PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **13** (2000), 173-178

PLISKA
STUDIA MATHEMATICA
BULGARICA

A SYSTEM FOR SIMULATION AND ESTIMATION OF BRANCHING PROCESSES*

Daniela Nitcheva, Nickolay M. Yanev

A computer code system for simulation and estimation of branching processes is proposed. Using the system, samples for some models with or without migration are generated. Over these samples we compare some properties of various estimators.

1 Introduction

The software system SimBP is developed to simulate different models branching processes and to estimate various parameters of these processes. Simulation algorithm is based on following definition of the branching process with random migration (BPRM).

Let us have on some probability space three independent sets of integer-valued random variables, i.i.d. in each set, $X = \{X_{t,i}\}$, $\eta = \{(\eta_{t,1}, \eta_{t,2})\}$, $I = \{(I_t, I_t^0)\}$. Then we define

$$Z_t = (\sum_{i=1}^{Z_{t-1}} X_{t,i} + M_t)^+, \quad t = 1, 2, \ldots, \quad Z_0 \geq 0, \tag{1}$$

where

$$M_t = \begin{cases} -(\sum_{i=1}^{\eta_{t,1}} X_{t,i} + \eta_{t,2}), & \text{with probability } p, \\ 0, & \text{with probability } q, \\ I_t \mathbf{1}_{\{Z_{t-1}>0\}} + I_t^0 \mathbf{1}_{\{Z_{t-1}=0\}}, & \text{with probability } r. \end{cases}$$

Here $p + q + r = 1$ and Z_0 is independent of X, η, I. As usual $a^+ = \max(0, a)$.

We refer to the random variables in (1) by the following way: $X_{t,i}$ is the offspring in the t-th generation of the i-th individual which exists in the $(t-1)$-th generation. In the t-th generation the following three situations are possible: (i) with probability p: $\eta_{t,1}$

*The work is supported by Bulgarian NFSI, grant MM-704/97

families emigrate (*family emigration*) that is $\eta_{t,1}$ individuals are eliminated in the $(t-1)$-th generation (before the reproduction) and do not take part in the further evolution, and additionally after the reproduction in the $(t-1)$-th generation $\eta_{t,2}$ individuals emigrate from the t-th generation who can be chosen randomly from different families (*individual emigration*); (ii) with probability q: the reproduction is according to the Bienayme-Galton-Watson process (BGW), i.e. without any migration; (iii) with probability r: a state dependent immigration of new individuals is possible - I_t individuals in the non-zero states or I_t^0 in the state zero.

We use the following notations: $m = \mathbf{E}X_{t,i}$ and $\sigma^2 = \mathbf{Var}X_{t,i}$ are the offspring distribution mean and variance; $e_1 = \mathbf{E}\eta_{t,1}$ is the family emigration distribution mean; $e_2 = \mathbf{E}\eta_{t,2}$ is the individual emigration distribution mean; $\lambda = \mathbf{E}I_t$ and $b^2 = \mathbf{Var}I_t$ are the immigration distribution mean and variance; $\lambda_0 = \mathbf{E}I_t^0$ and $b_0^2 = \mathbf{Var}I_t^0$ are the mean and the variance of the immigration distribution in state zero.

Many particular cases can be obtained from the model (1). When $q = 1$ the process $\{Z_t, t = 0, 1, \ldots\}$ is a classical BGW process. When $r = 1$ and I_t and I_t^0 are identically distributed the process is a branching process with immigration (BGWI). The process (1) with $p = 1$ i.e. the process with pure emigration was studied for $\eta_{t,2} = 0$ a.s. by Vatutin (1977) and Kaverin (1990) and for $\eta_{t,1} = 0$ a.s. by Grey (1988). Some results for the general process (1) are obtained by Yanev and Yanev (1995, 1996, 1997). Models with non-homogeneous migration, i.e. $p = p_t$, $q = q_t$ and $r = r_t$ were investigated by Yanev and Mitov (1985).

