PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **20** (2011), 45–52

PLISKA
STUDIA MATHEMATICA
BULGARICA

EM ESTIMATION OF THE OFFSPRING DISTRIBUTION IN MUTITYPE BRANCHING PROCESSES — A MODEL IN CELL KINETICS

Nina Daskalova[1] [2]

Multitype branching processes (MTBP) have been proven to be very useful models in cell kinetics. A typical example is the process of oligodendrocyte generation in cell culture, which is regarded as two-type branching process. Usually, such a process is not observable in the sense of the whole tree, but only as the "generation" at given moment in time, which consist of the number of cells of every type. An EM-type algorithm is used to obtain a maximum likelihood (ML) estimation of the offspring distribution. The performance of the presented algorithm is assessed using simulated data.

1. Introduction

Multitype branching processes (MTBP) are stochastic models in population dynamics, where particles are of different types. The theory and application of such processes could be found in a number of books [1, 2, 9, 14]. Statistical inference

[1]The research was partially supported by appropriated state funds for research allocated to Sofia University (contract 112/2010), Bulgaria.

[2]The author would like to express an appreciation to Nickolay Yanev for the meaningful and essential comments, which helped improving this work.

2000 *Mathematics Subject Classification:* 60J80, 60J85, 62P10, 92D25.

Key words: multitype branching processes, offspring distribution, maximum likelihood estimation, expectation maximization, stochastic context-free grammars, inside-outside algorithm, cell kinetics modelling.

in MTBP depends on the kind of observation available, whether the whole family tree has been observed, or only the particles existing at given moment t, or sometimes even the relative frequencies of types at that moment.

We consider a MTBP $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \dots Z_d(t))$, where $Z_k(t)$ denotes the number of particles of type T_k at time t, $k = 1, 2, \dots d$. Some estimators if the entire tree has been observed could be found in [8, 18], but usually we don't have such information about the process. Yakovlev and Yanev in [17] develop some statistical methods to obtain ML estimators for the offspring characteristics, based on observation on the relative frequencies of types at time t. Other approaches use simulation and Monte Carlo methods [7, 10, 11].

When the entire tree is not observed, but only the particles existing at given moment, an Expectation Maximization (EM) algorithm could be used, regarding the tree as the hidden data. Such algorithms exist for strictures, called Stochastic Context-free Grammars (SCFG). A number of sources point out the relation between MTBPs and SCFGs [6, 16].

We have proposed an approach to estimate offspring distribution probabilities in some MTBPs using the well developed methods for estimating parameters of SCFGs. The details are given in [3]. This approach is used here for the particular example of the process of oligodendrocyte generation in cell culture.

The paper is organized as follows. In Section 2 the algorithm is briefly explained. Section 3 defines the biological model. The algorithm was performed on simulated data and some results are shown in Section 4.

2. The Algorithm

The EM algorithm was explained and given its name in a paper by Dempster, Laird, and Rubin [4]. It is a method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. Let a statistical model is determined by parameters θ, x is the observation and Y is some "hidden" data, which determines the probability distribution of x. Then the joint probability of the "complete" observation is $P(x, Y|\theta)$ and the probability of the "incomplete" observation is the marginal probability $P(x|\theta) = \sum_y P(x, y|\theta)$. Write

$$Q(\theta|\theta^{(i)}) = \sum_y P(y|x, \theta^{(i)}) \log P(x, y|\theta).$$

The *Expectation Maximization Algorithm* is usually stated formally like this:

- *E-step:* Calculate function $Q(\theta|\theta^{(i)})$.

- *M-step:* Maximize $Q(\theta|\theta^{(i)})$ with respect to θ.

More about the theory and applications of the EM algorithm could be found in [13].

An EM algorithm for estimating the offspring probabilities in MTBP is easy to define. Let x be the observed set of particles, π is the unobserved tree structure and θ is the set of parameters - the offspring probabilities. Then the joint probability of the "complete" observation is:

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

where $T_v \to \mathcal{A}$ is the rule that a particle of type T_v produces the set of particles $\mathcal{A}$ and c is a count function. We have $\sum_{\mathcal{A}} p(T_v \to \mathcal{A}) = 1$. The probability of the "incomplete" observation is the marginal probability $P(x|\theta) = \sum_{\pi} P(x,\pi|\theta)$. Then

$$Q(\theta|\theta^{(i)}) = \sum_{T_v \to \mathcal{A}} E_{\theta^{(i)}} c(T_v \to \mathcal{A}) \log p(T_v \to \mathcal{A})$$

and directly maximizing it we get to the result that the re-estimating parameters are the normalized expected counts

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

where the expected number of times a particle of type T_v appears in the tree π is:

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

The M-step is explicitly solved, so no effort on maximization is needed. The problem is that in general enumerating all possible trees π is of exponential complexity. As cited above, we have proposed using the inside-outside algorithm for stochastic context-free grammars to reduce complexity.

