Controlled branching processes: applications in Biology

Miguel González[1,5], Cristina Gutiérrez[2], Rodrigo Martínez[3], and Inés del Puerto[4]
[1,2,4]Department of Mathematics. University of Extremadura, Badajoz, SPAIN
[3]Department of Mathematics. University of Extremadura, Plasencia, SPAIN
[5]Corresponding author: Miguel González, e-mail: mvelasco@unex.es

Abstract

Bayesian analysis of controlled branching processes is developed, considering a non-parametric offspring distribution and control distribution belonging to the power series family of distributions, depending on a single parameter, called control parameter. Mainly, inferences on the offspring distribution, offspring mean and on the control parameter are considered under two sampling schemes: first, the classical one in branching theory based on the observation of the entire family tree; secondly, the more realistic situation in which only the generation-by-generation population size is observed. In this latter case, the Dirichlet process and the Gibbs sampler are used to estimate the posterior density of the main parameters of interest.

Keywords: Controlled process, Bayesian inference, non–parametric offspring law, power series family of control distribution, Gibbs sampler, Dirichlet process.

1. Introduction

The branching model considered in the present work is the controlled branching process. This model is a generalization of the standard Bienaymé-Galton-Watson (BGW) branching process, and, in the terminology of population dynamics, is used to describe the evolution of populations in which a control of the population size at each generation is needed. This control consists of determining the number of individuals with reproductive capacity at each generation mathematically through a random process. In practice, this branching model could describe reasonably the probabilistic evolution of populations in which, for various reasons of an environmental, social, or other nature, there is a mechanism that establishes the number of progenitors which take part in each generation. For example, in an ecological context, one can think of an invasive animal species that is widely recognized as a threat to native ecosystems, but there is disagreement about plans to eradicate it, i.e., while the presence of the species is appreciated by a part of the society, if its numbers are left uncontrolled it is known to be very harmful to native ecosystems. In such a case, it is better to control the population to keep it between admissible limits even though this might mean periods when animals have to be culled. Another practical situation that can be modeled by this kind of process is the evolution of an animal population that is threatened by the existence of predators. So that, in each generation the survival of each animal (and therefore the possibility of giving new births) will be strongly affected by this factor, being necessary the introduction of a random mechanism to model the evolution of this kind of population.

Mathematically, a controlled branching process with random control function (CBP) is a discrete-time stochastic growth population model $\{Z_n\}_{n\geq 0}$ defined

recursively as

$$Z_0 = N \in \mathbb{N}, \quad Z_{n+1} = \sum_{j=1}^{\phi_n(Z_n)} X_{nj}, \quad n \geq 0, \tag{1}$$

where $\{X_{nj} : n = 0, 1, \ldots; j = 1, 2, \ldots\}$ and $\{\phi_n(k) : n, k = 0, 1, \ldots\}$ are two independent families of non–negative integer–valued random variables. Moreover, X_{nj}, $n = 0, 1, \ldots$; $j = 1, 2, \ldots$, are independent and identically distributed random variables and for each $n = 0, 1, \ldots$, $\{\phi_n(k)\}_{k\geq 0}$ are independent stochastic processes with equal one–dimensional probability distributions. The empty sum in (1) is defined to be 0. Let $\{p_k\}_{k\geq 0}$ denote the common probability distribution of the random variables X_{nj}, i.e. $p_k = P(X_{nj} = k)$, $k \geq 0$, and $m = E[X_{nj}]$ (assumed finite). Let also $\varepsilon(k) = E[\phi_n(k)]$, $k \geq 0$ (assumed finite for each k).

