and Applications (WBPA-2010) в рамките на XIV-th International Summer Conference on Probabilit and Statistics (ISCPS-2010, Sozopol, Bulgaria).

Списък от публикации по темата:

1. Daskalova, N. Using Inside-Outside Algorithm for Estimation of the Offspring Distribution in Multitype Branching Processes, *Serdica Journal of Computing*, Volume 4, Number 4, (2010), 463–474.

2. Daskalova, N. EM Estimation of the Offspring Distribution in Multitype Branching Processes — a Model in Cell Kinetics, *Pliska Stud. Math. Bulgar.*, Volume 20, (2011), 45–52.

3. Daskalova, N. Maximum Likelihood Estimation in Multitype Branching Processes with Terminal Types, *Compt. rend. Acad. bulg. Sci., (to appear)*, Volume 65, Number 5 (2012).

Авторска справка

Дисертационния труд е посветен на използването на ЕМ алгоритъм за намиране на МПО за разпределението на потомството в МТРП при непълни наблюдения. При изпълнението на поставената задача са постигнати резултати по отношение дефинирането на такъв алгоритъм, както и за приложенията му в конкретни случаи. Основните научни и научно приложни приноси могат да бъдат обобщени по следния начин:

- Направен е обзор на резултатите в областта на разклоняващите се процеси и EM-алгоритъма, който показва, че актуалното състояние на изследваната тематика и реалните задачи за моделиране от различни научно-изследователски области мотивират въвеждането на EM-алгоритъм за непараметрична оценка на индивидуалното разпределение в определен клас РП.

- Определена е постановка на задачата и извадкова схема, която разглежда оценяването в РП в терминологията на EM-рамката. Дефинирани са „пълно“ и „непълно“ наблюдение и съответните правдоподобия за поставения проблем.

- Разгледана е връзката на многотиповите разклоняващи се процеси (МТРП) със стохастичните контекстно-свободни граматики (СКСГ). Представяйки процеса чрез съответна граматика, е показано как могат да бъдат използвани техниките за оценяване параметрите на тази граматика, за да се оцени индивидуалното разпределение на процеса.

- Описан е модел на МТРП с терминални типове и е показано, че това е класът РП, в който е възможно да се направи EM-оценка на индивидуалните вероятности, ако е наблюдавано само състоянието на процеса в даден момент, а самата траектория е „скрита“. Изведен е общият вид на EM-алгоритъма в този случай и е изчислена аналитично M-стъпката. Предложена е рекурентна схема за изчисление на очакванията в E-стъпката. Действието на предложения алгоритъм е демонстрирано с конкретен числен пример.

- Предложеният метод е реализиран софтуерно на езика R за модел с два типа частици. Основните функции лесно могат да бъдат модифицирани и за други модели.

- Показано е, че за модели, в които има само един нетерминален тип и които се представят чрез двоични дървета, може да се намери EM-оценката в явен вид. В такива случаи EM-оценките съвпадат с оценките, получени от пълното наблюдение върху коя да е траектория на процеса. Този факт е използван за проверка на предложения рекурсивен алгоритъм, реализиран софтуерно.

- Получените резултати са приложени за един модел на клетъчна пролиферация върху експериментални данни. Проведеният лабораторен експеримент е описан посредством модел на многотипов разклоняващ се процес на Белман-Харис с терминални типове, за който са изведени уравнения за моментите. За получаване на ЕМ-оценка на индивидуалните вероятности е използван вложения процес на Биенеме-Галтон-Уотсън. Получени са оценките за двата предложени модела.

Благодарности

Авторът изказва своята благодарност на научния си ръководител проф. Николай Янев за подкрепата и ползотворните коментари по изследователските търсения и задачи в дисертацията. Благодаря още и на Марусия Божкова и Косто Митов за полезните критични бележки по част от работата по дисертацията, както и на Пламен Матеев, който имаше търпението да изслуша някои от идеите, преди още те да добият завършен вид. Благодарности и на всички колеги от секция Вероятности и статистика на ИМИ-БАН и катедра ВОИС на ФМИ-СУ за създадената благоприятна творческа атмосфера. И, не на последно място, искам да благодаря на семейството и на родителите си, за помощта, подкрепата и търпението през цялото време на работа по тази дисертация.

