МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2005
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2005
Proceedings of the Thirty Fourth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 6–9, 2005

ALGORITHMS FOR GENERATION AND ROBUST ESTIMATION OF BRANCHING PROCESSES WITH RANDOM NUMBER OF ANCESTORS*

Vessela K. Stoimenova, Dimitar V. Atanasov, Nickolay M. Yanev

Generation of sample paths in branching processes with random number of ancestors has its important role in studying the properties of the parametric estimators. Using different types of generated realizations and the asymptotic normality of some popular offspring mean estimators their robust modification is obtained. When the offspring distribution is considered to belong to the class of PSOD, robust estimators, based on several sample paths, are studied.

1. Introduction. We assume that on some probability space there exists a set of i.i.d. r.v. $\{\xi_i(t,n)\}$ with values in the set of nonnegative integers $N = \{0,1,2,\dots\}$ and that $\{\xi_i(t,n),\ i \in N\}$ are independent of $Z_0(n)$. Then for each $n = 1,2,\dots$ $Z(n) = \{Z_t(n), t = 0,1,\dots\}$ is a Bienayme-Galton-Watson process having a random number of ancestors $Z_0(n) \geq 1$, where

$$Z_t(n) = \begin{cases} \sum_{i=1}^{Z_{t-1}(n)} \xi_i(t,n) & \text{if } Z_{t-1}(n) > 0, \quad t = 1,2,\dots \\ 0, & \text{otherwise} \end{cases}$$

Such a process is denoted by BGWR.

Our main concern in this paper is the robust parametric estimation of a BGWR process with power series offspring distribution (PSOD) based on a sample $\{Z_0(n),\dots, Z_t(n)\}$ as both n and t tend to infinity (and thus $Z_0(n)$ in some sense). Naturally, the relative speed, at which n and $t \to \infty$, comes into play for all values of the offspring mean $m : 0 < m < \infty$.

Let $\{p_k\}$ be the common offspring distribution, i.e. $p_k = P(\xi = k) \geq 0$, $\sum p_k = 1$, $p_0 + p_1 < 1$ and put $m = E\xi$, $\sigma^2 = Var(\xi)$. We assume throughout that $0 < \sigma^2 < \infty$.

The individual distribution is said to belong to the class of power series offspring distributions, if $p_k = P(\xi = k) = \dfrac{a_k\theta^k}{A(\theta)}$, $\theta > 0$, $a_k \geq 0$, where $A(\theta) = \sum a_k\theta^k$ is a positive function.

Yakovlev and Yanev (1989) noted that branching processes with a large and often random number of ancestors occur naturally in the study of cell proliferation. Such is

*2000 Mathematics Subject Classification: 60J80

The paper is supported by the National Science Fund of Bulgaria, Grant No MM-1101/2001.

also the case in applications to nuclear chain reactions. Results about the nonparametric estimation of the offspring mean m and variance σ^2 in the BGWR process have been announced in [3], [2], the proofs of several results about the Harris estimator of the offspring mean $\widehat{m}_t(n) = \dfrac{Z_1(n) + \cdots + Z_t(n)}{Z_0(n) + \cdots + Z_t(n)}$ are given in [4]. A sequel to this work is the paper of [5], where the nonparametric m.l.e. and a family of l.s.e. for σ^2 are concerned and consistency and asymptotic normality of these estimators are obtained for all values of the mean m, $0 < m < \infty$.

Further on, we suppose that $n = n(t) \to \infty$ as $t \to \infty$ and use the following

Condition A. *$m > 1$ or $m = 1$, $t/n \to 0$ or $m < 1$, $nm^t \to \infty$.*

We use a robust extension of the maximum likelihood estimators (MLE) that possesses a high breakdown point, which was introduced in [9] and [10], the so called Weighted Least Trimmed estimator in order k (WLT(k)). As a measure of robustness we consider the Finite Sample Breakdown Point (BP), defined as the largest fraction of oservations from the original data, which can be replaced by arbitrary values (see [8], [1] and the references therein).

