PLISKA
STUDIA MATHEMATICA

ПЛИСКА
МАТЕМАТИЧЕСКИ
СТУДИИ

STATISTICAL ESTIMATION IN BRANCHING PROCESSES WITH BIVARIATE POISSON OFFSPRING DISTRIBUTION[*]

Ana Staneva, Vessela Stoimenova

We consider two-type branching stochastic processes with offspring distributions from the bivariate poisson distribution family. We study the parametric estimation under different sampling schemes – when the entire family tree is observed and when observations only on the generation sizes are made. We use a randomized algorithm to switch from the generation sizes scheme to family tree observations and apply it in the context of the Bayesian approach. The considered estimation methods are illustrated via a simulational and computational example.

1. Introduction

Branching processes form an important class of stochastic processes with numerous applications in different scientific and practical areas, many of them involving multitype modeling. Generally speaking, there is a number of objects, often called particles, cells, individuals, which, according to some probabilistic law, reproduce (or "branch") and die out. They can be of multiple types and may have different locations in space. Their evolution and generation may be independent or according to certain probabilistic laws.

One of the pioneering works on the formulation and handling of branching processes with several types of particles are the papers from 1947 by Kolmogorov and Dmitriev[10] and Kolmogorov and Sevastyanov[11] in the Markov case. Since then there is an impressive number of work in the area of branching processes theory and applications (see f.e. the books of Asmussen and Herring[1], Athreya and Ney[2], Harris[7], Jagers[9], Sevastyanov[12], Yakovlev and Yanev[21] and others).

Statistical estimation of the process' characteristics like the mean number of offspring, the criticality of the process, the offspring distribution and others, is an important issue in their study. Some of the most resent approaches devoted to the statistical inference for branching processes can be found in González et al.[6]. The work of Jacob[8] gives a comprehensive overview of the theoretical and statistical methods used in epidemiology. The importance of simulation, computing and more flexible statistical procedures can also be traced in González et al.[5]. Results on the asymptotics of the relative frequencies in multitype branching processes and applications in cell biology can be traced in Yakovlev and Yanev[22],[23] and Yakovlev et al[20].

As in other fields of statistics, there are different approaches for estimation – parametric, nonparametric and semiparametric settings. The parametric approach uses the exact offspring distribution in a specified parametric family like the multivariate power series. The estimation in the multivariate power series family is of inerest in itself. In the class of the univariate power series offspring distributions some topics of the parametric estimation are considered in Stoimenova and Yanev[19] and of the robust parametric estimation – in Stoimenova[17]. In the present paper we consider the maximum likelihood estimation for two-type branching processes with bivariate Poisson offspring distribution for both types of particles. Considering the task of generating the whole family tree based on data over the size of generations, we derive an interesting result related to the properties of the Poisson distribution.

1.1.   Definition and sampling schemes

Let us suppose that our model describes the evolution of a population with two types of particles, which reproduce independently of each other. Each particle of, say, type 1 may have a number of offspring of type 1 and of type 2 according to some bivariate probabilistic law.

Let us denote by $T = \{1, 2\}$ the set of particle types. Let $Z_i(t)$, $i = 1, 2$, $t = 0, 1, 2, \ldots$, be the number of particles of type i in generation t. Let $\xi_k^j(t, l)$ be a r.v., representing the number of offspring of type j, $j \in T$, in the generation

$t+1$, produced from the l-th particle of type k in the generation t. Each particle, say the l-th particle of type $k \in T$ living in the t-th generation ($t = 0, 1, 2, \dots$), is associated with a random vector $\overrightarrow{\xi}_k(t, l) = (\xi_k^1(t, l), \xi_k^2(t, l))$. The distribution of the random vector $\overrightarrow{\xi}_k(t, l)$ does not depend on the generation, where the parent particle lives, and on the index l. The offspring of the particles in the generation t forms the next generation $t + 1$. We denote by $\{p_{ij}^k\}$,

$$p_{ij}^k = P(\overrightarrow{\xi}_k(t, l) = (i, j)) = P(\xi_k^1(t, l) = i, \xi_k^2(t, l) = j),$$

$k = 1, 2; \ t = 0, 1, \dots, \ l = 1, 2, \dots, Z(t-1)$, the bivariate joint distribution (offspring distribution, offspring law) of the vector $\overrightarrow{\xi}_k(t, l)$.

A multitype disctere time branching process (MBP) $\mathbf{Z}(t)$ is defined as a sequence of random vectors

$$\{\mathbf{Z}(t) = (Z_1(t), Z_2(t))\},$$

$t \in N_0 = \{0, 1, 2, \dots\}$, where $Z_k(t)$, the number of particles of type $k \in T$ in generation t, satisfies the following recursive equations (the branching property):

$$Z_k(t+1) = \sum_{j=1}^{2} \sum_{l=1}^{Z_j(t)} \xi_j^k(t, l).$$

The main properties of the MBP processes have been thoroughly studied in many sources (see f.e. [1], [2], [7],[9] and others). One of the main problems considered in the study of a given type of a branching process is to determine the asymptotic behaviour of the process – whetheter it goes extinct or has an unlimited growth. The irreducible MBP processes are divided in three classes: subcritical, critical or supercritical, according to the magnitude of their real maximum-modulus eigenvalue ρ of the mean matrix $M = \{m_{ij}\}$, whose elemens m_{ij} are the mean numbers of offspring of type j of a descendant of type i. The extinction occurs with probability 1 iff the process is subcritical ($\rho < 1$) or critical ($\rho = 1$). Otherwise (in the supercritical case, when $\rho > 1$) it grows exponentially. The estimation of the the subcritical or critical MBP faces many difficulties because of the limited (small) amount of data which may be observed especially in cases of a rapid extinction.