Библиография

[1] F. Galton and H. W. Watson. On the probability of the extinction of families. *J. Anthrop. Soc. London*, 4:138–144, 1875.

[2] I.J. Bienaymé. De la loi de multiplication et de la duree de familles. *Soc. Philomat. Paris Extrats. Ser.*, 1845.

[3] А. Н. Колмогоров. К решению одной биологической задачи. *Известия Томского университета*, 2:7–12, 1938.

[4] А. Н. Колмогоров и Н. А. Дмитриев. Ветвящиеся случайные процессы. *Докл. Акад. Наук СССР*, 56:7–10, 1947.

[5] R. Bellman and T. E. Harris. On the theory of age-dependant stochastic branching processes. *Proc. Nat. Acad. Sci.*, 34:601–604, 1948.

[6] Б. А. Севастьянов. Ветвящиеся процессы с превращениями зависящими от возраста частиц. *Теория Вероятн. и Примен.*, 9:577–594, 1964.

[7] K. S. Crump and C. J. Mode. A general age-dependent branching process I. *J. Math. Anal. Appl.*, 24:496–508, 1968.

[8] K. S. Crump and C. J. Mode. A general age-dependent branching process II. *J. Math. Anal. Appl.*, 25:8–17, 1968.

[9] P. Jagers. *Branching Processes with Biological Applications.* London, Wiley, 1975.

[10] А. Н. Колмогоров и Б. А. Севастьянов. Вычисление финальных вероятностей для ветвящихся случайных процессов. *Докл. Акад. Наук СССР*, 56:783–786, 1947.

[11] P. E. Ney. Generalized branching processes I and II. *J. Math.*, 8:316–350, 1964.

[12] C. J. Mode. *Multitype Branching Processes: Theory and Applications*. Elsevier, New York, 1971.

[13] T. E. Harris. *Branching Processes*. Springer, New York, 1963.

[14] Б. А. Севастьянов. *Ветвящиеся процессы*. Мир, Москва, 1971.

[15] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, Berlin, 1972.

[16] S. Asmussen and H. Hering. *Branching Processes*. Birkhauser, Boston, 1983.

[17] М. Славчова-Божкова и Н. Янев. *Разклоняващи се стохастични процеси*. УИ „Св. Климент Охридски", 2008.

[18] Peter Jagers. Some Notes on the History of Branching Processes, from my Perspective, 2009. URL `http://www.math.chalmers.se/ jagers/`.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[20] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 2008.

[21] D. Sankoff. Branching processes with terminal types: Application to context-free grammars. *Journal of Applied Probability*, 8(2):233–240, 1971.

[22] Michael I. Miller and Joseph A. O'Sullivan. Entropies and combinatorics of random branching processes and context-free languages. *IEEE Transactions on Information Theory*, 38, 1992.

[23] S. Geman and M. Johnson. Probability and statistics in computational linguistics, a brief review. *Mathematical foundations of speech and language processing. Johnson, M.; Khudanpur, S.P.; Ostendorf, M.; Rosenfeld, R. (Eds.)*, 2004.

[24] C. F. J. Wu. On the convergence of the EM algorithm. *Ann. Stat.*, 11: 95–103, 1983.

[25] G. Casella and R. L. Berger. *Statistical Inference, second edition*. Duxbury, 2002.

[26] D. Sankoff. Probability and linguistic variation. *Synthese*, 1978.

[27] К. Манев. *Увод в дискретната математика*. НБУ, София, 1996.

[28] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Pearson Addison-Wesley, 2007.

[29] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 4:35–36, 1990.

[30] K. Lari and S.J. Young. Applications of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 5(3): 237–257, 1991.

[31] N. Daskalova. Using inside-outside algorithm for estimation of the offspring distribution in multitype branching processes. *Serdica Journal of Computing*, 4(4):463–474, 2010.

[32] N. Daskalova. EM estimation of the offspring distribution in multitype branching processes — a model in cell kinetics. *Pliska Stud. Math. Bulgar.*, 20, 2011.

[33] N. Daskalova. Maximum likelihood estimation in multitype branching processes with terminal types. *Compt. rend. Acad. bulg. Sci., (to appear)*, 65(5), 2012.