2. Robust modified nonparametric estimators. We apply the concept of the WLT(k) estimators for estimating the offspring mean in the BGWR processes. The study is focused on the well known estimators of Lotka-Nagaev and Harris using the technique for the classical Bienayme-Galton-Watson (BGW) process described in [6]. Let us suppose that we have a set of sample paths of a branching process. Using this set and the estimators mentioned above we can obtain a number of values for the offspring mean (for any sample path we have one offspring mean estimator). It is well known that, under certain conditions, these values are asymptotically normally distributed. If these conditions are not satisfied the estimated value is far from the real value of the offspring mean. The aim is to apply the theory of robustness in order to eliminate the cases, which do not satisfy these conditions, and to obtain an estimator of the offspring mean closer to the real value. The study of the robustness of the estimates of the offspring mean is based on the breakdown properties of the $WLT(k)$ estimators.

Following [6], let us define a robust estimator of the unknown parameter θ over the set of sample paths $\mathbf{Z} = \{\mathbf{Z^{(1)}(n)}, \ldots, \mathbf{Z^{(r)}(n)}\}$, where $\mathbf{Z^{(r)}(n)}\}$ is a single realization of a BGWR process with PSOD, $r = 1, 2, \ldots$, as

$$\bar{M}(\theta) = \underset{\theta \in R}{\operatorname{argmin}} \sum_{i=1}^{k} -w_i f(\mathrm{Est}\,(\mathbf{Z}^{(\nu(\mathbf{i}))}(\mathbf{n}), \theta)), \tag{1}$$

where k is the trimming factor, $f(x)$ is the logarithm of the density function of the standard normal distribution, ν is a permutation of the indexes, such that $f(\mathrm{Est}\,(\mathbf{Z}^{(\nu(\mathbf{1}))}(\mathbf{n}), \theta))$ $\geq f(\mathrm{Est}\,(\mathbf{Z}^{(\nu(\mathbf{2}))}(\mathbf{n}), \theta)) \geq \cdots \geq f(\mathrm{Est}\,(\mathbf{Z}^{(\nu(\mathbf{r}))}(\mathbf{n}), \theta))$, $\mathrm{Est}(\mathbf{Z^{(i)}(n)}, \theta)$ is the transformation of the estimator of θ, which gives us asymptotic normality.

In the case of Lotka-Nagaev estimator $\mathrm{Est}(\mathbf{Z^{(i)}(n)}, \mu)$ can be presented as follows:

$$\mathrm{Est}\,(\mathbf{Z^{(i)}(n)}, \mu) = \frac{\sqrt{Z_{t_i}^{(i)}(n)}}{\sigma}(\bar{m}_{t_i}^{(i)}(n) - \mu), \mu \in R$$

where $\bar{m}_{t_i}^{(i)}(n)$ is Lotka-Nagaev estimator for the $i-$th sample path. Here $Z_{t_i}^{(i)}(n)$ are the number of individuals in t_i-th generation in the sample path $\mathbf{Z^{(i)}(n)}$. The fixed parameter

σ represents the variance of the offspring distribution. By analogy, in the case of Harris estimator, we have

$$\text{Est}\,(\mathbf{Z^{(i)}}(\mathbf{n}),\mu) = \frac{\sqrt{U_{t_i}^{(i)}(n)}}{\sigma}(\hat{m}_{t_i}^{(i)}(n)-\mu), \mu \in R$$

where $U_{t_i}^{(i)}(n) = Z_0^{(i)}(n)+\cdots+Z_{t_i-1}^{(i)}(n)$. And, finally, for the offspring variance one gets:

$$\text{Est}\,(\mathbf{Z^{(i)}}(\mathbf{n}),\sigma) = \left(\frac{t_i}{2\sigma^4}\right)^{\frac{1}{2}}(\hat{\sigma}_{t_i}^{2,(i)}(n)-\sigma^2),$$

where

$$\hat{\sigma_{t_i}^2}(n) = \frac{1}{t_i}\sum_{k=0}^{t_i-1} Z_k(n)\left(\frac{Z_{k+1}(n)}{Z_k(n)}-m\right)^2$$

is the estimator of the individual variance and m is the true value of the individual mean if it is known, or the Lotka-Nagaev or Harris estimators of the mean if it is unknown. $\text{Est}(\mathbf{Z^{(i)}}(\mathbf{n}),\sigma)$ is asymptotically normal if Condition A holds, $E\xi^4<\infty$ and $\frac{Z_0(n)}{n} \vec{d}\ \nu$ (see [5]).

