Nonparametric Inference for Controlled Branching Processes with Deterministic Function

Miguel González[1], Carmen Minuesa[2], and Inés del Puerto[3,4]
[1,2,3]Department of Mathematics. University of Extremadura, Badajoz, SPAIN
[4]Corresponding author: Inés del Puerto, e-mail: idelpuerto@unex.es

Abstract

Controlled branching processes are stochastic growth population models in which the number of individuals with reproductive capacity in each generation is controlled by a deterministic function. The behaviour of these populations is strongly related to the main parameters of the offspring distribution. In practice these values are unknown and their estimation is necessary. Usually it must be observed the whole family tree up to a given generation in order to estimate the offspring distribution. In this work, we deal with the problem of estimating the main parameters of the model assuming that the only observable data are the total number of individuals in each generation. We set out the problem in a nonparametric framework and obtain the maximum likelihood estimator of the offspring distribution using the expectation-maximization algorithm.

Keywords: Controlled process, offspring distribution, maximum likelihood estimation, expectation-maximization algorithm.

1. Introduction

The standard branching process (also called Bienaymé-Galton-Watson process) is not always adequate to describe actual phenomena. There are many variants that have been proposed to deal with particular problems. One of them is the Controlled Branching Process (CBP) in which the number of individuals (particles) with reproductive capacity in each generation is governed by a control function ϕ (which could be deterministic or random). These processes were introduced by Sevastyanov and Zubkov (1974), and include as particular cases, some other modifications of the standard branching process (e.g. the Bienaymé Galton-Watson process with immigration) or even the standard branching process.

The probabilistic theory of CBPs, in particular the study of its extinction problem and its limiting behaviour, has been extensively investigated, see for example Bagley (1986), González et al. (2005) (and references therein) and Sevastyanov and Zubkov (1974). The presence of the control mechanism makes complex the study of this kind of process, nevertheless it allows to model a much greater variety of behaviours than the standard branching process.

In practice, the offspring distribution is usually unknown, and need to be estimated to guarantee the applicability of these models. Inferential studies for CBPs from a frequentist viewpoint may be found in González et al. (2004), González et al. (2005) and Sriram et al. (2007).

The purpose of this article is to consider the estimation of the offspring law (and some derived parameters) of a controlled model (with deterministic control function) from the general nonparametric outlook. We begin in Section 2 with a brief description on CBP, where some notation and basic results are provided. In

Section 3 we consider the estimation of the offspring distribution. This section is split in two subsections. The first one is devoted to exposing how to obtain maximum likelihood estimators (MLE) of the parameters based on the observation of the entire family tree. Actually we review the results given in González et al. (2004). In the second part, since usually it is not possible to observe in practice the entire family tree, a more realistic sample is considered, that given only by the total number of individuals and progenitors in each generation. The problem of obtaining MLEs based on this sample is tackled as an incomplete data problem. To deal with it, we use the Expectation-Maximization (EM) algorithm (see McLachlan and Krishnan (2008)).

2. Probability Model

A CBP is defined recursively as follows:

$$Z_0 = N, \quad Z_{n+1} = \sum_{j=1}^{\phi(Z_n)} X_{nj} \quad n = 0, 1, \ldots \tag{1}$$

where the empty sum is considered to be 0, N is a non–negative integer, $\{X_{nj} : n = 0, 1, \ldots; j = 1, 2, \ldots\}$ is a sequence of i.i.d. non-negative integer-valued random variables, with $\{p_k\}_{k\geq 0}$ being the common probability distribution (reproduction law or offspring distribution) and with ϕ being a function that is non–negative and integer–valued for integer–valued arguments.

Intuitively, Z_n denotes the number of individuals (particles) in the n-th generation. Thus, if $\phi(Z_n) < Z_n$ then $Z_n - \phi(Z_n)$ individuals are artificially removed from the population and, therefore, they do not participate in the future evolution of the process. If $\phi(Z_n) > Z_n$ then $\phi(Z_n) - Z_n$ new individuals of the same type are added to the population participating under the same conditions as the others. No control is applied to the population when $\phi(Z_n) = Z_n$. Obviously, if $\phi(n) = n$ for all n, we obtain the standard branching process.

