МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2006
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2006
Proceedings of the Thirty Fifth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 5–8, 2006

NONPARAMETRIC ROBUST ESTIMATION OF THE INDIVIDUAL DISTRIBUTION IN BRANCHING PROCESSES WITH A RANDOM NUMBER OF ANCESTORS*

Vessela K. Stoimenova, Dimitar V. Atanasov

In the present paper we consider the Bienayme–Galton–Watson process with a random number of ancestors. The asymptotic normality of the estimators of the individual distribution is combined with the general idea of the trimmed and weighted maximum likelihood. As a result, a robust modification of the estimators of the individual probabilities is proposed. It is based on several realizations of the entire family tree and is studied *via* their simulation and numerical results.

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 the nonnegative integers $N = \{0, 1, 2, \dots\}$ and that $\{\xi_i(t,n),\ i \in N\}$ are independent of the positive integer-valued r.v. $Z_0(n)$ and can be considered as independent copies of some r.v. ξ. Then, for each $n = 1, 2, \dots$ $\mathbf{Z}(\mathbf{n}) = \{Z_t(n), t = 0, 1, \dots\}$ is a Bienayme-Galton-Watson process having a random number of ancestors $Z_0(n) \geq 1$, where

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

Such a process is denoted by BGWR.

Our main purpose in this paper is the robust nonparametric estimation of the individual distribution of a BGWR process, based on several realizations of the entire family tree.

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 the paper that $0 < \sigma^2 < \infty$, $Z_0(n)/n \xrightarrow{P} \nu$, where ν is a positive r.v.; $n, t \to \infty$ and $n/t \to \infty$ in the critical case $m = 1$.

Yakovlev and Yanev (1989) (see e.g. [6]) noticed that branching processes with large and often random number of ancestors occur naturally in the study of cell proliferation.

*The paper is supported by the National Science Fund of Bulgaria, Grant No VU-MI -105/2005.
2000 Mathematics Subject Classification: 60J80

Similar case appears in applications to nuclear chain reactions. Results on the nonparametric estimation of the offspring mean m and variance σ^2 in the BGWR process are announced in [3], [4], [5] and [6].

In the present paper we use a robust extension of the maximum likelihood estimators (MLE) that possesses a high breakdown point, which is introduced in [11] and [12]. It is the so called Weighted Least Trimmed estimator of order k ($WLT(k)$ estimator). The notation and the definition of this estimator is presented in the next Section 2. The robust modification of the classical nonparametric estimators of the individual distribution is introduced in Section 3. Section 4 contains simulations of BGWR processes and numerical results about the proposed estimators.

2. The robust estimation. According to Vandev and Neykov [12], the $WLT(k)$ estimator $\widehat{\theta}$ for the unknown parameter $\theta \in \Theta^p$ is defined as

$$\widehat{\theta} = \underset{\theta\in\Theta^p}{\operatorname{argmin}} \sum_{i=1}^{k} w_i f_{\nu(i)}\left(\theta\right). \tag{1}$$

Here $f_{\nu(1)}\left(\theta\right) \leq f_{\nu(2)}\left(\theta\right) \leq \cdots \leq f_{\nu(n)}\left(\theta\right)$ are the ordered values of $f_i = -\log\varphi\left(x_i, \theta\right)$ at θ, $\varphi\left(x_i, \theta\right)$ is a probability density, θ is an unknown parameter and $\nu = (\nu(1), \ldots, \nu(n))$ is the corresponding permutation of the indices, which may depend on θ. The weights w_i are nonnegative and at least k of them are strictly positive.

As a measure of robustness of a given estimator Hampel et al. [7] propose a finite sample breakdown point. According to Vandev [10] this breakdown point for a given estimator T is defined as $\varepsilon(T) = \dfrac{1}{n}\max\{m : \sup\|T(X_m)\| < \infty\}$, where X_m is a sample, obtained from the sample X by replacing any m of the observations by arbitrary values.

According to the definitions, given by Vandev [9], a function $g(\theta)$ is subcompact if its Lesbegue sets $L_g\left(C\right) = \{\theta : g\left(\theta\right) \leq C\}$ are compact for any real constant C. A finite set $F = \{f_i(\theta)\}_{i=1}^{n}$ of n functions is called d-full, if for each subset $J \subset \{i = 1, \ldots, n\}$ of cardinality d $(|J| = d)$, the supremum $g\left(\theta\right) = \sup\{f_i(\theta)\}$ is a subcompact function.