Proposition 1. *The estimator $\bar{M}(\mu)$, defined by* (1)*, exists and its BP is not less than $(r-k)/r$ if $r \geq 3$, $(r+1)/2 \leq k \leq r-1$.* $\square$

The proof is omitted because it is similar to the case of the classical BGW process (see [6]).

Proposition 2. *The estimator $\bar{M}(\sigma)$, defined by* (1)*, exists and its BP is not less than $(r-k)/r$ if $r \geq 3$, $(r+1)/2 \leq k \leq r-1$.*

Proof. To prove this proposition we have to find out the index of fullness of the set $F = \{f(Est(\mathbf{Z^{(i)}}(\mathbf{n}),\sigma)), i = 1,\cdots,r\}$. Let us consider the function $g(\sigma) = f(Est(\mathbf{Z^{(i)}}(\mathbf{n}),\sigma))$ for a given sample path $\mathbf{Z^{(i)}}(\mathbf{n})$. The study of the function $g(\hat{\sigma}^2)$ shows that it can be presented as follows:

$$g(\hat{\sigma}_{t_i}^{2,i}(n)) = \log\frac{1}{\sqrt{2\pi}} - \frac{C^2(\hat{\sigma}_{t_i}^{2,i}-\sigma^2)^2}{\sigma^4},$$

where $C = \frac{\sqrt{t_i}}{2}$. It is obvious that $g(\hat{\sigma}^2)$ tends to minus infinity when σ^2 tends to zero and has an asymptote as σ^2 tends to infinity. It does not satisfy the conditions of the criterion for subcompactness , but is still subcompact in the sense of the theory of the generalized d-fullness and its breakdown point is not less than $(r-k)/r$ if $r \geq 3$, $(r+1)/2 \leq k \leq r-1$. $\square$

One should notice that the reliability of the robust estimator depends extremely on the rate of convergence to the limiting distribution of the initial estimator. We have simulated 100 sample paths of a 50 generations BGW process with individual Poisson distribution with mean 1.2 (the supercritical case) and have calculated the estimators of the offspring distribution when the mean is known. Our aim was to test the normality of the obtained sample of transformations $Est(S_i,\sigma)$. We performed a Jarque-Bera test for goodness-of-fit to a normal distribution. It showed that the normality assumption could be rejected even at significance level 0.5. This can be seen at the normal probability plot shown bellow (Fig. 1). The normal probability plot performs like those of the heavy-

tailed distributions. For instance, we have compared it with the normal probability plot of the Cauchy distribution (Fig. 2).

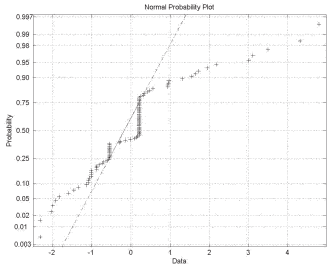


Fig. 1. Estimates of σ

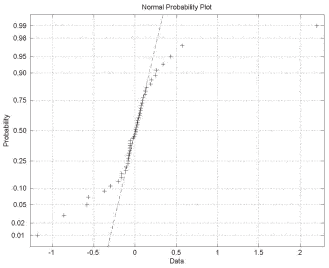


Fig. 2. Cauchy distribution

The robust modification of the variance estimator of a BGW process proposed here can be used, but to obtain good results, as in the case of offspring mean estimators, one has to use a large data set with thousands of generations, which is often not useful from a practical point of view.

