Inferential problems for a Y-linked gene branching model with blind choice

Cristina Gutiérrez
Department of Mathematics. University of Extremadura. 06006 Badajoz.
Spain. e-mail: cgutierrez@unex.es

Abstract

Estimation problems for a two-type bisexual branching process with blind choice are studied. Such a model is adequate for analyzing the evolution of Y-linked genes from generation to generation in a two-sex monogamic population, where it is assumed that females and males mate under a blind mating scheme. A non-parametric frequentist framework is considered. Moreover, it is also assumed that the only available data are the total number of females, the total number of males of each genotype and the total number of each type of couple in each generation. The problem of estimate the main parameters of the model is then tackled as an incomplete data problem and the maximum likelihood estimators for the main parameters of the model are derived using the expectation-maximization method.

Keywords: Sex-linked inheritance, bidimensional bisexual stochastic model, maximum likelihood estimators, expectation-maximization method

1. Introduction

In González et al. (2009) was introduced a stochastic model to analyze the evolution of Y-linked genes from generation to generation in a two-sex monogamic population. This model assumes that females and males mate in order to form couples under a blind mating scheme which means that females choose their mates without caring about what their genotypes are, i.e. each female makes a blind choice of the genotype of her mate. The model, which was called Y-linked Bisexual Branching Process with blind choice, is a Multitype Bisexual Branching Process in discrete time. The authors considered a perfect fidelity mating mechanism, that is, an individual may mate with no more than one individual of the opposite sex. The Y-linked genes considered in such model have two alleles R and r (r could mean simply the absence of R).

In this paper, we present first the definition of the Y-linked Bisexual Branching Process with blind choice. Then, considering a frequentist outlook in a non-parametric framework, we deal with the estimation problem. At the beginning, maximum likelihood estimators (MLEs) of the main parameters of the model are obtained assuming that we can observe the complete family tree up to some generation. Later, we assume that the observed sample is only given by the total number of females, and the total number of males and couples of each type up to some generation. The problem is considered as estimation with incomplete data and the expectation-maximization (EM) algorithm is applied.

2. The Y-linked Bisexual Branching Process with Blind Choice

Definition 1 *Let* $\{(FR_{ni}, MR_{ni}) : i = 1, 2, ...; n = 0, 1, ...\}$ *and* $\{(Fr_{nj}\ , Mr_{nj}) : j = 1, 2, ...; n = 0, 1, ...\}$ *be two independent sequences of i.i.d., non-negative and integer-valued bivariate random vectors on the same probability triple* $(\Omega, \mathcal{F}, P)$*. The following sequences of r.v.* $\{(FR_{n+1}, MR_{n+1})\}_{n\geq 0}$*,* $\{(Fr_{n+1},\ Mr_{n+1})\}_{n\geq 0}$ *and* $\{(ZR_{n+1}, Zr_{n+1})\}_{n\geq 0}$ *are defined recursively as follows: Let* $(ZR_0, Zr_0) = (a, b)$ *be, with* $a, b \in \mathbb{N}, (a, b) \neq (0, 0)$*, and assume* $\sum_1^0 = 0$*, then, for* $n = 0, 1, ...$

$$(FR_{n+1}, MR_{n+1}) = \sum_{i=1}^{ZR_n} (FR_{ni}, MR_{ni}), \quad (Fr_{n+1}, Mr_{n+1}) = \sum_{j=1}^{Zr_n} (Fr_{nj}, Mr_{nj}), \quad (1)$$

$$F_{n+1} = FR_{n+1} + Fr_{n+1}, \quad M_{n+1} = MR_{n+1} + Mr_{n+1}.$$

$$\begin{array}{llll} \textit{If } F_{n+1} \geq M_{n+1}, & \textit{then} & ZR_{n+1} = MR_{n+1} \quad \textit{and} \quad Zr_{n+1} = Mr_{n+1}. \\ \textit{If } F_{n+1} < M_{n+1}, & \textit{then} & ZR_{n+1} \sim H(F_{n+1}, M_{n+1}, MR_{n+1}),^{1} \\ & & Zr_{n+1} = F_{n+1} - ZR_{n+1}. \end{array}$$