Applying the theory of d-fullness [9], Vandev and Neykov [12] proved that the finite sample breakdown point of the $WLT(k)$ estimators is not less than $(n-k)/n$ if $n \geq 3d$, $(n+d)/2 \leq k \leq n-d$, when Θ^p is a topological space and the set $F = \{f_i(\theta), i = 1, \ldots, n\}$ is d-full.

A simpler and easier criterion for subcompactness is given in [2], where it is proved that the real valued continuous function $g\left(\theta\right)$, defined on an open subset of $D \subset R^n$, is subcompact if and only if for any sequence $\theta_i \to \theta_0$, where θ_0 belongs to the boundary of D, $g\left(\theta_i\right) \to \infty$ when $i \to \infty$. Thus, if D is a compact set, then any continuous function, defined on D, is subcompact.

3. Robust modified nonparametric estimators. We apply the concept of the WLT(k) estimators in order to estimate the offspring distribution in the BGWR processes. Suppose that we have a set of sample paths of the entire family tree of a branching process. Using this set and the above mentioned estimators we obtain a number of values for the offspring distribution. Under the conditions of Theorem 2.2 [6] these values are asymptotically normal. If these conditions are not satisfied, then the estimated value is far from the value of the particular probability. The aim is to apply the weighted and trimmed likelihood in order to eliminate the cases which do not satisfy these conditions, and to obtain estimators of the individual probabilities, closer to the faithful values.

Let us consider the set $\mathbf{Z} = \{\mathbf{Z^{(1)}}(\mathbf{n}),\ldots,\mathbf{Z^{(r)}}(\mathbf{n})\}$, where $\{\mathbf{Z^{(i)}}(\mathbf{n})\}$ is a single realization of a BGWR process, such that the number of offspring of each particle is available, $i = 1, 2, \ldots, r$

Let

$$(2) \qquad \widehat{p}_k^{(i)}(n,t) \;=\; \sum_{j=0}^{t-1} \vartheta_k^{(i)}(j,t) / \sum_{k=0}^{\infty}\sum_{j=0}^{t-1} \vartheta_k^{(i)}(j,t) = \vartheta_k^{(i)}(t)/Y_t(n),$$

$i = 1, 2, \ldots, r$, be the estimator of p_k for the sample path $\mathbf{Z^{(i)}}(\mathbf{n})$, introduced in [6]; $\vartheta_k^{(i)}(t)$ and $\vartheta_k^{(i)}(j,t)$ be the number of particles in the i-th sample path with k offspring in the first $t+1$ generations and in the j-th generation, respectively, and $Y_t(n)$ be the total number of particles in the first t generations.

Let

$$Est(\mathbf{Z^{(i)}}(\mathbf{n}), p_k) = \sqrt{Y_t^{(i)}(n)}(\widehat{p}_k^{(i)}(n,t) - p_k)/\sqrt{p_k(1-p_k)}), \quad 0 \le p_k \le 1,$$

be the transformation of the estimator of p_k, which is asymptotically normal [6]:

$$Est(\mathbf{Z^{(i)}}(\mathbf{n}), p_k) \xrightarrow{d} N(0,1) \;\; as \;\; t \to \infty.$$

Following [8], for a given set of family trees $\mathbf{Z} = \{\mathbf{Z^{(1)}}(\mathbf{n}),\ldots,\mathbf{Z^{(r)}}(\mathbf{n})\}$ let us introduce a trimmed estimator based on a sample of Harris estimates of the unknown individual distribution in a BGWR process $\{p_k, k \ge 0\}$. Then, the estimator is presented as follows:

$$(3) \qquad \widehat{p}_k^{T}(n,t) = \operatorname*{argmin}_{p_k \in (0,1)} \sum_{i=1}^{T} -w_i f(Est(\mathbf{Z^{(\nu(i))}}(\mathbf{n}), p_k)),$$

where T is the trimming factor, $f(x)$ is the log-density of the standard normal distribution; ν is a permutation of the indices, such that

$$f(Est(\mathbf{Z^{(\nu(1))}}(\mathbf{n}),\theta)) \ge f(Est(\mathbf{Z^{(\nu(2))}}(\mathbf{n}),\theta)) \ge \cdots \ge f(Est(\mathbf{Z^{(\nu(T))}}(\mathbf{n}),\theta)).$$

Let us denote by $N_0^k = \sharp\{i = 1, 2, \ldots, r : \;\; \widehat{p}_k^{(i)}(n,t) = 0\}$, i.e. the number of the estimators of the individual probability p_k equal to zero, and by $N_1^k = \sharp\{i = 1, 2, \ldots, r : \widehat{p}_k^{(i)}(n,t) = 1\}$ – the number of estimators of p_k that are equal to 1.