3. Robust parametric estimation. In this section we construct robust estimators of the parameter of the BGWR process with PSOD, based on the entire family tree and on the generation sizes. We use results from [7], where the BP of these models is studied according to the properties of the processes.

Let us first consider the situation, when we are able to observe the entire family tree. Let ϑ_k be the number of particles with k offspring. Then, if M is the total progeny, the log likelihood has the form $L_M(\theta) = \left(\frac{a_0\theta^0}{A(\theta)}\right)^{\vartheta_0} \cdot \left(\frac{a_1\theta^1}{A(\theta)}\right)^{\vartheta_1} \cdot \ldots \cdot \left(\frac{a_N\theta^N}{A(\theta)}\right)^{\vartheta_N}$.

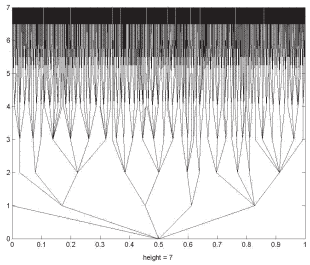
In order to check the properties of the proposed estimators, BGWR processes have to be simulated, where the whole family trees are observed. On the graphics below we have shown simulations of some family trees with one ancestor and different offspring distributions:

In the next model we are not able to observe the entire family tree and information only about the generation sizes is available. The proposed method for constructing robust estimators uses several sample paths over the process. Let us have at our disposal r independent realizations $Z^{(i)}(n)$, $i = 1, \ldots, r$, from a BGWR process with one and the same PSOD and number of generations, equal to t_i. Then the likelihood function, based on all r realizations, has the form:

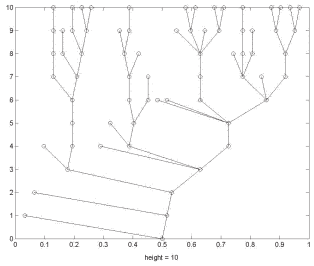
$$L_r(\theta) = \prod_{i=1}^{r} \mathbf{P}(Z_0^{(i)}(n)) \prod_{l=1}^{r} \left\{ \prod_{i=1}^{t_1} \left[\sum_{s_1+\cdots+s_{Z_{i-1}^{(l)}(n)}=Z_i^{(l)}(n)} \prod_{j=1}^{Z_{i-1}^{(l)}} a_{s_j} \right] \right\} \prod_{i=1}^{r} \frac{\theta^{\sum_{j=1}^{t_i} Z_j^{(i)}(n)}}{(A(\theta))^{\sum_{j=0}^{t_i-1} Z_j^{(i)}(n)}}.$$

The behaviour and the properties of WLT(k) estimators, based on this model, are studied in [7]. Let l be the number of sample paths with extinction at time 1. There is shown, that in the case of Poisson offspring distribution the WLT(k) estimator, based on

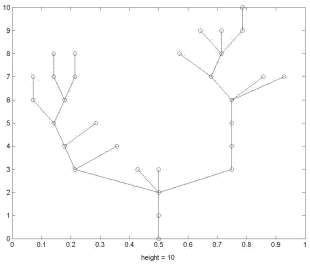
Bi(10;0,14)



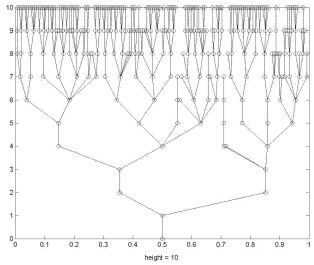
Bi(10;0,18)



Ge(1/22)



Po(1,5)



r realizations, exists, possesses BP not less than $(r-k)/r$ if $r \geq 3(l+1)$, $(r+l+1)/2 \leq k \leq r-l-1$ and is consistent if $m_{(i)} < \infty$ and $Z_0^{(i)}(n) \xrightarrow{P} \infty$ or $\sigma^2_{(i)} < \infty$, $Z_0^{(i)}(n)/n. \xrightarrow{d} \nu$ and Condition A holds for $i = 1, \ldots, r$.