2 The system SimBP

Simulation of each model is determined by some basic parameters. They are: probability distributions of random variables which the process consists of; the initial number of particles Z_0; the length of the simulated path, which is the number of generations.

The system allows to generate the most popular discrete distributions - Binomial, Poisson, Geometric - according to the parameters given by the user. The user also can give an arbitrary discrete distribution with finite set of values. The algorithms for generating each of these distributions are based on a generator of independent random numbers uniformly distributed over interval $(0, 1)$. The random process simulation follows exactly the described constructive definition (1).

The system SimBP works also with data given by the user.

When the system gets the data, it shows a graphic presentation of the process path. That enables the user to have a visual idea about the process progress.

3 Estimators

Further, the user can get estimators of process parameters. The sample $Z_0, Z_1, \ldots, Z_t$ by which the estimators are computed consists of generation sizes. That is why in SimBP are made many of known non-parametric estimators for different models branching processes.

To compute other estimators is possible to use other sets of observations like the full tree of generations; two successful generations $Z_t - 1, Z_t$; the initial and another observation Z_0, Z_t; censored observations $Z_L(n), Z_{L+1}(n), \dots, Z_{L+T}(n)$.

When the process is simulated its parameter values are known. Then we can compare the properties of various estimators. As a criterion for comparison we use the empirical variance

$$\Delta_N(\hat{\theta}_t, \theta) = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_t(j) - \theta)^2,$$

where θ is the true parameter value, $\hat{\theta}_t(j)$ is the estimator obtained from j-th sample path and N is the number of simulated sample paths.

The most simple branching process is Galton-Watson process (BGW). Its asymptotic behavior depends on the offspring mean m. This arises the problem of parameter m estimating. Many different estimators for m and the offspring variance σ^2 are discussed in [3]. The most popular estimators for the m are:

$$m_t^* = (Z_t)^{\frac{1}{t}} \quad \text{known as Heyde estimator;}$$

$$\overline{m}_t = \begin{cases} \dfrac{Z_t}{Z_{t-1}}, & Z_{t-1} > 0, \\ 1, & Z_{t-1} = 0 \end{cases} \quad \text{known as Lotka - Nagaev estimator;}$$

(2) $$\hat{m}_t = \frac{Z_1 + Z_2 + \cdots + Z_t}{Z_0 + Z_1 + \cdots + Z_{t-1}} \quad \text{known as Harris estimator.}$$

All three estimators are strongly consistent over set of nonextinction paths if $1 < m < \infty$ and $\sigma^2 < \infty$.

To compare the estimators (2) the system SimBP simulates 1000 sample paths each with number of generations t of Galton-Watson process with Poisson offspring distribution with mean $m = 1.2$ and variance $\sigma^2 = 1.2$. From each path we obtain the three estimators and compute their empirical variance. Let t take the values 20, 30 and 50. The values of Δ_N are shown in Table 1.

t	m_t^*	$\overline{m}_t$	$\hat{m}_t$
20	0.0069	0.0470	0.0034
30	0.0030	0.0747	0.0009
50	0.0014	0.0006	0.0002

TABLE 1. Values of Δ_N for mean estimators for a Galton-Watson process.

Note that the maximum Likelyhood estimator $\hat{m}_t$ has the smallest empirical variance.

Estimators of offspring mean m and immigration mean λ for BGWI process are proposed first by Heyde and Seneta (1972). They use the conditional least squares method. The estimators are:

$$\overline{m}_t = \frac{\sum_{i=1}^{t} Z_i \sum_{i=1}^{t} Z_{i-1} - t\sum_{i=1}^{t} Z_i Z_{i-1}}{(\sum_{i=1}^{t} Z_{i-1})^2 - t\sum_{i=1}^{t} Z_{i-1}^2},$$

(3) $$\overline{\lambda}_t = \frac{\sum_{i=1}^{t} Z_i Z_{i-1} \sum_{i=1}^{t} Z_{i-1} - \sum_{i=1}^{t} Z_{i-1}^2 \sum_{i=1}^{t} Z_i}{(\sum_{i=1}^{t} Z_{i-1})^2 - t\sum_{i=1}^{t} Z_{i-1}^2}.$$