Usually three sampling schemes are used for the estimation of the MBP: Let us denote by

$$\widetilde{\widetilde{\Im}}(n) = \{\xi_j^k(t, l), \ k, j = 1, 2, \ t = 0, 1, \dots, n-1, \ l = 1, 2, \dots, Z_j(t)\},$$

the observations over the entire family tree (one can observe the number of off-spring of each particle untill the generation n).

There are situations, where one can observe only the number of particles with an equal number of a given offspring. We denote this set by

$$\widetilde{\Im}(n) = \{Z_k(s,(i,j)), s = 0, 1, \ldots, n-1, \ k = 1, 2, \ (i,j) \in \Im_k\},$$

where

$$Z_k(s,(i,j)) = \sum_{h=1}^{Z_k(s)} I\{(\overrightarrow{\xi}_k(s,h) = (i,j)\}$$

is the number of particles of type k in generation s with i offspring of type 1 and j offspring of type 2, $\Im_k$ is the set of all possible values of the random vector $\overrightarrow{\xi}_k(s,l) = (\xi_1^k(s,l), \xi_2^k(s,l))$, $I(\cdot)$ is the indicator variable.

Finally, let

$$\Im(n) = \{\mathbf{Z}(0), \ldots, \mathbf{Z}(n)\}$$

be the sample over the generation sizes.

1.2. Multivariate power series distributions

Let the offspring distributions be defined by:

$$(1) \qquad p_{ij}^k = \frac{a_k(i,j)\theta_{1k}^i \theta_{2k}^j}{A_k(\theta_{1k}, \theta_{2k})},$$

where $k = 1, 2$ is the particle type, $(i,j) \in \Im_k$ is in the set of possible number of offspring of type 1 and 2, $\theta_{1k}, \theta_{2k} \in \Theta_k \subset R^+$ are unknown parameters, $a_k(i,j) > 0$ and

$$(2) \qquad A_k(\theta_{1k}, \theta_{2k}) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_k(i,j)\theta_{1k}^i \theta_{2k}^j < \infty.$$

We recall that the family of distributions having probability mass functions of the form (1) is called bivariate power series distribution family (bivariate PSOD). The function $A_k(\theta_{1k}, \theta_{2k})$ is called the defining function of the distribution. Note that the form (2) is the second-order Taylor expansion of the scalar-valued function of more than one variable $A_k(\theta_{1k}, \theta_{2k})$ in a bivariate power series form. The coefficient $a_k(i,j)$ in the expansion is called the coefficient function.

The bivariate power series distribution family in (1) is a natural generalization of the univariate power series distribution family and a subclass of the multivariate power series distribution family. There are many sources concerning the properties and applications of the multivariate power series distributions. Among them we mention the pioneering papers of Katri[14], Patil[16], Gerstenkorn[4] and the thorough books on discrete multivariate distributions of Johnson et.al.[13] and discrete bivariate distributions of Kocherlakota[15]

The multivariate PSOD form a subclass of the multivariate discrete exponential family, hence inheriting its properties for the moments, cumulants, covarances, additiveness and so on.

One of the commonly used distributions of this class of bivariate distributions are the trinomial distributions (the positive multinomial), the negative binomial distributions (negative multinomial in the bivarate case), the logarithmic series distributions and the multivarate Poisson (double Poisson) distributions. The latter is of a special interest in this paper due to its properties, described in the next Section 2.1. The double Poisson distribution has the following representation:

$$A_k(\theta_{1k}, \theta_{2k}) = \exp\{\theta_{1k} + \theta_{2k}\}, \ \ \theta_{ik} > 0,$$

$$a_k(i, j) = 1/i!j!,$$

hence

$$p_{ij}^k \ \ = \ \ \frac{a_k(i,j)\theta_{1k}^i \theta_{2k}^j}{\exp\{\theta_{1k} + \theta_{2k}\}} = \frac{\theta_{1k}^i e^{-\theta_{1k}}}{i!} \frac{\theta_{2k}^j e^{-\theta_{2k}}}{j!}.$$

Further on we keep the following assumptions: the distributions of the particles of type 1 and 2 are of the bivariate PSOD family and in particular from the double Poisson distribution. Our goal is to estimate the four parameters, arising from the two distributions.

2. Maximum likelihood estimation

The first step for more complex analyses like the Bayesian and the trimmed estimation is to find the form of the likelihood function and the maximum likelihood estimator (mle).