Proposition 1. *Assume that in a BGWR stochastic process $0 < \sigma^2 < \infty$, $Z_0(n)/n \xrightarrow{P} \nu$, where ν is a positive r.v., $n, t \to \infty$ and $n/t \to \infty$ in the critical case $m = 1$. Let the random variable ξ (interpreted as the number of offspring of one particle) be not degenerate and take the values $k_1, k_2, \ldots, k_N$, where N is a positive integer or infinity, with positive probabilities. Then, the estimator $\widehat{p}_{k_s}^{T}(n,t)$, $s = 1, 2, \ldots, N$ of the individual probability p_{k_s}, defined by* (3), *exists and its breakdown point is not less than $(r-T)/r$, if $r \ge 3(\max\{N_0^{k_s}, N_1^{k_s}\}+1)$, $(r + \max\{N_0^{k_s}, N_1^{k_s}\} + 1)/2 \le T \le r - \max\{N_0^{k_s}, N_1^{k_s}\} - 1$.*

Proof. Note that under the conditions $0 < \sigma^2 < \infty$, $Z_0(n)/n \xrightarrow{P} \nu$, $n, t \to \infty$ and $n/t \to \infty$ in the critical case $m = 1$, the transformations $Est(\mathbf{Z^{(i)}}(\mathbf{n}), p_{k_s})$ are asymptotically normal [6].

We have to find out the index of fullness of the set

$$F = \{-f(Est(\mathbf{Z^{(i)}(n)}, p_{k_s})), i = 1, \dots, n\}.$$

The conditions of the proposition ensure that $0 < p_{k_s} < 1$. Let us consider the function $g^{(i)}(p_{k_s}) = f(Est(\mathbf{Z^{(i)}(n)}, p_{k_s}))$ for the sample path $\mathbf{Z^{(i)}(n)}$, $i = 1, 2 \dots, r$. It holds that

$$\begin{aligned} g^{(i)}(p_{k_s}) &= \log \frac{1}{\sqrt{2\pi}} - \frac{Est^2(\mathbf{Z^{(i)}(n)}, p_{k_s})}{2} = \\ &= \log \frac{1}{\sqrt{2\pi}} - \frac{C^2 \left((\widehat{p}^{(i)}_{k_s}(n,t) - p_{k_s}) / \sqrt{p_{k_s}(1 - p_{k_s})} \right)^2}{2}, \end{aligned}$$

where the constant $C = \sqrt{Y_t^{(i)}(n)}$ does not depend on p_{k_s}.

When $\widehat{p}^{(i)}_{k_s}(n,t) \neq 0$ or 1, the function $-g^{(i)}(p_{k_s})$ satisfies the conditions for the subcompactness, because $\lim\limits_{p_{k_s} \to 0^+} Est^2(\mathbf{Z^{(i)}(n)}, \mathbf{p_{k_s}}) = \lim\limits_{p_k \to 1^-} Est^2(\mathbf{Z^{(i)}(n)}, \mathbf{p_{k_s}}) = \infty$. From this follows the subcompactness of $-g^{(i)}(p_{k_s})$ for $0 < p_{k_s} < 1$.

Let $\widehat{p}^{(i)}_{k_s}(n,t) = 0$. Then $g_0^{(i)}(p_{k_s}) = \log \frac{1}{\sqrt{2\pi}} - \frac{C^2}{2} \frac{p_{k_s}}{1 - p_{k_s}}$ and $\lim\limits_{p_{k_s} \to 0^+} g_0^{(i)}(p_{k_s}) = \log \frac{1}{\sqrt{2\pi}}$, $\lim\limits_{p_{k_s} \to 1^-} g_0^{(i)}(p_{k_s}) = -\infty$.