Using the conditional least squares method, Wei and Winnicki [4] obtain the estimators:

$$\tilde{m}_t = \frac{\sum_{i=1}^{t} Z_i \sum_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t\sum_{i=1}^{t} \frac{Z_i}{1+Z_{i-1}}}{\sum_{i=1}^{t}(1+Z_{i-1})\sum_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t^2},$$

(4) $$\tilde{\lambda}_t = \frac{\sum_{i=1}^{t} Z_{i-1} \sum_{i=1}^{t} \frac{Z_i}{1+Z_{i-1}} - \sum_{i=1}^{t} Z_i \sum_{i=1}^{t} \frac{Z_{i-1}}{1+Z_{i-1}}}{\sum_{i=1}^{t}(1+Z_{i-1})\sum_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t^2}.$$

The estimators $\overline{m}_t$ and $\tilde{m}_t$ are consistent if $0 < m < \infty$, $\overline{\lambda}_t$ is consistent if $0 < m < 1$ and $\tilde{\lambda}_t$ is consistent if $0 < m \le 1$.

To compare the estimators (3) and (4) again we simulate 1000 sample paths with number of generations t. The offspring distribution is Geometric with mean $m = 0.5$ and variance $\sigma^2 = 0.75$ and immigration distribution Poisson with mean $\lambda = 3$ and variance $b^2 = 3$. Let t take the values 100, 200 and 300. The empirical variances are shown in Table 2.

t	$\overline{m}_t$	$\tilde{m}_t$	$\overline{\lambda}_t$	$\tilde{\lambda}_t$
100	0.0102	0.0096	0.3565	0.2892
200	0.0058	0.0062	0.1747	0.1887
300	0.0040	0.0040	0.1266	0.1164

TABLE 2. Values of Δ_N for mean estimators for a process with immigration.

The estimators obtained by both methods are comparable.

Here we want to propose estimators of the parameters of a branching process with random migration. Let I_t and I_t^0 have the same distribution with mean λ. Then one can prove that

$$(Z_{t+1}|Z_t) = mZ_t - p(me_1 + e_2) + r\lambda - p(mZ_t - me_1 - e_2)\bowtie_{\{A_t\}},$$

where A_t is the event $\{\sum_{i=1}^{Z_t} X_{t,i} \le \sum_{i=1}^{\eta_{t,1}} X_{t,i} + \eta_{t,2}\}$. We use the notations $M = r\lambda - p(me_1 + e_2)$ for the mean of a random migration component in each generation. Let $\delta_t = p(mZ_t - me_1 - e_2)\bowtie_{\{A_t\}}$. This yields $(Z_{t+1}|Z_t) = mZ_t + M - \delta_t$. We call M mean of random migration.

Let $Z_0, Z_1, \ldots, Z_t$ be the sample of the generation sizes of BPRM. Using the conditional least squares method, we obtain the estimators:

$$\overline{m}_t = \frac{\sum\limits_{i=1}^{t} Z_i \sum\limits_{i=1}^{t} Z_{i-1} - t\sum\limits_{i=1}^{t} Z_i Z_{i-1}}{(\sum\limits_{i=1}^{t} Z_{i-1})^2 - t\sum\limits_{i=1}^{t} Z_{i-1}^2},$$

$$\overline{M}_t = \frac{\sum\limits_{i=1}^{t} Z_i Z_{i-1} \sum\limits_{i=1}^{t} Z_{i-1} - \sum\limits_{i=1}^{t} Z_{i-1}^2 \sum\limits_{i=1}^{t} Z_i}{(\sum\limits_{i=1}^{t} Z_{i-1})^2 - t\sum\limits_{i=1}^{t} Z_{i-1}^2}. \tag{5}$$