2.1. The entire family tree scheme

When the entire family tree is observed, the likelihood function has the form

$$L(\widetilde{\widetilde{\Im}}(n)|\theta_{1k}, \theta_{2k}, k = 1, 2) = L(\widetilde{\Im}(n)|\theta_{1k}, \theta_{2k}, k = 1, 2) =$$

$$= \prod_{k=1}^{2} \prod_{s=1}^{n-1} \prod_{(i,j)\in\Im_k} p_{ij}^k = \prod_{k=1}^{2} \prod_{s=1}^{n-1} \prod_{(i,j)\in\Im_k} \left[\frac{a_k(i,j)}{A_k(\theta_{1k},\theta_{2k})} \theta_{1k}^i \theta_{2k}^j \right]^{Z_k(s,(i,j))}$$

Due to the properties of the power series moments , f.e.

$$E\xi_k^1(s,l) = \frac{\theta_{1k}}{A_k(\theta_{1k},\theta_{2k})} \frac{\partial A_k(\theta_{1k},\theta_{2k})}{\partial \theta_{1k}},$$

one derives the following Harris type estimators for the mean number of offspring of a given type i from a father of type k

$$\widehat{m}_{ik} = E\xi_k^i(t,l) = \frac{\sum\limits_{s=1}^{n} Z_i^k(s)}{\sum\limits_{n=0}^{n-1} Z_k(n)},$$

where $Z_i^k(s)$ is the number of children of type i in generation s+1, whose father is of type k.

Due to the independence of evolutions one can consider the likelihood function $L(\widetilde{\Im}(n)|\theta_{1k},\theta_{2k},k=1,2)$ as a product of two likelihood functions, depending on different parameters:

$$
\begin{aligned}
L(\widetilde{\Im}(n)|\theta_{1k},\theta_{2k},k=1,2) &= \prod_{s=1}^{n-1} \prod_{(i,j)\in\Im_1} \left[\frac{a_1(i,j)}{A_1(\theta_{11},\theta_{21})} \theta_{11}^i \theta_{21}^j \right]^{Z_1(s,(i,j))} \times \\
&\times \prod_{s=1}^{n-1} \prod_{(i,j)\in\Im_2} \left[\frac{a_2(i,j)}{A_2(\theta_{12},\theta_{22})} \theta_{12}^i \theta_{22}^j \right]^{Z_2(s,(i,j))} = \\
&= L(\widetilde{\Im}(n)|\theta_{11},\theta_{21}) L(\widetilde{\Im}(n)|\theta_{12},\theta_{22})
\end{aligned}
$$

(3)

2.2.   The generation sizes scheme

Let us suppose, that we are able to observe the generation sizes only. The observations form the set

$$\Im(n) = \{\mathbf{Z}(0),\ldots,\mathbf{Z}(n)\}$$

The following statements are valid (see González et al.[5]):

$$P(\widetilde{\Im}(n)|\Im(n),\theta_{11},\theta_{21}) = \prod_{s=0}^{n-1} P(\widetilde{\Im}(s)|\mathbf{Z}(s),\mathbf{Z}(s+1),\theta_{11},\theta_{21})$$

and

$$P(\widetilde{\Im}(n)|\Im(n),\theta_{11},\theta_{21}) = \prod_{s=0}^{n-1} P(\widetilde{\Im}(s)|\mathbf{Z}(s),\mathbf{Z}(s+1),\theta_{11},\theta_{21}).$$

Applying the idea in [5] in the situation of power series with finite support $|\Im_1|,|\Im_2| < \infty$ one can use the multinomial distribution to 'reconstruct' the family tree:

$$P(\widetilde{\Im}(s)|\mathbf{Z}(s),\mathbf{Z}(s+1),\theta_{11},\theta_{21}) = \frac{\prod\limits_{k=1,2} \frac{[Z_k(s)]!}{\prod\limits_{(i,j)}[Z_k(n,(i,j))]!} \prod\limits_{(i,j)}[p_{ij}^k]^{Z_k(s,(i,j))}}{P(\mathbf{Z}(s+1)|\mathbf{Z}(s))}$$

Note that the family tree is reconstructed in 'strips', using every two consecutive generation sizes.

When the support of the distribution is not finite, i.e. $|\Im_1|,|\Im_2| = \infty$, for some distributions (in particular for the double Poisson distribution) one can directly compute the expression

$$P(\widetilde{\widetilde{\Im}}(s)|\mathbf{Z}(s),\mathbf{Z}(s+1),\theta_{11},\theta_{21}) = \frac{\prod\limits_{k=1,2} \prod\limits_{l=1}^{Z_k(s)} p_{\xi_1^k(s,l),\xi_2^k(s,l)}^k}{P(\mathbf{Z}(s+1)|\mathbf{Z}(s))}.$$