We have simulated 10 sample paths of BGW processes with Poisson offspring distribution and offspring mean values shown in the first rows of the tables given below and with 5 outliers with Poisson offspring distribution with mean 2 (supercritical situation on Table 1) and mean 0.5 (subcritical on Table 2).

It is easy to be seen that the estimated value of the unknown parameter is close to the generated one. The estimator is stable to the presence of outliers in the data.

offspring mean	estimate
0.8	0.8139
0.9	0.9500
1.0	1.1020
1.1	1.1257
1.3	1.2169
1.5	1.4103
1.8	1.8052

Table 1

offspring mean	estimate
0.7	0.7222
0.9	0.8657
1.1	1.0597
1.5	1.4726
1.8	1.8227

Table 2

REFERENCES

[1] D. Atanasov, N. Neykov. On the Finite Sample Breakdown Point of the WLTE(k) and d-fullness of a Set of Continuous Functions. *Proceedings of the VI International Conference "Computer Data Analysis And Modeling"*, Minsk, Belarus (2002).
[2] J. P. Dion. Statistical Inference for Discrete Time Branching Processes. Proc. 7th International Summer School on Prob.Th.& Math.Statist. Varna 1991, *Sci. Cult. Tech. Publ.*, Singapore, (1993), 60–121.
[3] J. P. Dion, N. M. Yanev. Limiting Distributions of a Galton-Watson Branching Process with a Random Number of Ancestors. *Compt. rend. Acad. bulg. Sci.*, **44**, 1991, No 3, 23–26.
[4] J. P. Dion, N. M. Yanev. Estimation Theory for Branching Processes with or without Immigration. *Compt. rend. Acad. bulg. Sci.*, **44**, 1991, No 4, 19–22.
[5] 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.
[6] V. Stoimenova, D. Atanasov, N. Yanev. Robust Estimation and Simulation of Branching Processes, *Compt. rend. Acad. bulg. Sci.*, **5**, 2004, 19–22.
[7] V. Stoimenova. Robust Parametric Estimation of Branching Processes with Random Number of Ancestors, *Pliska Stud. Math.*, 2005 (in print).
[8] D. Vandev. A Note on Breakdown Point of the Least Median of Squares and Least Trimmed Estimators. *Statistics and Probability Letters*, **16**, 1993, 117–119.
[9] D. Vandev, N. Neykov. Robust Maximum Likelihood in the Gaussian Case, *New Directions in Statistical Data Analysis and Robustness*, S. Morgenthaler, E. Ronchetti, and W. A. Stahel (eds.), Birkhauser Verlag, Basel, 1993.
[10] D. Vandev, N. Neykov. About Regression Estimators with High Breakdown Point. *Statistics*, **32**, 1998, 111–129.

Vessela Stoimenova
Dimitar Atanasov
Faculty of Mathematics and Informatics
Sofia University
5, J. Boucher Str.
1164 Sofia, Bulgaria
e-mail: stoimenova@fmi.uni-sofia.bg
datansov@fmi.uni-sofia.bg

Nickolay Yanev
Bulgarian Academy of Science
Acad. G. Bontchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: yanev@math.bas.bg

АЛГОРИТМИ ЗА ГЕНЕРИРАНЕ И РОБАСТНО ОЦЕНЯВАНЕ НА РАЗКЛОНЯВАЩИ СЕ ПРОЦЕСИ СЪС СЛУЧАЕН БРОЙ НАЧАЛНИ ЧАСТИЦИ

Весела Стоименова, Димитър В. Атанасов, Николай М. Янев

Генерирането на траектории разклоняващи се процеси със случаен брой начални частици играе важна роля в изучаването на свойствата на параметричните и непараметричните оценки. Получени са модификации на някои известни оценки на индивидуалните средно и дисперсия като е използвана тяхната асимптотична нормалност и различни типове генерирани реализации. Когато индивидуалното разпределение принадлежи на класа на разпределенията тип степенен ред са разгледани робастни оценки, базирани на една и няколко извадъчни траектории.