The weighted conditional least squares estimators are

$$\tilde{m}_t = \frac{\sum\limits_{i=1}^{t} Z_i \sum\limits_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t\sum\limits_{i=1}^{t} \frac{Z_i}{1+Z_{i-1}}}{\sum\limits_{i=1}^{t}(1+Z_{i-1})\sum\limits_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t^2},$$

$$\tilde{M}_t = \frac{\sum\limits_{i=1}^{t} Z_{i-1} \sum\limits_{i=1}^{t} \frac{Z_i}{1+Z_{i-1}} - \sum\limits_{i=1}^{t} Z_i \sum\limits_{i=1}^{t} \frac{Z_{i-1}}{1+Z_{i-1}}}{\sum\limits_{i=1}^{t}(1+Z_{i-1})\sum\limits_{i=1}^{t} \frac{1}{1+Z_{i-1}} - t^2}. \tag{6}$$

The analytical investigation of these estimators is very difficult and we do not have any analytical results for them up to now and this is an open problem.

We use the empirical distance

$$\epsilon_N(\hat{\theta}_t, \theta) = |\overline{\theta}_N(t) - \theta|, \text{ where } \overline{\theta}_N(t) = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_t(j),$$

to study which estimator (5) or (6) is closer to the true parameter value. Again we simulate 1000 paths for each number of generations $t = 50, 100, 200$. The offspring mean is $m = 0.6$ and migration mean is $M = 1.28$. The results are in Table 3.

t	$\overline{m}_t$	$\tilde{m}_t$	$\overline{M}_t$	$\tilde{M}_t$
50	0.0645	0.0630	0.2070	0.1950
100	0.0362	0.0376	0.1320	0.1366
200	0.0256	0.0291	0.1092	0.1260

TABLE 3. Values of ϵ_N for a process with migration.

It can be seen that when the number of generations increases the difference between the estimators and the true parameter values decreases.

In the case $m > 1$ it is interesting to use Harris estimator (2). We make 100 sample paths simulations of branching process with migration which has offspring mean $m = 2$ and migration mean $M = -1.7$. Each sample path consists of t generations. Let t takes values 15, 20 and 25. The results are presented in Table 4.

t	$\hat{m}_t$
15	0.0069
20	0.0030
25	0.0014

TABLE 4. Values of ϵ_N for Harris estimator for a process with migration.

It can be seen that even if $M < 0$ ϵ_N decreases quickly.

Finally, we would like to mention that all results from the system SimBP can be obtained very easy and quickly by the user. The obtained results confirm the theoretical conclusions of Dion and Yanev (1995, 1997).

The SimBP will be developed for more complicated models of branching processes.

Bibliography

[1] J. P. Dion, N. M. Yanev Statistical inference for branching processes with an increasing number of ancestors. *J. Statistical Planning & Inference*, **39** (1994), 329-359.

[2] J. P. Dion, N. M. Yanev Limit theorems and estimation theory for branching processes with an increasing number of ancestors. *J. Appl. Prob.*, **34** (1997), 309-327.

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

[4] J. Winnicki. Estimation theory for the branching processes with immigration. *Contemporary Mathematics*. **80** (1988), 301-322.

[5] G. P. Yanev, N. M. Yanev. Critical branching process with random migration. In: Branching Processes (ed. C. C. Heyde), Lecture Notes in Statistics, vol. **99**, (Springer-Verlag, New York) 1995, 36-46.

[6] G. P. Yanev, N. M. Yanev. Branching processes with two types emigration and state-dependent immigration. In: Lecture Notes in Statistics, vol. **114**, (Springer, New York) 1996, 216-228.

[7] G. P. Yanev, N. M. Yanev. Limit theorems for branching processes with random migration stopped at zero. *The IMA volumes in Math. and its Appl.*, **84**, (Springer, New York), 1997, 323-336,

Daniela Nitcheva
Dept.of Probability and Statistics
Faculty of Mathematics, University of Sofia
5, James Bourcher str.
1126 Sofia, Bulgaria
*e-mail:*daniela@fmi.uni-sofia.bg

Nickolay Yanev
Dept.of Probability and Statistics
Institute of Mathematics
Bulgarian Academy of Sciences
Acad. G.Bontchev str., bl. 8
1113 Sofia,Bulgaria
e-mail: yanev@math.bas.bg