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 BGW branching process. For example, it can model the evolution of populations that, although their individuals have a supercritical offspring law (i.e. $m > 1$), they die out with probability one. Indeed, the behaviour of the CBPs depends on their mean growth rates, that is the sequence of values: $\tau_m(k) = k^{-1}E[Z_{n+1} \mid Z_n = k] = k^{-1}\varepsilon(k)m$, $k \geq 1$. A simplified framework is obtained when there exists the limit of this sequence, which is called the asymptotic mean growth rate of the process and is denoted by τ. In this case, τ determines the extinction problem and the asymptotic behaviour of a CBP. Hence the importance of making inference on the offspring mean and on the asymptotic mean growth rate. However, up to now there are only few papers devoted to this subject. A first approach from a Bayesian standpoint was considered in Martínez et al. (2009) in a parametric context and in González et al. (2012) for the particular case of a deterministic control function. The present paper is a continuation of this line of research by developing the inferential theory from a non–parametric framework for the offspring law and in a parametric setting for the control distributions, depending on a single parameter, called control parameter. We address the inference of the control parameter, of the offspring distribution and of the offspring mean, as well as the asymptotic mean growth rate. To this end, Section 2 begins by assuming that the entire family tree up to some given generation can be observed. A Dirichlet process is introduced to model the prior distribution of the offspring law avoiding assumptions on the cardinal of its support. Nowadays, in most populations, it is not possible to observe this data, and only the population size at each generation can be recorded. To deal with the Bayesian inference in this case, a Markov chain Monte-Carlo method is used, concretely the Gibbs sampler algorithm, to approximate the main parameters of interest. In the current work, the implementation of such an algorithm generalizes the results in González et al. (2008).

2. Bayesian Analysis

For the purpose of this paper, we consider a CBP with an offspring distribution $p = \{p_k, k \geq 0\}$, without assuming any knowledge about the cardinal of its support. Respect to the random control mechanism, notice that we have different probability distributions for each population size $k \geq 0$, that corresponding to

$\phi_n(k)$. Consequently, from a finite sample, it is not possible to deal with the inference problems arising from this model (at least for the control distributions) without considering that some structure is stable along time. So that we consider a parametric scheme for the control process. Concretely, we address with a CBP with the control distributions belonging to the power series family of distributions (introduced formally below).

2.1 Analysis based on the entire family tree

We consider a CBP given by (1) with control distributions belonging to the power series family of distributions, i.e. for each k,

$$P(\phi_n(k) = j) = a_k(j)\theta^j / A_k(\theta), \quad j = 0, 1, \ldots; \theta \in \Theta, \tag{2}$$

with $a_k(j)$ known nonnegative values, $A_k(\theta) = \sum_{j=0}^{\infty} a_k(j)\theta^j$ and $\Theta = \{\theta > 0 : A_k(\theta) < \infty\}$ being an open subset of $\mathbb{R}$. Moreover, we assume the following regularity condition:

$$\prod_{k \in B} A_k(\theta) = A_{\sum_{k \in B} k}(\theta), \text{ for every } B \subseteq \mathbb{N},\ \theta \in \Theta. \tag{3}$$

Hence, the control distributions in the model depend on a single parameter θ, the control parameter, and on the size of the population, say k.

Remark 1 *Condition (3) is a technical hypothesis, satisfied by a wide family of probability distributions, assumed to deduce later on (4) in a more elegant way, allowing the use of conjugate families of distributions to obtain (7).*

Remark 2 *It is worthwhile remarking that from a practical viewpoint, in most of the situations, the choice of the control process, whatever be its law belonging to the power series family, should be a prior specification based on the knowledge of the development of the population.*

We consider that the entire family tree up to the current nth generation can be observed, i.e., $\{X_{lj} : j = 1, \ldots, \phi_l(Z_l); l = 0, 1, \ldots, n-1\}$ or at least the variables $Z^*_{n,k} = \sum_{l=0}^{n-1} Z_l(k)$, where $Z_l(k) = \sum_{j=1}^{\phi_l(Z_l)} I_{\{X_{lj}=k\}}$, $k \in \mathcal{S}$, 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, and therefore $Z^*_{n,k}$ is the accumulated number up to generation n of progenitors that give rise to exactly k offspring. Let denote $\mathcal{Z}^*_n = \{Z_l(k),\ k \in \mathcal{S},\ l = 0, 1, \ldots, n-1\}$. Moreover, let us introduce the following variables $Y_n = \sum_{l=0}^{n-1} Z_l$ and $Y^*_n = \sum_{l=0}^{n-1} \phi_l(Z_l)$, i.e. Y_n and Y^*_n represent, respectively, the total number of individuals and progenitors in the population up to $(n-1)$th generation. One can deduce, using (2) and (3), that the likelihood based on the sample $\mathcal{Z}^*_n$ verifies