It is easy to verify that $\{Z_n\}_{n\geq 0}$ is a Markov chain with stationary transition probabilities.

In the following, we assume that $p_0 > 0$ and $\phi(k) = 0$ if and only if $k = 0$, in consequence 0 is an absorbing state and the states $k = 1, 2, \ldots$ are transient. Whence it is verified the *extinction-explosion dichotomy*, that is $P[Z_n \to 0] + P[Z_n \to \infty] = 1$.

We also suppose that the mean and variance of the reproduction law are finite, i.e. $m = \sum_{k=0}^{\infty} k p_k < \infty$ and $\sigma^2 = \sum_{k=0}^{\infty} (k-m)^2 p_k < \infty$.

3. Maximum Likelihood Estimators of the Offspring Distribution

We consider in this section the MLEs of the probabilities p_k, $k = 0, 1, \ldots$, i.e., of the reproduction law. First, we assume that the entire family tree is observed until a given generation. Second, we consider that only the total number of individuals and progenitors in each generations are observed.

3.1 Based on the entire family tree

We consider that the entire family tree up to the current nth generation can be observed, i.e., $\{X_{lj} : j = 1, \ldots, \phi_l(Z_l); l = 0, 1, \ldots, n-1\}$. Let $Z_l(k) = \sum_{j=1}^{\phi(Z_l)} I_{\{X_{lj}=k\}}$, $k \geq 0$, with I_A standing for the indicator function of the set A.

Intuitively, $Z_l(k)$ represents the number of progenitors at the lth generation with exactly k offspring. It is deduced that

$$\phi(Z_l) = \sum_{k=0}^{\infty} Z_l(k) \quad \text{and} \quad Z_{l+1} = \sum_{k=0}^{\infty} kZ_l(k), \quad l \geq 0$$

Let also $Y_j(k) = \sum_{l=0}^{j} Z_l(k)$, $j \geq 0$, that is, the accumulated number up to generation j of progenitors that give rise to exactly k offspring. Moreover, denote $Y_n = \sum_{k=0}^{n} Z_k$ and $\Delta_n = \sum_{k=0}^{n} \phi(Z_k)$, $n \geq 0$, that is the total progeny and the total number of progenitors, respectively, up to generation n.

Finally, let denote $\mathcal{Z}_n^* = \{Z_l(k),\ k \in \mathcal{S},\ l = 0, 1, \ldots, n-1\}$.

The following result, given in González et al. (2004), provides us the MLE of the offspring law as well as the MLEs of the offspring mean and variance.

Theorem 1 *Let $\{Z_n\}_{n\geq 0}$ be a CBP. The MLE of p_k for $k \geq 0$, based on $\mathcal{Z}_n^* = \{Z_l(k) : l = 0, \ldots, n-1; k = 0, 1, \ldots\}$ is:*

$$\widehat{p_k} = \frac{Y_{n-1}(k)}{\Delta_{n-1}}, \qquad k = 0, 1, \ldots \tag{2}$$

Moreover, the MLEs of the parameters m and σ^2, are, respectively

$$\widehat{m} = \frac{Y_n - Z_0}{\Delta_{n-1}} \quad \textit{and} \quad \widehat{\sigma}^2 = \sum_{k=0}^{\infty} (k - \widehat{m})^2 \widehat{p_k}.$$

Remark 1

1. *The strong consistency and asymptotic normality on the non-extinction set of these estimators were established in González et al. (2004, 2005).*
2. *It can be proved (see González et al. (2004)) that $\widehat{m}$ is also the MLE of m based on the sample $\{Z_0, \phi(Z_l), Z_{l+1}, l = 0, \ldots, n-1\}$.*