The two-dimensional process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ *is called Y-linked Bisexual Branching Process with blind choice (Y-BBP with blind choice).*

Intuitive interpretation

Intuitively, for N fixed, the r.v. (ZR_N, Zr_N) represents, respectively, the total number of couples of type R and r (that means, the male who forms the couple has genotype R and r, resp.) at generation N. To describe the evolution of the population from this generation on, two phases are considered: reproduction and mating. In the reproduction phase, each couple, independently of the others, generates females and males according to some probability distribution depending on its type. So, (FR_{Ni}, MR_{Ni}) denotes the total number of females and males given by the i-th R couple at the generation N and analogously for (Fr_{Nj}, Mr_{Nj}). According to equation (1), we obtain the total number of females and males generated by all R (r) couples of generation N: (FR_{N+1}, MR_{N+1}) $((Fr_{N+1}, Mr_{N+1}))$. Moreover, F_{N+1} denotes the total number of females at the generation $(N + 1)$, (respectively, M_{N+1}, denotes the total number of males at the generation $(N + 1)$).

In the mating phase, it is considered perfect fidelity mating and that females choose their mates blindly, respect to the allele they carry. If the total number of females is greater than or equal to the total number of males, all males mate so the total number of couples of each type is equal to the total number of males of that type. On the other hand, if the total number of females is less than the total number of males, all females mate. Since females make a blind choice between the males, the total number of R couples, at the generation $(N+1)$, is given by a hypergeometric distribution with parameters $(F_{N+1}, M_{N+1}, MR_{N+1})$. The rest of the couples will have r genotype.

[1] $H(n, N, k)$ is the hypergeometric distribution with parameters $n, N, k \in \mathbb{N}, n \leq N$, which is the probability law of the number of red balls when drawing n balls at random without replacement from an urn containing a total number of N balls of which k are red and $N - k$ are black.

Remark 1 In the reproduction phase, in order to determine the distribution of the vectors (FR_{ni}, MR_{ni}) (resp. (Fr_{nj}, Mr_{nj})) we assume, throughout the paper, the scheme introduce in Daley (1968). So we consider the following two sequences of i.i.d., non-negative and integer-valued r.v. $\{TR_{ni} := FR_{ni} + MR_{ni} : i = 1, 2...; n = 0, 1, ...\}$ and $\{Tr_{nj} := Fr_{nj} + Mr_{nj} : j = 1, 2...; n = 0, 1, ...\}$ representing the total number of individuals given by the i-th R couple and j-th r couple, respectively at generation n, for $n = 0, 1, ...$ Taking into account that the probability distribution is the same for all the couples with a given genotype, irrespective of the generation they belong, we talk about the reproduction law associated to R genotype as $p^R = \{p^R_k\}_{k\in S^R}$, with $p^R_k = P(TR_{01} = k), k \in S^R$, being $S^R \subseteq \mathbb{Z}_+$ its support. Analogously, the reproduction law associated to r genotype is $p^r = \{p^r_l\}_{l\in S^r}$, with $p^r_l = P(Tr_{01} = l), l \in S^r$, being $S^r \subseteq \mathbb{Z}_+$ its support. Both distributions are assumed with finite variances. Moreover, m_R and m_r are the average number of individuals generated by a couple of type R and r, respectively. Now, let α $(0 < \alpha < 1)$ be the probability for an offspring of any genotype to be female, being α the same for both genotypes. These sex designations are made independently among the offspring of any couple. Then $FR_{ni}|TR_{ni} = k$ follows a Binomial distribution of parameters (k, α) $(FR_{ni}|TR_{ni} = k \sim B(k, \alpha))$. Respectively, $Fr_{nj}|Tr_{nj} = l \sim B(l, \alpha)$, $MR_{ni}|TR_{ni} = k \sim B(k, 1 - \alpha)$ and $Mr_{nj}|Tr_{nj} = l \sim B(l, 1 - \alpha)$. ∎