In the case of double Poisson offspring disribution it can easily be seen that

$$P(\widetilde{\Im}(n)|\mathbf{Z}(n),\mathbf{Z}(n+1),\theta_{11},\theta_{21}) =$$

$$= \left[\frac{Z_1(n+1)!}{\left[\sum_{l=1}^{Z_1(n)}\xi_1^1(n,l)\right]!\left[\sum_{l=1}^{Z_2(n)}\xi_2^1(n,l)\right]!}\left[\frac{Z_1(n)\theta_{11}}{Z_1(n)\theta_{11}+Z_2(n)\theta_{21}}\right]^{\sum_{l=1}^{Z_1(n)}\xi_1^1(n,l)}\left[\frac{Z_2(n)\theta_{21}}{Z_1(n)\theta_{11}+Z_2(n)\theta_{21}}\right]^{\sum_{l=1}^{Z_2(n)}\xi_2^1(n,l)}\right]$$

$$\times\left[\frac{Z_2(n+1)!}{\left[\sum_{l=1}^{Z_1(n)}\xi_1^2(n,l)\right]!\left[\sum_{l=1}^{Z_2(n)}\xi_2^2(n,l)\right]!}\left[\frac{Z_1(n)\theta_{12}}{Z_1(n)\theta_{12}+Z_2(n)\theta_{22}}\right]^{\sum_{l=1}^{Z_1(n)}\xi_1^2(n,l)}\left[\frac{Z_2(n)\theta_{22}}{Z_1(n)\theta_{12}+Z_2(n)\theta_{22}}\right]^{\sum_{l=1}^{Z_2(n)}\xi_2^2(n,l)}\right]$$

$$\times\left[\frac{\left[\sum_{l=1}^{Z_1(n)}\xi_1^1(n,l)\right]!}{\prod_{l=1}^{Z_1(n)}\xi_1^1(n,l)!}\left[\frac{1}{Z_1(n)}\right]^{\sum_{l=1}^{Z_1(n)}\xi_1^1(n,l)}\right]\times\left[\frac{\left[\sum_{l=1}^{Z_2(n)}\xi_2^1(n,l)\right]!}{\prod_{l=1}^{Z_2(n)}\xi_2^1(n,l)!}\left[\frac{1}{Z_2(n)}\right]^{\sum_{l=1}^{Z_2(n)}\xi_2^1(n,l)}\right]\times$$

$$\times\left[\frac{\left[\sum_{l=1}^{Z_1(n)}\xi_1^2(n,l)\right]!}{\prod_{l=1}^{Z_1(n)}\xi_1^2(n,l)!}\left[\frac{1}{Z_1(n)}\right]^{\sum_{l=1}^{Z_1(n)}\xi_1^2(n,l)}\right]\times\left[\frac{\left[\sum_{l=1}^{Z_2(n)}\xi_2^2(n,l)\right]!}{\prod_{s=1}^{Z_2(n)}\xi_2^2(s)!}\left[\frac{1}{Z_2(n)}\right]^{\sum_{l=1}^{Z_2(n)}\xi_2^2(n,l)}\right]$$

Note that this expression consists of the multiplied probabilities from six multinomial distributions, hence giving us the following method for simulation of the offspring numbers in each generation:

1. Generate the number of type 1 particles $\sum_{l=1}^{Z_1(n)} \xi_1^1(n,l)$ and $\sum_{l=1}^{Z_2(n)} \xi_2^1(n,l)$ in generation $n+1$, whose ancestors in generation n is of type 1 and type 2 correspondingly, using the multinomial (in our situation of a two-type process this is in fact binomial) distribution with $Z_1(n+1)$ number of trials and event probabilities $\frac{Z_1(n)\theta_{11}}{Z_1(n)\theta_{11}+Z_2(n)\theta_{21}}$ and $\frac{Z_2(n)\theta_{21}}{Z_1(n)\theta_{11}+Z_2(n)\theta_{21}}$. Note that the following relation holds:

$$Z_1(n+1) = \sum_{l=1}^{Z_1(n)} \xi_1^1(n,l) + \sum_{l=1}^{Z_2(n)} \xi_2^1(n,l)$$

2. Generate the number of type 2 particles $\sum_{l=1}^{Z_1(n)} \xi_1^2(n,l)$ and $\sum_{l=1}^{Z_2(n)} \xi_2^2(n,l)$ in generation $n+1$, whose ancestors in generation n are of type 1 and 2 respectively, using the multinomial distribution with $Z_2(n+1)$ number of trials and event probabilities $\frac{Z_1(n)\theta_{12}}{Z_1(n)\theta_{12}+Z_2(n)\theta_{22}}$ and $\frac{Z_2(n)\theta_{22}}{Z_1(n)\theta_{12}+Z_2(n)\theta_{22}}$. Again it is true that

$$Z_2(n+1) = \sum_{l=1}^{Z_1(n)} \xi_1^2(n,l) + \sum_{l=1}^{Z_2(n)} \xi_2^2(n,l)$$

3. Using the generated number $\sum_{l=1}^{Z_1(n)} \xi_1^1(n,l)$, generate the number of offspring $\xi_1^1(n,l)$, $l = 1, 2, \ldots Z_1(n)$ for each of the $Z_1(n)$ particles of type 1, whose offspring is of type 1, using the multinomial distribution with $\sum_{l=1}^{Z_1(n)} \xi_1^1(n,l)$ trials and $Z_1(n)$ events with equal event probabilities $\frac{1}{Z_1(n)}$.

4. Using the generated number $\sum_{l=1}^{Z_2(n)} \xi_2^1(n,l)$, generate the number of offspring $\xi_2^1(n,l)$, $l = 1, 2, \ldots Z_2(n)$ for each of the $Z_2(n)$ particles of type 1, whose offspring is of type 1, using the multinomial distribution with $\sum_{l=1}^{Z_2(n)} \xi_2^1(n,l)$ trials and $Z_2(n)$ events with equal event probabilities $\frac{1}{Z_2(n)}$.