3.2 Based on the population size in each generation: EM-Algorithm

In real situations it is difficult to observe the whole family tree up to the current generation or even the random variables $Z_l(k)$, $k \geq 0$, $l = 0, \ldots, n-1$. Hence, in this subsection we assume the more realistic requirement that these are unobservable, being the observable data only the total number of individuals and progenitors in each generation, that is $\overline{\mathcal{Z}}_n = \{Z_0, \phi(Z_l), Z_{l+1}, l = 0, \ldots, n-1\}$. Then, one is faced with an incomplete data estimation problem. In such a case, it seems appropriate to use an Expectation-Maximization (EM) algorithm (see McLachlan and Krishnan (2008)), in order to obtain MLEs. In our case, this algorithm is an iterative method which starts with certain initial values of the parameters $p = \{p_k\}_{k\geq 0}$ and gives rise to a sequence of vectors which, under certain conditions, converges to the MLEs based on the sample $\overline{\mathcal{Z}}_n$. Each iteration of the method consists of two steps. In the first step (E step), the expectation of the complete log-likelihood is calculated using the distribution of the unobserved data. The second step (M step) consists of finding the values of the parameters which maximize the expectation that had been calculated in the E step. The E and M steps are repeated until convergence is attained. In our case, starting

with initial values $p^{(0)} = \{p_k^{(0)}\}_{k\geq 0}$, we shall obtain a sequence $\{p^{(i)}\}_{i\geq 0}$ which is updated in each iteration of the method, as will be described in the following.

3.2.1 Step E

We shall develop the E step of the EM algorithm in the $(i+1)$-th iteration. Let, for each i, $p^{(i)} = \{p_k^{(i)}\}_{k\geq 0}$ be the vector obtained in the i-th iteration and $\mathcal{Z}_n^*|(p^{(i)}, \overline{\mathcal{Z}}_n)$ the distribution of the latent vector $\mathcal{Z}_n^*$ given the sample $\overline{\mathcal{Z}}_n$ and the parameters of the model $p^{(i)}$. For simplicity, we shall write

$$E_i^*[\cdot] := E_{\mathcal{Z}_n^*|(p^{(i)},\overline{\mathcal{Z}}_n)}[\cdot]$$

It is not hard to obtain the complete log-likelihood

$$\ell(p\,|\mathcal{Z}_n^*,\ \overline{\mathcal{Z}}_n) = \ell(p\,|\mathcal{Z}_n^*) = \sum_{l=0}^{n-1} \log\left(\frac{\phi(Z_l)!}{\prod_{k=0}^{\infty} Z_l(k)!}\right) + \sum_{l=0}^{n-1}\sum_{k=0}^{\infty} Z_l(k)\log p_k,$$

which depends on the variables $Z_l(k)$, $l = 0,\ldots,n-1$, $k \geq 0$, which are not observed. Then the expected value of the complete log-likelihood with respect to the available data $(p^{(i)}, \overline{\mathcal{Z}}_n)$ is given by:

$$\begin{aligned} E_i^*[\ell(p\,|\mathcal{Z}_n^*,\,\overline{\mathcal{Z}}_n)] &= E_i^*\left[\sum_{l=0}^{n-1} \log\Big(\frac{\phi(Z_l)!}{\prod_{k=0}^{\infty} Z_l(k)!}\Big) + \sum_{l=0}^{n-1}\sum_{k=0}^{\infty} Z_l(k)\log p_k\right] \\ &= \sum_{l=0}^{n-1} E_i^*\left[\log\Big(\frac{\phi(Z_l)!}{\prod_{k=0}^{\infty} Z_l(k)!}\Big)\right] + \sum_{l=0}^{n-1}\sum_{k=0}^{\infty} E_i^*\big[Z_l(k)\big]\log p_k \end{aligned} \tag{3}$$

Therefore, in order to obtain the expected value of the complete log-likelihood, we must obtain the distribution of $\mathcal{Z}_n^*$ given $p^{(i)}$ and $\overline{\mathcal{Z}}_n$.