3. Maximum Likelihood Estimators

To guarantee the applicability of this model, it is necessary to develop its estimation theory. In this sense, we consider in this section the maximum likelihood estimation of the following parameters of a Y-BBP with blind choice: α, the probability for an individual to be female, $p^R = \{p^R_k\}_{k\in S^R}$ $(p^r = \{p^r_l\}_{l\in S^r})$, probability distribution of R (r) genotype, m_R (m_r), mean number of descendants per R (r) couple. We assume that we can observe the entire family tree up to generation N, that is the random vector

$$\{(FR_{ni}, MR_{ni}), i = 1, ..., ZR_n; (Fr_{nj}, Mr_{nj}), j = 1, ..., Zr_n; n = 0, ..., N - 1\},$$

or at least, for $n = 0, ..., N - 1$, the random variables

$$ZR_n(k_1, k_2) = \sum_{i=1}^{ZR_n} I_{\{FR_{ni}=k_1, MR_{ni}=k_2\}}, \quad \text{and} \quad Zr_n(l_1, l_2) = \sum_{j=1}^{Zr_n} I_{\{Fr_{nj}=l_1, Mr_{nj}=l_2\}},$$

which denote the number of R couples (resp. r couples) in the generation n which have generated k_1 females (resp. l_1 females) and k_2 males (resp. l_2 males), with $k_1, k_2 \in S^R$ (resp. $l_1, l_2 \in S^r$). So, we know the values of the following set of variables

$$\mathcal{Z}_N(S^R, S^r) = \{ZR_n(k_1, k_2),\ k_1, k_2 \in S^R;\ Zr_n(l_1, l_2),\ l_1, l_2 \in S^r;\ n = 0, ..., N-1\}.$$

Theorem 1 *The MLEs of* $(p^R, p^r, m_R, m_r, \alpha)$ *based on the sample* $\mathcal{Z}_N(S^R, S^r)$ *are, respectively,*

$$\hat{p}_k^R = \frac{\sum_{n=0}^{N-1} ZR_n(k)}{\sum_{n=0}^{N-1} ZR_n}, \ \hat{p}_l^r = \frac{\sum_{n=0}^{N-1} Zr_n(l)}{\sum_{n=0}^{N-1} Zr_n}, \ \hat{m}_R = \frac{\sum_{n=1}^{N} kZR_n(k)}{\sum_{n=0}^{N-1} ZR_n},$$

$$\hat{m}_r = \frac{\sum_{n=1}^{N} lZr_n(l)}{\sum_{n=0}^{N-1} Zr_n}, \ \text{and } \hat{\alpha} = \frac{\sum_{n=0}^{N-1} F_{n+1}}{\sum_{n=0}^{N-1}(F_{n+1} + MR_{n+1} + Mr_{n+1})},$$

with

$$ZR_n(k) = \sum_{\substack{k_1, k_2 \in S^R \\ k_1 + k_2 = k}} ZR_n(k_1, k_2), k \in S^R \ \text{and } Zr_n(l) = \sum_{\substack{l_1, l_2 \in S^r \\ l_1 + l_2 = l}} Zr_n(l_1, l_2), l \in S^r.$$

Remark 2 All of the estimators in Theorem 1 verify some properties related to their asymptotic behaviour. Specifically, on the non-extinction set, each estimator is strongly consistent, and, suitably normalized, converges in distribution to a standard normal distribution. ▮