Grammars are well developed tool for modelling strings of symbols in computational linguistics. Stochastic grammars give a probabilistic approach to the problems in that field. A stochastic context-free grammar (SCFG) consists of a number of symbols and a number of production rules of the form $\alpha \to \beta$, where α and β are sequences of these symbols. The symbols could be two kinds – abstract nonterminal and terminal that actually appear in an observation. There are also probabilities assigned to the rules. For a SCFG to be in Chomsky normal form it

is necessary rules to be of the form $X \to YZ$ or $X \to a$, where X, Y, Z are nonterminals and a is a terminal symbol. Every CFG could be represented in Chomsky normal form. For such grammars there exist an EM-type algorithm, called the inside-outside algorithm [12], which finds a ML estimator of the parameters θ of that grammar, namely the probabilities of the rules, called the transition and emission probabilities respectively for the first and the second type of rules above. It is a three dimensional dynamic programming algorithm.

A MTBP could be represented as a SCFG the following way. First our process have to be represented only with "rules" of the form

$$X \xrightarrow{p} \{Y, Z\},$$

which means that a particle of type X could produce two particles of types Y and Z with probability p. For every such rule in the process, the corresponding SCFG will include nonterminals $\{X, Y, Z, Y^T, Z^T\}$, terminals $\{y, z\}$ and rules

$$X \xrightarrow{p_1} YZ|ZY, \quad X \xrightarrow{p_2} Y^T Z|ZY^T, \quad X \xrightarrow{p_3} YZ^T|Z^T Y,$$

$$X \xrightarrow{p_4} Y^T Z^T|Z^T Y^T, \quad Y^T \xrightarrow{1} y, \quad Z^T \xrightarrow{1} z,$$

and $p_1 + p_2 + p_3 + p_4 = p$.

Here Y^T and Z^T are nonterminals of "terminal" type, meaning that they transform into terminals y and z only. We regard these terminals like the observed particles, and the other nonterminals represent the hidden structure of the process. Thus, for a single rule in the process there are six rules in the grammar and the number of types doubles.

In general, to use the Inside-Outside Algorithm for MTBP, we take the following steps:

1. Construct the corresponding SCFG.
2. Estimate parameters for SCFG using as observed sequences all possible permutations of the observed set of particles. Thus, if we have observed 2 particles of type X and 1 of type Y, we use as "observed sequences" all xxy, xyx and yxx.
3. If the number of permutations is large, a Monte Carlo sample approach could be used to obtain the estimate.
4. Calculate probabilities in MTBP summing up ones estimated in SCFG.

3. Biological Model

Oligodendrocyte type-2 astrocyte (O-2A) progenitor cells are known to be precursors of oligodendrocytes in the developing central nervous system. When plated *in vitro* and stimulated to divide by purified cortical astrocytes or by platelet-derived growth factor, these cells grow in clones giving rise to oligodendrocytes. An O-2A progenitor cell is partially committed to differentiation into an olygodendrocyte but it retains the ability to proliferate. Oligodendrocytes are terminally differentiated (mature) cells and they do not divide under normal conditions. At different time points over a period of several days after plating, the composition of each clone is examined microscopically to count the numbers of O-2A progenitor cells and oligodendrocytes per clone. A certain number N of cell clones, each originating from a single initiator cell, are followed-up with the observation process being either longitudinal or serial sacrifice, depending on the experimental design (see [17] for details).

We consider a MTBP with two types of particles T_1 (progenitor cells) and T_2 (oligodendrocytes), where the second type is terminal – a particle of this type does not reproduce. The productions allowed are:

$$T_1 \stackrel{p_1}{\to} \{T_1, T_1\}, \quad T_1 \stackrel{p_2}{\to} T_2,$$

where $p_1 + p_2 = 1$.

The corresponding SCFG has nonterminals T_1, T_2, T_1^T, T_2^T, terminals t_1, t_2 , and rules:

$$T_1 \to T_1T_1 | T_1^TT_1 | T_1T_1^T | T_1^TT_1^T | T_2 | T_2^T,$$

$$T_1^T \stackrel{p_1}{\to} t_1, \quad T_1^T \stackrel{p_2}{\to} t_2, \quad T_2^T \stackrel{1}{\to} t_2,$$

And in Chomsky normal form the grammar is:

$$T_1 \to T_1T_1 | T_1^TT_1 | T_1T_1^T | T_1^TT_1^T,$$

$$T_1^T \stackrel{p_1}{\to} t_1, \quad T_1^T \stackrel{p_2}{\to} t_2.$$

A simulation of the process has been performed and several sets of independent observations have been generated. Using the approach described above estimates of p_1 and p_2 have been obtained and compared to the initial values. Calculations are made in R (see [15]).