5. Using the generated number $\sum_{l=1}^{Z_1(n)} \xi_1^2(n,l)$, generate the number of offspring $\xi_1^2(n,l)$, $l = 1, 2, \ldots Z_1(n)$ for each of the $Z_1(n)$ particles of type 1, whose offspring is of type 2, using the multinomial distribution with $\sum_{l=1}^{Z_1(n)} \xi_1^2(n,l)$ trials and $Z_1(n)$ events with equal event probabilities $\frac{1}{Z_1(n)}$.

6. Using the generated number $\sum_{l=1}^{Z_2(n)} \xi_2^2(n,l)$, generate the number of offspring $\xi_2^2(n,l)$, $l = 1, 2, \ldots Z_2(n)$ for each of the $Z_2(n)$ particles of type 2, whose offspring is of type 2, using the multinomial distribution with $\sum_{l=1}^{Z_2(n)} \xi_2^2(n,l)$ trials and $Z_2(n)$ events with equal event probabilities $\frac{1}{Z_2(n)}$.

This partition in the $n + 1$-st generation is based on the known generation sizes $Z_1(n)$, $Z_2(n)$, $Z_1(n+1)$ and $Z_2(n+1)$ and the estimated parameters (for instance via the coinsiding Harris estimators $\widehat{m}_{ik}$ of the mean numbers). Finding the value of the parameters θ_{ik}, $i, k = 1, 2$ is an important part of the generation of the family tree. There are different approaches to simultaneously solve the two problems - to generate the family tree on the basis of the generation sizes and to estimate the unknown parameters. Sometimes, especially when the process starts with a small number of particles, one is able to observe the family tree for the first two or three generations and hence to estimate as a starting point the mean numbers of offspring for both particle types. In cases when the prior distribution of the parameters can be appropriately chosen, one can generate their values. One of the most refined iterative method used in the context of the Bayesian framework is the Gibbs sampler, proposed by Gonzaléz et al. [5] to alternatively generate the family tree on the basis of the current value of the parameters and generate new parameter values using the posterior distribution conditional on the family tree.

3. The Bayesian estimation

Since the likelihood function is proportional to

$$L \propto \prod_{k=1}^{2} \theta_{1k}^{\sum_{n=0}^{N-1} Z_1^k(n+1)} e^{-\theta_{1k} \sum_{n=0}^{N-1} Z_k(n)} \theta_{2k}^{\sum_{n=0}^{N-1} Z_2^k(n+1)} e^{-\theta_{2k} \sum_{n=0}^{N-1} Z_k(n)},$$

we choose the conjugate prior to be the joint distribution of four independent Gamma distributed random variables:

$$\theta_{1k} \sim Gamma(\alpha_{1k}, \beta_{1k}), \qquad \theta_{2k} \sim Gamma(\alpha_{2k}, \beta_{2k}), \quad k = 1, 2.$$

Hence the posterior distribution has the form

$$f(\theta|\tilde{\mathcal{J}}(n), \alpha, \beta) \propto \prod_{k=1}^{2} \theta_{1k}^{\sum_{n=0}^{N-1} Z_1^k(n+1)} e^{-\theta_{1k} \sum_{n=0}^{N-1} Z_k(n)} \theta_{2k}^{\sum_{n=0}^{N-1} Z_2^k(n+1)} e^{-\theta_{2k} \sum_{n=0}^{N-1} Z_k(n)}$$

$$\times \frac{\beta_{1k}^{\alpha_{1k}}}{\Gamma(\alpha_{1k})} \theta_{1k}^{\alpha_{1k}-1} e^{-\beta_{1k}\theta_{1k}} \frac{\beta_{2k}^{\alpha_{2k}}}{\Gamma(\alpha_{2k})} \theta_{2k}^{\alpha_{2k}-1} e^{-\beta_{2k}\theta_{2k}} \propto$$

$$\propto \prod_{k=1}^{2} \theta_{1k}^{\sum_{n=0}^{N-1} Z_1^k(n+1)+\alpha_{1k}-1} e^{-\theta_{1k}(\sum_{n=0}^{N-1} Z_k(n)+\beta_{1k})} \theta_{2k}^{\sum_{n=0}^{N-1} Z_2^k(n+1)+\alpha_{2k}-1} e^{-\theta_{2k}(\sum_{n=0}^{N-1} Z_k(n)+\beta_{2k})}$$

The posterior distribution can be considered as the joint distribution of four independent Gamma distributed r.v. with

$$Gamma(\sum_{n=0}^{N-1} Z_1^k(n+1) + \alpha_{1k}, \sum_{n=0}^{N-1} Z_k(n) + \beta_{1k}), \quad k = 1, 2.$$

and

$$Gamma(\sum_{n=0}^{N-1} Z_2^k(n+1) + \alpha_{2k}, \sum_{n=0}^{N-1} Z_k(n) + \beta_{2k}), \quad k = 1, 2.$$

From here one can calculate the posterior maximum lielihood estimators of the parameters and the posterior mean and variance.