4. Maximum Likelihood Estimators with Incomplete Data

Observing the sample $\mathcal{Z}_N(S^R, S^r)$, we have obtained the estimators given in Theorem 1. Moreover, the MLEs of $(p^R, p^r, m_R, m_r, \alpha)$ are also the MLEs based on the sample (see Jagers (1975)) given by the sets

$$\mathcal{FM}_N = \{ZR_0, Zr_0, F_n, MR_n, Mr_n, n = 1, ..., N\}$$

and

$$\mathcal{Z}_N = \{ZR_n(k), Zr_n(l), k \in S^R, l \in S^r, n = 0, ..., N-1\}.$$

In real situations to observe the set $\mathcal{Z}_N$ it is not easy. Hence, an interesting problem of estimation arises from supposing that instead of $\mathcal{Z}_N$ we can only observe the total number of couples of each type in each generation. If we denote, for $n = 0, \ldots, N-1$, $ZFM_n = \{ZR_n, Zr_n, F_{n+1}, MR_{n+1}, Mr_{n+1}\}$ then, we are assuming that we only observe the sample $\mathcal{ZFM}_N = \{ZFM_0, ..., ZFM_{N-1}\}$. Based on the assumption that $\mathcal{Z}_N$ is unknown and only the total number of individuals and couples are observed, we find an incomplete data estimation problem. In that case, it seems appropriate to use the Expectation-Maximization (EM) algorithm, which is a widely used method for dealing with maximum likelihood calculations when there are missing or incomplete data. In our case, the procedure leads to an iterative computing algorithm in which, starting with some initial values $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$ we obtain a sequence $\{(p^{R(i)}, p^{r(i)}, \alpha^{(i)})\}_{i\geq 0}$ which is updated in each iteration of the method. Under certain conditions, this sequence converge to the MLEs of (p^R, p^r, α) based on the sample $\mathcal{ZFM}_N$. The algorithm consists of two steps called E (expectation) and M (maximization), which we describe in the following subsections.

4.1 The E step

In this subsection we develop the E step of the EM algorithm in the iteration $(i+1)$. Let $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$ be, the values obtained in the i-th iteration (being $p^{R(i)} = \{p_k^{R(i)}\}_{k\in S^R}$ and $p^{r(i)} = \{p_l^{r(i)}\}_{l\in S^r}$). Then, the E step starts calculating the expectation of the log-likelihood with respect to available values: $(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_N)$. That expectation is given by the expression

$$E_{\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_N)}\left[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{ZFM}_N)\right]. \tag{2}$$

Given $\mathcal{ZFM}_N$ and the vector $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$ obtained in the i-th iteration, we calculate the distribution $\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_N)$. It can be proved that to sample from $\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_N)$ it is enough to sample generation-by-generation. Then, fixed $n = 0, \ldots, N-1$ and given ZFM_n, it can be shown that this can be made by conveniently normalizing the probabilities given by independent multinomial distributions with sizes ZR_n and Zr_n and probabilities $p^{R(i)}$ and $p^{r(i)}$, respectively, and independent binomial distributions with size the total number of descendants generated by all mating units of each type and probability $1-\alpha^{(i)}$. (See González et al. (2012) for more details). Once we know how to determine the distribution of $\mathcal{Z}_N$, the expectation in (2) is equal, for certain constant C, to the following expression:

$$\begin{aligned} C &+ \sum_{n=0}^{N-1}(F_{n+1}\log\alpha + (MR_{n+1}+Mr_{n+1})\log(1-\alpha)) \\ &+ \sum_{n=0}^{N-1}\left(\sum_{k\in S^R} E_i^*[ZR_n(k)]\log p_k^R + \sum_{l\in S^r} E_i^*[Zr_n(l)]\log p_l^r\right), \end{aligned}$$

with $E_i^*[\cdot]$ denoting $E_{\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_N)}[\cdot]$.