Let $z_l(k) \geq 0$, $l = 0,\ldots,n-1$, $k \geq 0$, be non-negative integers, $z_{l+1} = \sum_{k=0}^{\infty} z_l(k)$, $\phi_l^* = \sum_{k=0}^{\infty} z_l(k) = \phi(z_l)$, then

$$\begin{aligned} &P\big[Z_l(k) = z_l(k), k\geq 0, l = 0,\ldots,n-1 \big| Z_0 = z_0, Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*, l = 0,\ldots,n-1\big] = \\ &\quad = \frac{P\big[\{Z_0 = z_0\} \cap \bigcap_{l=0}^{n-1}\{Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*, Z_l(k) = z_l(k), k\geq 0\}\big]}{P\big[\{Z_0 = z_0\} \cap \bigcap_{l=0}^{n}\{Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*\}\big]} \\ &\quad = \prod_{l=0}^{n-1} \frac{P\big[A_l | \bigcap_{j=0}^{l-1} A_j \cap A\big]}{P\big[B_l | \bigcap_{j=0}^{l-1} B_j \cap A\big]} \end{aligned} \tag{4}$$

where $A = \{Z_0 = z_0\}$ and for each $l = 0,\ldots,n-1$,

$$\begin{aligned} A_l &= \{Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*, Z_l(k) = z_l(k), k\geq 0\} \\ &= \{Z_l(k) = z_l(k), k\geq 0\} \\ B_l &= \{Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*\}. \end{aligned}$$

Now

$$P\left[A_l | \bigcap_{j=0}^{l-1} A_j \cap A\right] = P\big[Z_l(k) = z_l(k), k\geq 0 | Z_l = z_l\big], \tag{5}$$

and

$$P\left[B_l | \bigcap_{j=0}^{l-1} B_j \cap A\right] = P\big[Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^* | Z_l = z_l\big]. \tag{6}$$

From (5) and (6) we obtain:

$$\begin{aligned}
&P\big[Z_l(k) = z_l(k), k \geq 0, l = 0, \ldots, n-1 \big| Z_0 = z_0, Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^*, l = 0, \ldots, n-1\big] = \\
&= \prod_{l=0}^{n-1} \frac{P\big[Z_l(k) = z_l(k), k \geq 0 | Z_l = z_l\big]}{P\big[Z_{l+1} = z_{l+1}, \phi(Z_l) = \phi_l^* | Z_l = z_l\big]} \\
&= \prod_{l=0}^{n-1} \frac{P\big[\sum_{i=0}^{\phi_l^*} I_{\{X_{li}=k\}} = z_l(k), k \geq 0\big]}{P\big[\sum_{i=0}^{\phi_l^*} X_{li} = z_{l+1}\big]} \\
&= \prod_{l=0}^{n-1} \frac{1}{P\big[\sum_{i=0}^{\phi_l^*} X_{li} = z_{l+1}\big]} \cdot \frac{\phi_l^*!}{\prod_{k=0}^{\infty} z_l(k)!} \prod_{k=0}^{\infty} {p_k^{(i)}}^{z_l(k)}
\end{aligned}$$

Taking into account that $z_{l+1} = \sum_{k=0}^{\infty} k z_l(k)$ and $\phi_l^* = \sum_{k=0}^{\infty} z_l(k)$, for each $l = 0, \ldots, n-1$, the infinite products on the latter expression are actually finite, because almost all $z_l(k)$ are null.

Computationally, to sample from $\mathcal{Z}_n^* | (p^{(i)}, \overline{\mathcal{Z}}_n)$ it is enough to sample generation-by-generation. With $l = 0, \ldots, n-1$ fixed, and given Z_{l+1} and $\phi(Z_l)$, we have shown that this can be done by suitably normalizing the probabilities of a multinomial distribution of size $\phi(Z_l)$ and probability $p^{(i)}$.

3.2.2 Step M

The M step consists of finding the values of the parameters $p = \{p_k\}_{k \geq 0}$ which maximize the expectation of the complete log-likelihood. This expectation has been calculated previously in the E step. In our case, we must find the vector $p^{(i+1)} = \{p_k^{(i+1)}\}_{k \geq 0}$ which maximizes the expression (3) subject to the constraints $\sum_{k=0}^{\infty} p_k^{(i+1)} = 1$, $p_k^{(i+1)} \geq 0$, $k = 0, 1, \ldots$. Following a similar argument to that used in the calculation of the MLEs based on the observation of the complete family tree (see González et al. (2004)), one obtains that $p^{(i+1)} = \{p_k^{(i+1)}\}_{k \geq 0}$, is given by, for each $k \geq 0$

$$p_k^{(i+1)} = \frac{\sum_{l=0}^{n-1} E_i^* \left[Z_l(k)\right]}{\sum_{k=0}^{\infty} \sum_{l=0}^{n-1} E_i^* \left[Z_l(k)\right]} = \frac{\sum_{l=0}^{n-1} E_i^* \left[Z_l(k)\right]}{\sum_{l=0}^{n-1} \phi(Z_l)}.$$

Intuitively, $p_k^{(i+1)}$ is the ratio of the average number of progenitors which have generated k offspring to the total number of progenitors.