3.1. A numerical example

In this section we generate a 'slightly' supercritical two-type branching process with eigenvalue of 1.005 and two bivariate poisson offspring distributions with parameters $\theta_{11} = 0.35$, $\theta_{12} = 0.45$, $\theta_{21} = 0.06$ and $\theta_{22} = 0.98$.

We choose the parameters of the prior distribution to be $\alpha_i = 1$, $\beta_i = 1.5$.

For the simulations we use the free software environment for statistical computing and graphics R.

On Table 1 the simulated generation sizes for the two particle types are shown. They are plotted on Figure 1.

We use the following procedure to generate the family tree, simultaneousely updating the values of the parameter estimates:

Table 1. Generation
sizes

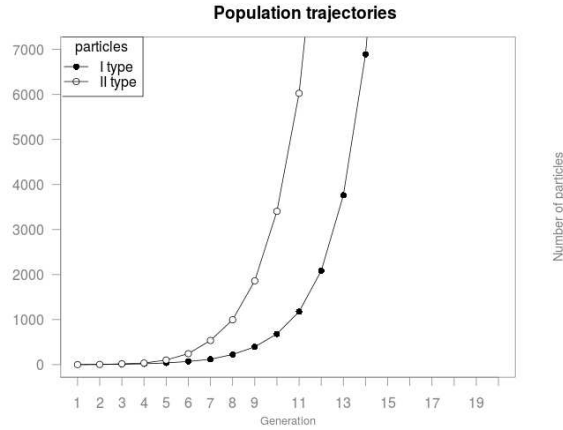| n	$Z_1(n)$	$Z_2(n)$
1	1	1
2	4	4
3	8	18
4	18	35
5	38	102
6	70	243
7	118	536
8	224	996
9	393	1861
10	682	3405
11	1176	6026
12	2082	10691
13	3762	18995
14	6892	34116
15	12547	61479
16	22758	110831
17	41142	200920
18	74796	363314
19	135482	658797
20	245113	1194947

Figure 1. The sample trajectory of the process

Step 0: Initialize $n = 0$ generation.

Step 1: Fix the initial values of the parameters $\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22}$ as generated numbers from the prior distribution

Step 2: Iterate on the generations

- $n = n + 1$ moving to the next generation
- For $k = 1, 2$ generate the partition of the generation sizes using the multinomial distribution

$$Multinomial\left(Z_k(n+1), \left\{ \frac{Z_1(n)\theta_{k1}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}}, \frac{Z_2(n)\theta_{k2}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}} \right\} \right)$$

- Generation of the new values of the parameters from the posterior density.

Step 3: Go to Step 2.

On Step 1 of the algorithm we need to find the event probabilities

$$p_1^k = \frac{Z_1(n)\theta_{k1}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}} \quad p_2^k = \frac{Z_2(n)\theta_{k2}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}}$$

from the multinomial distribution

$$Multinomial\left(Z_k(n+1), \left\{\frac{Z_1(n)\theta_{k1}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}}, \frac{Z_2(n)\theta_{k2}}{Z_1(n)\theta_{k1} + Z_2(n)\theta_{k2}}\right\}\right), k = 1, 2.$$

They can be seen in Table 2 together with the generated quantities $Z_1^1(n+1)$, $Z_2^1(n+1)$, $Z_1^2(n+1)$, $Z_2^2(n+1)$:

Table 2. Offspring generations

$Z_1(n)$	$Z_2(n)$	$Z_1^1(n+1)$	$Z_2^1(n+1)$	$Z_1^2(n+1)$	$Z_2^2(n+1)$	p11	p21	p12	p22
1	1								
4	4	3	1	0	4	0.9859628282	0.0140371718	0.7450198119	0.2549801881
8	18	8	0	5	13	0.64741091	0.35258909	0.0025112333	0.9974887667
18	35	16	2	14	21	0.4608802939	0.5391197061	0.0540584266	0.9459415734
38	102	31	7	32	70	0.6688375056	0.3311624944	0.1695060125	0.8304939875
70	243	54	16	64	179	0.6616636055	0.3383363945	0.1406046427	0.8593953573
118	536	102	16	153	383	0.6197203556	0.3802796444	0.1070917754	0.8929082246
224	996	196	28	352	644	0.4951972343	0.5048027657	0.083225946	0.916774054
393	1861	337	56	624	1237	0.5372199758	0.4627800242	0.097926924	0.902073076
682	3405	596	86	1095	2310	0.5440858083	0.4559141917	0.0903006036	0.9096993964
1176	6026	1031	145	2000	4026	0.5523584683	0.4476415317	0.0886432134	0.9113567866
2082	10691	1803	279	3452	7239	0.5584014987	0.4415985013	0.0850290894	0.9149709106
3762	18995	3261	501	6036	12959	0.5773197397	0.4226802603	0.0860027228	0.9139972772
6892	34116	5988	904	10931	23185	0.5660859183	0.4339140817	0.0861329229	0.9138670771
12547	61479	10902	1645	19620	41859	0.5703392073	0.4296607927	0.0874802642	0.9125197358
22758	110831	19713	3045	35411	75420	0.5761959394	0.4238040606	0.0882114478	0.9117885522
41142	200920	35615	5527	63939	136981	0.5733560978	0.4266439022	0.0883937578	0.9116062422
74796	363314	64700	10096	116227	247087	0.5708433366	0.4291566634	0.0873338852	0.9126661148
135482	658797	117377	18105	211376	447421	0.5710974723	0.4289025277	0.0884341078	0.9115658922
245113	1194947	212283	32830	383517	811430	0.568764931	0.431235069	0.0883023592	0.9116976408