4.2 The M step

The M step consists of finding the values of the parameters which maximize the expectation of the log-likelihood. This expectation has been calculated previously in the E step. In our case, we must find the values $(p^{R(i+1)}, p^{r(i+1)}, \alpha^{(i+1)})$ which maximize the expression in (2). Following a similar argument to that given to obtain the MLEs based on the observation of the complete family tree, the values for α in the iteration $(i+1)$ is

$$\alpha^{(i+1)} = \frac{\sum_{n=0}^{N-1} F_{n+1}}{\sum_{n=0}^{N-1}(F_{n+1}+MR_{n+1}+Mr_{n+1})}.$$

Notice that, $\alpha^{(i+1)}$ does not depend on the iteration i because it is based on $\mathcal{ZFM}_N$ which is observed. Then, the sequence $\{\alpha^{(i)}\}_{i\geq 0}$ is constant in all iterations of the method and it is called $\hat{\alpha}_{EM,N}$. Moreover, this value coincides with the MLE given in Theorem 1 after observing all entire family tree. The values for each p_k^R with $k \in S^R$ and each p_l^r with $l \in S^r$ in the iteration $(i+1)$ are

$$p_k^{R(i+1)} = \frac{\sum_{n=0}^{N-1} E_i^*[ZR_n(k)]}{\sum_{n=0}^{N-1} ZR_n}, k \in S^R \text{ and } p_l^{r(i+1)} = \frac{\sum_{n=0}^{N-1} E_i^*[Zr_n(l)]}{\sum_{n=0}^{N-1} Zr_n}, l \in S^r.$$

The values found on this M step, $(p_k^{R(i+1)}, p_l^{r(i+1)}, \alpha^{(i+1)})$ are used to begin another E step, and the process is repeated until some convergence criterion is verified, in that case, the process stops and the final values are denoted $(\hat{p}^R_{EM,N}, \hat{p}^r_{EM,N},$

$\hat{\alpha}_{EM,N}$). McLachlan and Krishnan (2008) showed that, under general conditions of derivability and continuity, estimates obtained using the EM algorithm converge to a stationary point of the incomplete data likelihood function. The multinomial structure of our likelihood function, often leads us to verify such conditions and also conclude that the incomplete data likelihood function is unimodal with only one stationary point. Then, that point is the "expected MLE", which maximizes the incomplete data likelihood function. Assuming those conditions, $(\hat{p}^R_{EM,N}, \hat{p}^r_{EM,N}, \hat{\alpha}_{EM,N})$ are the "expected MLEs" of (p^R, p^r, α) given the sample $\mathcal{ZFM}_N$. We call these estimators expectation-maximization MLEs.
The following summarizes our proposed EM algorithm to estimate the parameters of the model:

> *Step 0.* $i = 0$. Set each component of $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$ take some strictly positive value.
>
> *Step 1* (**E Step**). Based on $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$,
>
> (a) Determine $\mathcal{Z}_{\mathcal{N}}|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{ZFM}_{\mathcal{N}})$
>
> (b) Calculate $E^*_i\left[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{ZFM}_N)\right]$
>
> *Step 2* (**M Step**). Obtain the vector
>
> $$(p^{R(i+1)}, p^{r(i+1)}, \alpha^{(i+1)}) = \arg\max_{(p^R,p^r,\alpha)} E^*_i[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{FM}_N)].$$
>
> *Step 3.* If $\max\{|p_k^{R(i+1)} - p_k^{R(i)}|, k \in S^R, |p_l^{r(i+1)} - p_l^{r(i)}|, l \in S^r, |\alpha^{(i+1)} - \alpha^{(i)}|\}$ is less than some convergence criterion, stop and denote $(\hat{p}^R_{EM}, \hat{p}^r_{EM}, \hat{\alpha}_{EM})$ to these final values. Otherwise, increase i by 1 and repeat steps $1 - 3$.

References

Daley, D. J., 1968. Extinction conditions for certain bisexual Galton-Watson branching processes. Z. Wahrscheinlichkeitsth. 9, 315–322.

González, M., Gutiérrez, C., Martínez, R., 2012. Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes. J. Comp. Biol. 19(9), 1015–1026.

González, M., Martínez, R., Mota, M., 2009. Bisexual branching processes to model extinction conditions for Y-linked genes. J. Theor. Biol. 258, 478–488.

Jagers, P., 1975. Branching Processes with Biological Applications. John Wiley and Sons, Inc.

McLachlan, G. J., Krishnan, T., 2008. The EM Algorithm and Extensions. John Wiley and Sons, Inc.