If $\widehat{p}^{(i)}_{k_s}(n,t) = 1$, then $g_1^{(i)}(p_{k_s}) = \log \frac{1}{\sqrt{2\pi}} - \frac{C^2}{2} \frac{1 - p_{k_s}}{p_{k_s}}$ and $\lim\limits_{p_{k_s} \to 0^+} g_1^{(i)}(p_{k_s}) = -\infty$, $\lim\limits_{p_{k_s} \to 1^-} g_1^{(i)}(p_{k_s}) = \log \frac{1}{\sqrt{2\pi}}$. But

$$\begin{aligned} \lim_{p_{k_s} \to 0^+} 1/2[g_0^{(i)}(p_{k_s}) + g_1^{(i)}(p_{k_s})] &= \lim_{p_{k_s} \to 0^+} 1/2[g_0^{(i)}(p_{k_s}) + g^{(i)}(p_{k_s})] = \\ \lim_{p_{k_s} \to 0^+} 1/2[g^{(i)}(p_{k_s}) + g_1^{(i)}(p_{k_s})] &= \lim_{p_{k_s} \to 1^-} 1/2[g_0^{(i)}(p_{k_s}) + g_1^{(i)}(p_{k_s})] = \\ \lim_{p_{k_s} \to 1^-} 1/2[g_0^{(i)}(p_{k_s}) + g^{(i)}(p_{k_s})] &= \lim_{p_{k_s} \to 1^-} 1/2[g_1^{(i)}(p_{k_s}) + g^{(i)}(p_{k_s})] = -\infty. \end{aligned}$$

Note that the average of n functions is always smaller than their supremum. Therefore, the supremum of any two functions from the set $\{-g_0^{(i)}(p_{k_s}),\ -g_1^{(i)}(p_{k_s}),\ -g^{(i)}(p_{k_s})\}$ is a subcompact function. Consequently, the supremum of any $\max\{N_0^{k_s}, N_1^{k_s}\} + 1$ functions from the set F is a subcompact function while the supremum of $\max\{N_0^{k_s}, N_1^{k_s}\}$ functions from the set F may be not subcompact according to the choice of the functions in it. Hence, the index of fullnes of the set F is $\max\{N_0^{k_s}, N_1^{k_s}\} + 1$. The existence of the estimator follows from the statements in Vandev [9]. Now, applying the theorem of Vandev and Neykov [12], the proposition is proved. □

4. Computational results. In order to study the applicability of the proposed estimators of the individual probabilities we realized a number of simulations. We simulated 10 family trees, each of them with 30 generations and Poisson individual distribution $Po(1.5)$, and 3 family trees, each of them with 10 generations and individual distribution $Po(5)$. These 3 BGWR processes are the outliers.

The calculated individual nonparametric (not robust) estimates for the individual distribution of each tree (obtained by formula (2)) are given in the next table (trees 11, 12 and 13 are the outlier trees, the first column denotes the number of the simulated tree):

No	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7
1	0.22728	0.32721	0.25839	0.1254	0.044085	0.012001	0.003673	0.001469
2	0.21952	0.33166	0.25336	0.1264	0.050615	0.013423	0.00392	0.001118
3	0.221	0.34122	0.23817	0.13966	0.041154	0.016202	0.002592	0
4	0.23143	0.32078	0.26431	0.12149	0.048444	0.010291	0.00301	0.00025
5	0.5	0.5	0	0	0	0	0	0
6	0.2286	0.34057	0.23662	0.12464	0.047065	0.016033	0.005430	0.000776
7	0.22826	0.32652	0.25328	0.12115	0.051572	0.014037	0.003357	0.001526
8	1	0	0	0	0	0	0	0
9	0.22429	0.32933	0.25	0.1288	0.048482	0.014936	0.003673	0.000245
10	1	0	0	0	0	0	0	0
11	0.0075291	0.030116	0.080082	0.14066	0.17625	0.17659	0.1448	0.1105
12	0.0048622	0.038088	0.08752	0.14182	0.17261	0.16613	0.14506	0.10778
13	0.0067164	0.028358	0.077612	0.13955	0.18209	0.1806	0.15746	0.09627

In the next table the following results are presented:

Column 2 – Theoretical probabilities: gives the values of the probabilities from the Poisson distribution with mean 1.5;

Column 3 – Trimmed esimates: includes the robust esimates of the probabilities in Column 2, obtained by formula (3) from all 13 trees (including the outliers); The values are calculated using the algorithm given by Atanasov [1].

Column 4 – $\widehat{p}_k(13, 30)$: gives the classical nonparametric estimates of the probabilities in Column 2, obtained by formula (2) from the data from all 13 trees (i.e. obtained over 1 BGWR tree, starting with 13 ancestors);

Column 5 – $\widehat{p}_k(10, 30)$: gives the classical nonparametric estimates of the probabilities in Column 2, obtained by formula (2) from the data from all 10 trees, which are not outliers (i.e. obtained over 1 BGWR tree, starting with 10 ancestors);

Column 6 – $\widehat{p}_k(3, 10)$: gives the classical nonparametric estimates of the probabilities in Column 2, obtained by formula (2) from the data from the 3 trees, which are outliers (i.e. obtained over 1 BGWR tree, starting with 3 ancestors);

Column 7 – Standard error: describes the standard errors (the inverse of the second derivative of the likelihood function at the point of minimum) of the robust estimates in Column 3.