4. Results and Conclusions

A Galton-Watson process with two types of particles have been simulated and the population in the fifth generation has been observed. The offspring probabilities have been set to $p_1 = P(T_1 \to \{T_1, T_1\}) = 2/3$, $p_2 = P(T_1 \to T_2) = 1/3$. A set of hundred observations has been generated and several subsets have been randomly taken from it to form the test samples. The results for three of them are shown in Table 1.

First sample consists of following sets: $\{3\ T_1, 3\ T_2\}$, $\{6\ T_1, 3\ T_2\}$, $\{8\ T_1, 1\ T_2\}$, $\{5\ T_1, 3\ T_2\}$, $\{6\ T_1, 2\ T_2\}$, (5 observations).

Second sample is: $\{8\ T_1, 1\ T_2\}$, $\{4\ T_1, 4\ T_2\}$, $\{4\ T_1, 2\ T_2\}$, $\{8\ T_1, 3\ T_2\}$, $\{6\ T_1, 5\ T_2\}$, (5 observations).

And the third sample consists of the sets included in the first and second samples plus 5 more sets: $\{9\ T_1, 3\ T_2\}$, $\{6\ T_1, 3\ T_2\}$, $\{10\ T_1, 2\ T_2\}$, $\{1\ T_1, 4\ T_2\}$, $\{2\ T_1, 6\ T_2\}$, (15 observations).

	real values	sample 1	sample 2	sample 3
p_1	0.667	0.665	0.590	0.642
p_2	0.333	0.335	0.410	0.358

Table 1: Estimates of the offspring probabilities for three samples compared to their real values.

Five more samples have been used and the resulting estimates can be seen in Table 2. The mean and standard deviation of the estimates from the eight samples have been calculated and the result is very close to the original values used in the simulation.

	s. 1	s. 2	s. 3	s. 4	s. 5	s. 6	s. 7	s. 8	mean	st.dev.
p_1	0.665	0.59	0.642	0.72	0.63	0.5	0.78	0.68	0.651	0.084
p_2	0.335	0.41	0.358	0.28	0.37	0.5	0.22	0.32	0.349	0.084

Table 2: Estimates of the offspring probabilities for all eight samples.

The results of that simulation experiment show that the estimates obtained through the algorithm described above could be used in practice where branching process models occur. They are obtainable in reasonable time. (It took several seconds for each sample for this model on a contemporary PC, though more complex models will need more time.) Being ML estimates, they have the

drawback to be sensitive to outliers when the sample size is small (see sample 2 for example), but with larger samples they become consistent.

REFERENCES

[1] S. Asmussen, H. Hering. Branching Processes, Birkhauser, Boston, 1983.

[2] K. B. Athreya, P. E. Ney. Branching Processes, Springer, Berlin, 1972.

[3] N. Daskalova. Using Inside-Outside Algorithm for Estimation of the Offspring Distribution in Mutitype Branching Processes. *Serdica Journal of Computing*, **4** (2010), 463–474.

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

[5] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison. Biological sequence analysis. Cambridge University Press, 1998.

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

[7] M. González, J. Martín, R. Martínez, M. Mota. Non-parametric Bayesian estimation for multitype branching processes through simulation-based methods. *Computational Statistics & Data Analysis,* **52** (2008), No 3, 1281–1291.

[8] P. Guttorp. Statistical inference for branching processes, New York: Wiley, 1991.

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

[10] O. Hyrien. Pseudo-likelihood estimation for discretely observed multitype Bellman-Harris branching processes. *Journal of Statistical Planning and Inference,* **137** (2007), No 4, 1375–1388.

[11] M. Kimmel, D. E. Axelrod. Branching Processes in Biology, Springer, New York, 2002.

[12] K. Lari, S. J. Young. The Estimation of Stochastic Context-Free Grammars Using the Inside-Outside Algorithm. *Computer Speech and Language,* **4** (1990), 35–36.

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

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

[15] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010, URL `http://www.R-project.org`.

[16] D. Sankoff. Branching Processes with Terminal Types: Application to Context-Free Grammars. *Journal of Applied Probability* **8** (1971), No 2, 233–240.

[17] A. Y. Yakovlev, N. M. Yanev. Relative frequencies in multitype branching processes. *Ann. Appl. Probab.*, **19(1)** (2009), 1–14.

[18] N. M. Yanev. Statistical Inference for Branching Processes. In: M. Ahsanullah, G. P. Yanev (Eds), Records and Branching Processes, Nova Sci. Publishers, Inc, 2008, 143–168.

Nina Daskalova
Sofia University "St.Kliment Ohridski",
Faculty of Mathematics and Informatics,
Sofia, Bulgaria,
e-mail: `ninad@fmi.uni-sofia.bg`