$$f(\mathcal{Z}^*_n \mid p, \theta) \propto \prod_{k \geq 0} p_k^{Z^*_{n,k}} \theta^{Y^*_n} / A_{Y_n}(\theta). \tag{4}$$

Taking into account (4), no restriction has been imposed on the cardinal of support of the reproduction law, which is considered unknown, and that the offspring and control distributions are independent, an appropriate conjugate class of prior distributions for (p, θ) is $\pi(p, \theta) = \pi(p)\pi(\theta)$, with $\pi(p)$ the distribution corresponding to

$$p \sim \mathrm{DP}(p(0), \alpha), \tag{5}$$

where DP denotes the Dirichlet process, being $p(0) = \{p_k(0),\ k \geq 0\}$ the base measure and α the concentration parameter, with $\alpha > 0$ and $\pi(\theta)$ the distribution given by the density

$$\varphi(a,b)^{-1}\theta^a/A_b(\theta), \tag{6}$$

with

$$\varphi(a,b) = \int_\Theta \theta^a/A_b(\theta)d\theta,$$

where $a, b \geq 0$.

Then using (4)-(6) one has that the posterior distribution

$$\pi(p,\theta \mid \mathcal{Z}_n^*) \propto \pi(p)\pi(\theta)f(\mathcal{Z}_n^* \mid p,\theta),$$

is given by

$$\pi(p,\theta \mid \mathcal{Z}_n^*) \propto \pi(p \mid \mathcal{Z}_n^*)\varphi(a+Y_n^*, b+Y_n)^{-1}\theta^{a+Y_n^*}/A_{b+Y_n}(\theta), \tag{7}$$

being $\pi(p \mid \mathcal{Z}_n^*)$ the distribution of

$$p \mid \mathcal{Z}_n^* \sim \text{DP}\left(\frac{\alpha}{\alpha+Y_n^*}p(0) + \frac{1}{\alpha+Y_n^*}\sum_{k\geq 0} Z_{n,k}^*\delta_k, \alpha+Y_n^*\right),$$

with δ_k a Dirac delta at k, $k \geq 0$. From (7), using Dirichlet process properties and considering the squared error loss function, it follows straightforwardly that the Bayes estimator for the offspring distribution and θ are, respectively:

$$\widehat{p}_k = (\alpha p_k(0) + Z_{n,k}^*)/(\alpha+Y_n^*) \quad k \geq 0,$$

and

$$\widehat{\theta} = \varphi(a+Y_n^*+1, b+Y_n)/\varphi(a+Y_n^*, b+Y_n).$$

As a consequence, one obtains that the Bayes estimator, under squared error loss, for the offspring mean based on the sample $\mathcal{Z}_n^*$ is given by

$$\widetilde{m} = (\alpha m^{(0)} + Y_n + Z_n - Z_0)/(\alpha+Y_n^*), \tag{8}$$

being $m^{(0)}$ the mean of $p(0)$.