Z_1(19,(x,y))
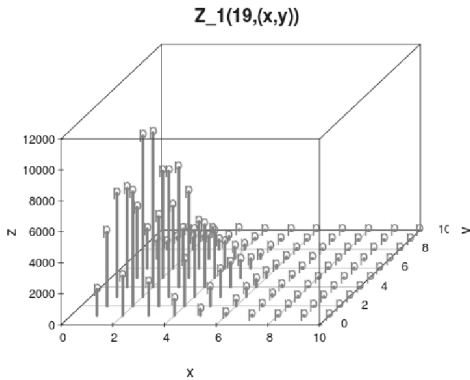

Z_2(19,(x,y))
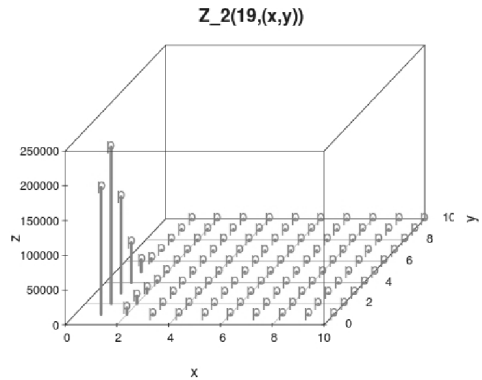
Figure 2   Figure 3

On Step 2 one can find the size of the groups with the same offspring in the given generation, using the four multinomial distributions $Multinomial\left(\sum_{s=1}^{Z_1(n)} X_{1s}^k, \left\{\frac{1}{Z_1(n)}\right\}\right)$ and $Multinomial\left(\sum_{s=1}^{Z_2(n)} X_{2s}^k, \left\{\frac{1}{Z_2(n)}\right\}\right)$, $k = 1, 2$. The generated distributions of the number of particles in the 19-th generation with (x, y) offspring in the last generation 20 are shown on Figures 2 and 3.

The estimated values of the parameters $\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22}$ after each step of the procedure are shown in Table 3.

Table 3. Estimated values of the parameters after each step of the algorithm

θ_{11}	θ_{21}	θ_{12}	θ_{22}
1.692343	3.311847	0.02409392	1.133467
1.2433578393	0.0040156536	0.677150172	1.5950606456
0.5442331297	0.1323649615	0.2829433719	1.0294172253
0.6405153977	0.4188438742	0.1631002256	1.0553779448
0.4959341094	0.3746628203	0.0944754293	0.8531342249
0.5422815116	0.3747062416	0.0958570108	0.8999811194
0.444306513	0.3964146992	0.0997110782	0.9613256732
0.3926370388	0.4562282658	0.0760680699	0.9451711923
0.3698688852	0.4476836331	0.0654500057	0.9524115377
0.3601579657	0.4667649357	0.0584614576	0.961188183
0.3515466355	0.4599826581	0.0542552983	0.9659605985
0.360354812	0.4622511379	0.0513793115	0.956694078
0.3487456345	0.4569921	0.0529431364	0.9602891736
0.3489409117	0.4571277641	0.0531045279	0.9632897579
0.3540131627	0.4566339075	0.0531407199	0.9632753467
0.3559856268	0.4554843365	0.0543934257	0.964566071
0.3557574092	0.4488779472	0.0547665153	0.9605492678
0.3571897688	0.4525538575	0.0552260037	0.9603615836
0.3572799833	0.452945235	0.0557083034	0.9617323071
0.3571188151	0.4531998739	0.0554473883	0.9614578868

The convergence of the estimates with the generation is shown on Figure 4.

The obtained results show that the "reconstruction" of the family tree on the basis of the generation sizes can easily and effective be done using realizations of multinomial random variables. The estimation of the parameters shows that the procedure stabilizes after a small number of generations. The most delicate point in this setting is the choice of the initial parameters of the prior distribution. In fact they may affect the final estimates, but in the case of a supercritical process their influence is reduced in the last generations due to the huge generation sizes.
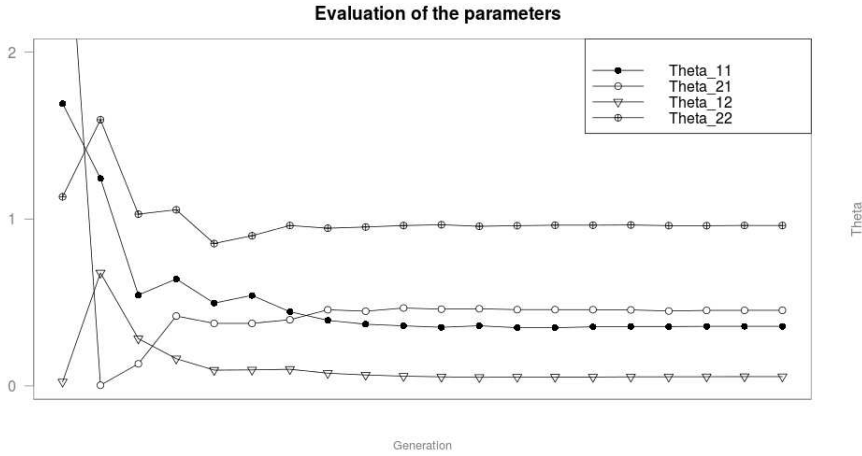
Figure 4. The behaviour of the estimates according to the generation numbers

REFERENCES

[1] S. Asmussen, H. Herring. Branching Processes. Birkhauser, Boston, 1983.