The values obtained in the M step, $p^{(i+1)} = \{p_k^{(i+1)}\}_{k \geq 0}$, are used to begin another E step and the process is repeated until some convergence criterion is verified, in which case the process stops and the final values are denoted by $\widehat{p}_{EM}$. In McLachlan and Krishnan (2008) it is shown that, under general conditions of differentiability and continuity of the expectation of the complete log-likelihood function, estimates obtained using the EM algorithm converge to a stationary point of the incomplete data likelihood function. The multinomial structure of our complete likelihood function means that usually those conditions are verified, and also that the incomplete data likelihood function is unimodal. Then, in this case, $\widehat{p}_{EM}$ is the MLE of p based on $\overline{\mathcal{Z}}_n$, which we call expectation-maximization MLE.

The following summarizes our proposed EM algorithm to estimate the parameters of the model:

Step 0 *i=0.* Set each value $0 < p_k^{(0)} < 1$.

Step 1 *E Step.* Based on $p^{(i)}$,

(a) determine $\mathcal{Z}_n^*|(p^{(i)}, \overline{\mathcal{Z}}_n)$,

(b) calculate $E_i^*[\ell(p\ |\mathcal{Z}_n^*,\ \overline{\mathcal{Z}}_n)]$.

Step 2 *M Step.* Calculate $p^{(i+1)} = \arg\max_p\ E_i^*[\ell(p\ |\mathcal{Z}_n^*,\ \overline{\mathcal{Z}}_n)]$

Step 3 If $\max\{|p_k^{(i+1)} - p_k^{(i)}|, k \geq 0\}$ is less than some convergence criterion, stop and denote by $\widehat{p}_{EM}$ these final estimates. Otherwise, increment i by 1 and repeat steps 1-3.

Finally, we would point out that since m and σ^2 are obtained from p, then, from $\widehat{p}_{EM}$ one can obtain the expectation-maximization MLEs for m and σ^2 based on $\overline{\mathcal{Z}}_n$, which will be denoted by $\hat{m}_{EM}$ (notice that this is the same that $\hat{m}$ given in Theorem 1) and $\hat{\sigma}^2_{EM}$, respectively. Also, one can obtain a sample of the distribution of Z_{n+s} knowing $\overline{\mathcal{Z}}_n$ for any $s > 0$ by simulating, through the Monte-Carlo method, s generations of a CBP starting with Z_n and considering $\widehat{p}_{EM}$ as the parameters of the model. This allows one to forecast the number of individuals and couples for unobserved generations.

Acknowledgment

This research was supported by the Ministerio de Economía y Competitividad and the FEDER through the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, grant MTM2012-31235 and GR10118.

References

Bagley, J.H. (1986) On the almost sure convergence of controlled branching processes. *J. Appl. Probab.* **23**, 827–831

González, M., Martínez, R., del Puerto, I. (2004) Nonparametric estimation of the offspring distribution and mean for a controlled branching process. *Test* **13**, 465–479

González, M., Martínez, R., del Puerto, I. (2005) Estimation of the variance for a controlled branching process. *Test* **14**, 199–213

González, M., Molina, M., del Puerto, I. (2005) Asymptotic behaviour for the critical controlled branching process with random control function. *J. Appl. Probab.* **42**, 463–477

McLachlan, G.J. and Krishnan, T. (2008). *The EM algorithm and extensions.* John Wiley and Sons, Inc.

Sevastyanov, B.A., Zubkov, A. (1974) Controlled branching processes. *Theor. Prob. Appl.* **19**, 14–24

Sriram, T., Bhattacharya, A., González, M., Martínez, R., del Puerto, I. (2007) Estimation of the offspring mean in a controlled branching process with a random control function. *Stoch. Proc. Appl.* **117**, 928–946