2.2 Analysis based on the population size in each generation: Gibbs sampler

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 $\mathcal{Z}_n = \{Z_0, \ldots, Z_n\}$. Given the definition of the model, an expression of the posterior for (p,θ) after observing $\mathcal{Z}_n$ can not be displayed in a closed form. Consequently, we describe an algorithm based on the Gibbs sampler to approximate it only by observing $\mathcal{Z}_n$. To this end, it is necessary to consider the unobservable variables $Z_l(k)$, $k \geq 0$, $l = 0, 1, \ldots, n-1$ as latent variables and consider the augmented parameter vector $(p,\theta,\mathcal{Z}_n^*)$. Let $\pi(p,\theta \mid \mathcal{Z}_n)$ denote the posterior distribution of (p,θ) after observing $\mathcal{Z}_n$. We shall approximate the posterior distribution of $(p,\theta,\mathcal{Z}_n^*)$ after observing $\mathcal{Z}_n$, denoted by $\pi(p,\theta,\mathcal{Z}_n^* \mid \mathcal{Z}_n)$, and from this obtain an approximation for $\pi(p,\theta \mid \mathcal{Z}_n)$. To use the Gibbs sampler, first, it is necessary to obtain the conditional posterior distribution of (p,θ) after observing $\mathcal{Z}_n$ and $\mathcal{Z}_n^*$ denoted by $\pi(p,\theta \mid \mathcal{Z}_n, \mathcal{Z}_n^*)$ and

the conditional posterior distribution of $\mathcal{Z}_n^*$ after observing $(p, \theta, \mathcal{Z}_n)$, denoted by $f(\mathcal{Z}_n^* \mid p, \theta, \mathcal{Z}_n)$.

Taking into account that, for $l = 0, \ldots, n-1$, $Z_{l+1} = \sum_{k\geq 0} kZ_l(k)$, then $\pi(p, \theta \mid \mathcal{Z}_n, \mathcal{Z}_n^*)$ is the same as $\pi(p, \theta \mid \mathcal{Z}_n^*)$ given in (7). Let now consider $f(\mathcal{Z}_n^* \mid p, \theta, \mathcal{Z}_n)$.

It can be proved that,

$$f(\mathcal{Z}_n^* \mid p, \theta, \mathcal{Z}_n) = \prod_{l=0}^{n-1} f(Z_l(k), k \geq 0 | p, \theta, Z_l, Z_{l+1}),$$

where $f(Z_l(k), k \geq 0 | p, \theta, Z_l, Z_{l+1})$ denotes the conditional distribution of the random sequence $\{Z_l(k), k \geq 0\}$, given p, θ, Z_l, and Z_{l+1}. Now, for $z_l(k) \in \mathbb{N} \cup \{0\}$, $k \geq 0$, $l = 0, 1, \ldots, n-1$, $z_l \in \mathbb{N}$, $l = 0, \ldots, n$, satisfying the constraints $z_l = \sum_{k\geq 0} kz_{l-1}(k)$, $l = 1, \ldots, n$, and denoting $\phi_l^* = \sum_{k\geq 0} z_l(k)$,

$$\begin{aligned} &P(Z_l(k) = z_l(k),\ k \geq 0 \mid Z_l = z_l, Z_{l+1} = z_{l+1}) \\ &\quad = \frac{1}{P(Z_{l+1} = z_{l+1} \mid Z_l = z_l)} \frac{\phi_l^*!}{\prod_{k\geq 0} z_l(k)!} \prod_{k\geq 0} p_k^{z_l(k)} a_{z_l}(\phi_l^*)\, \theta^{\phi_l^*} / A_{z_l}(\theta). \end{aligned}$$

Thus, from a computational viewpoint an appropriate way to obtain a sample from $f(\mathcal{Z}_n^* \mid p, \theta, \mathcal{Z}_n)$ is the following: Given the known sample $\{z_0, \ldots, z_n\}$ and known values of θ and p we sample, for each $l = 0, 1, \ldots, n-1$, a value $\phi_l^*(z_l)$ from the distribution of the variable $\phi_l(z_l)$ given by (2). Then, for each $l = 0, 1, \ldots, n-1$, we sample a sequence $\{z_l(k),\ k \geq 0\}$, from the probabilities $\frac{\phi_l^*(z_l)!}{\prod_{k\geq 0} z_l(k)!} \prod_{k\geq 0} p_k^{z_l(k)}$, $k \geq 0$, normalized by considering the constraint $z_{l+1} = \sum_{k\geq 0} kz_l(k)$. Notice that, although the cardinal of the support of the reproduction law may be infinite, once z_{l+1} is known, only a finite number of the coordinates of sequence $\{z_l(k),\ k \geq 0\}$ is non-null. Indeed, $z_l(k) = 0$ for all $k \geq z_{l+1}$.