	Theoretical probabilities	Robust estimates	$\widehat{p}_k(13, 30)$	$\widehat{p}_k(10, 30)$	$\widehat{p}_k(3, 10)$	Standard error
p_0	0.2231	0.2843	0.1877	0.2260	0.0067	0.018
p_1	0.3347	0.2743	0.2784	0.3307	0.0315	0.0022
p_2	0.2510	0.2539	0.2213	0.2510	0.0811	0.0036
p_3	0.1255	0.1341	0.1290	0.1265	0.1406	0.0029
p_4	0.0470	0.0498	0.0700	0.0474	0.1769	0.0018
p_5	0.0141	0.0145	0.0420	0.0138	0.1752	0.0010
p_6	0.0035	0.0036	0.0289	0.0037	0.1479	0.00049
p_7	0.00075	0.0009	0.0192	0.0008	0.1064	0.00029
p_8	0.00014	0.00018	0.0108	0.0002	0.0608	0.00023

Comments. *One should notice that in the presented example the classical nonparametric estimates of the individual distribution over all realizations of the process behave very well in the absence of outliers. However, in the presence of outliers the estimates may be seriously affected if the corresponding probabilities of the outliers are 'far' from the true value. In these situations the proposed trimmed estimates are more 'adequate' to the true stochastic model. The difference between the robust and the classical estimates is not so clear in the cases when the true probability is close to that of the outliers.*

Acknowledgements The authors would like to express their appreciation to Professor N. Yanev for the invaluable help and advises. The authors are also grateful to the referee for the constructive comments and suggestions that have led to significant improvement of the paper.

REFERENCES

[1] D. Atanasov. About the Concept of Weights of $WLTE(k)$ Estimators. *Pliska Stud. Math. Bulgar.*, **14** (2003), 5–14.

[2] 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, 2001.

[3] J.-P. Dion. Statistical Inference for Discrete Time Branching Processes. In: Proc. 7th International Summer School on Prob. Th. & Math.Statist, Varna, 1991. Sci. Cult. Tech. Publ., Singapore, 1993, 60–121.

[4] 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.

[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] J.-P. Dion, N. M. Yanev. Limit Theorems and Estimation Theory for Branching Processes with an Increasing Random Number of Ancestors. *J. Appl. Prob.*, **34** (1997), 309–327.

[7] F. R. Hampel, E. Ronchetti, P. Rousseeuw, W. Stahel. Robust Statistics: The Approach Based on influence Functions. John Wilew and Sons, New York, 1986.

[8] V. Stoimenova, D. Atanasov, N. Yanev. Robust Estimation and Simulation of Branching Processes. *Compt. rend. Acad. bulg. Sci.*, **57** (2004), No 5, 19–22.

[9] 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.

[10] D. Vandev. Robust Methods in Industrial Statistics. Proceedings of the 1st International Conference MII 2003, Thessaloniki, Greece, 2003, 318–334.

[11] D. Vandev, N. Neykov. Robust Maximum Likelihood in the Gaussian Case. In: New Directions in Statistical Data Analysis and Robustness (Eds S. Morgenthaler, E. Ronchetti, W. A. Stahel), Birkhauser Verlag, Basel, 1993.

[12] D. Vandev, N. Neykov. About Regression Estimators with High Breakdown Point. *Statistics*, **32** (1998), 111–129.

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

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

НЕПАРАМЕТРИЧНО РОБАСТНО ОЦЕНЯВАНЕ НА ИНДИВИДУАЛНОТО РАЗПРЕДЕЛЕНИЕ В РАЗКЛОНЯВАЩ СЕ ПРОЦЕС СЪС СЛУЧАЕН БРОЙ НАЧАЛНИ ЧАСТИЦИ

Весела К. Стоименова, Димитър В. Атанасов

В настоящата статия асимптотичната нормалност на класическите оценки на индивидуалното разпределение в разклоняващ се стохастичен процес на Биенеме–Галтон–Уотсън със случаен брой начални частици е съчетана с идеята на претегленото и орязано правдоподобие. Като резултат е предложена робастна модификация на оценките на индивидуалните вероятности. Тя е основана на няколко реализации на фамилни дървета и е изследвана посредством техни симулации и числени резултати.