[2] K. B. Athreya, P. E. Ney. Branching Processes. Springer-Verlag, Berlin 1972.

[3] D. Atanasov, V. Stoimenova. Simulation, Estimation and Robustification against Outliers in Branching Processes – a Computational Approach. *Proceedings COMPSTAT*, 2012, 53–65.

[4] T. Gerstenkorn. On multivariate power series distributions. *Revue Roumaine de Mathématiques Pures et Appliquées*, **26** (1981), 247–266.

[5] M. González, J. Martín, R. Martínez, M. Mota. Non-parametric Bayesian estimation for multitype branching processes through simulation-based methods. *Computational statistics & data analysis*, **52** (2008), 1281–1291.

[6] M. González, I. M. del Puerto, R. Martínez, M. Mota. Workshop on Branching Processes and Their Applications. Lecture Notes in Statistics, vol. **197**, Springer Verlag, 2010.

[7] T. E. Harris. The Theory of Branching Processes. Springer-Verlag, Berlin, 1963.

[8] C. JACOB. Branching processes: their role in epidemiology. *Int. J. Environ. Res. Public Health* **7** (2010), 11861204.

[9] P. JAGERS. Branching Processes with Biological Applications. Wiley, London, 1975.

[10] A. N. KOLMOGOROV, N. A. DMITRIEV. Branching random processes. *Dokl. Akad. Nauk (Proc. Acad. Sci. USSR)*, **56** (1947), 7–10 (in Russian).

[11] A. N. KOLMOGOROV, B. A. SEVASTYANOV. Calculation of final probabilities of branching random processes. *Dokl. Akad. Nauk (Proc. Acad. Sci. USSR)*, **56** (1947), 783–786 (in Russian).

[12] B. A. SEVASTYANOV. Branching Processes, Nauka, Moscow, 1971 (in Russian).

[13] N. JOHNSON, S. KOTZ, N. BALAKRISHTNAN. Discrete multivariate distributions. Wiley Series in Probability and Statistics. New York, 1997.

[14] C. G. KHATRI. On multivariate contagious distributions. *Sankhyia*, Series B, **33** (1971), 197–216.

[15] S. KOCHERLAKOTA, K. KOCHERLAKOTA. Bivariate discrete distributions, Marcel Dekker Inc., New York, 1992

[16] G. P. PATIL. On multivariate generalized power series distribution and its applications to the multinomial and negative multinomial. In: Clussical und Contugious Discrete Distributions (Ed. G. P. Patil) Proceedings of International Symposium at McGill University on August 15–20, Calcutta: Statistical Publishing Society; Oxford: Pergamon Press, 1965, 183–194.

[17] V. STOIMENOVA. Robust Parametric Estimation of Branching Processes with Random Number of Ancestors. *Serdica Math. J.* **31** *3* (2005), 243–262.

[18] V. STOIMENOVA, D. ATANASOV. Robust estimation in multitype branching processes based on their asymptotic properties. *Pliska Stud. Math. Bulgar.* **21** (2011), 203–220.

[19] V. STOIMENOVA, N. YANEV. Parametric Estimation in Branching Processes with an Increasing Random Number of Ancestors. *Pliska Stud. Math. Bulgar.* **17** (2005), 295–312.

[20] A. Y. YAKOVLEV, V. K. STOIMENOVA, N. M. YANEV. Branching Processes as Models of Progenitor Cell Populations and Estimation of the Offspring Distributions. *J. American Statistical Assoc*, **103** (1998), 1357–1366.

[21] A. YU. YAKOVLEV, N. M. YANEV. Transient Processes in Cell Proliferation Kinetics. Springer Verlag, Berlin, 1989.

[22] A. YU. YAKOVLEV, N. M. YANEV. Relative frequencies in multitype branching processes. *Ann. Appl. Probab.*, **19**, *1* (2009), 1–14.

[23] A. YU. YAKOVLEV, N. M. YANEV. Limiting distributions in multitype branching processes. *Stochastic Analysis and Applications*, **28**, *6* (2010), 1040–1060.

Ana Staneva
Technical University of Sofia
Bulgaria, Sofia 1000,
blvd Kliment Ohridski 8
and
Department of Probability,
OR and Statistics,
Faculty of Mathematics
and Informatics
Sofia University
5, James Bourchier Blvd
1164 Sofia, Bulgaria
e-mail: `anastaneva@gmail.com`

Vessela Stoimenova
Department of Probability,
OR and Statistics
Faculty of Mathematics and Informatics
Sofia University
5 James Bourchier Blvd
1164 Sofia, Bulgaria
and
Bulgarian Academy of Sciences
Institute of Mathematics and Informatics
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: `stoimenova@fmi.uni-sofia.bg`