Once it is known how to obtain samples from the distributions $\pi(p, \theta \mid \mathcal{Z}_n, \mathcal{Z}_n^*)$ and $f(\mathcal{Z}_n^* \mid p, \theta, \mathcal{Z}_n)$, the Gibbs sampler algorithm works in the following way:

`Initialize` $l = 0$

`Generate` $p^{(0)} \sim$ `DP`$(p(0), \alpha)$

`Generate` $\theta^{(0)}$ `from (6)`

`Iterate`

$l = l + 1$

`Generate` $\mathcal{Z}_n^{*(l)} \sim f(\mathcal{Z}_n^* \mid p^{(l-1)}, \theta^{(l-1)}, \mathcal{Z}_n)$

`Generate` $(p^{(l)}, \theta^{(l)}) \sim \pi(p, \theta \mid \mathcal{Z}_n^{*(l)})$

The sequence $\{(p^{(l)}, \theta^{(l)}, \mathcal{Z}_n^{*(l)})\}_{l\geq 0}$ comprises an ergodic Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint distribution, $\pi(p, \theta, \mathcal{Z}_n^* \mid \mathcal{Z}_n)$. Several practical implementation issues must be taken into account to be successful with the sample obtained by the method described above. Common approaches to reach the equilibrium distribution as well as to reduce the autocorrelation in the sample are to choose a sufficient burn-in period, N, and to thin the output by storing only every G-th value after the burn-in period (G is known as the batch size). Thus, for a run of the sequence $\{(p^{(l)}, \theta^{(l)}, \mathcal{Z}_n^{*(l)})\}_{l\geq 0}$, we choose $Q + 1$ vectors $\{(p^{(N)}, \theta^{(N)}),\ (p^{(N+G)}, \theta^{(N+G)}),$

$\ldots, (p^{(N+QG)}, \theta^{(N+QG)})\}$. These vectors are approximately independent sampled values of the distribution $\pi(p, \theta|\mathcal{Z}_n)$ if G and N are large enough (see Tierney (1994)). Since these vectors could be affected by the initial state $(p^{(0)}, \theta^{(0)})$, we apply the algorithm T times, obtaining a final sample of length $T(Q+1)$. To determine N, G, and T in practice we shall make use of the Gelman-Rubin-Brooks and autocorrelation diagnostics (see Brooks and Gelman (1998)). From this sample one can estimate $\pi(p, \theta \mid \mathcal{Z}_n)$ and its marginal distributions, $\pi(p \mid \mathcal{Z}_n)$ and $\pi(\theta \mid \mathcal{Z}_n)$, making use of kernel density estimators. These posterior densities can be used to calculate numerically HPD credible sets for the respective parameters providing sets in which there is a high probability of finding them. In general, if $\Psi(p, \theta)$ denotes a function of offspring law and the control parameter,

$$\pi(\Psi \mid \mathcal{Z}_n) = \int \pi(\Psi \mid \mathcal{Z}_n, p, \theta)\pi(p, \theta \mid \mathcal{Z}_n)dpd\theta.$$

Using again kernel density estimators, $\pi(\Psi \mid \mathcal{Z}_n)$ can be also approximated and calculate its HPD set. This allows to make inference on m and τ.

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.

References

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

Brooks, S., Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7**, 434–455

González, M., Gutiérrez, C., Martínez, R., del Puerto, I. (2012) Bayesian inference for controlled branching processes through MCMC and ABC methodologies. *Rev. R. Acad. Cien. Serie A. Mat.* DOI 10.1007/s13398-012-0072-8

González, M., Martín, J., Martínez, R., Mota, M. (2008) Non-parametric Bayesian estimation for multitype branching processes through simulation-based methods. *Comput. Statist. Data Anal.* **52**, 1281–1291

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

Martínez, R., Mota, M., del Puerto, I. (2009) On asymptotic posterior normality for controlled branching processes with random control function. *Statistics* **43**, 367-378

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

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1762