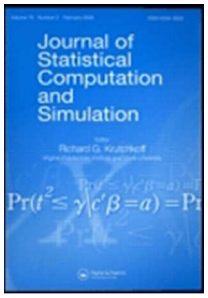

Non-parametric Bayesian inference through MCMC method for Y-linked two-sex branching processes with blind choice

Journal:	*Journal of Statistical Computation and Simulation*
Manuscript ID	GSCS-2018-0204.R1
Manuscript Type:	Original Paper
Areas of Interest:	APPLIED PROBABILITY, APPLIED STOCHASTIC PROCESSES, BAYESIAN INFERENCE, EM ALGORITHM, MARKOV PROCESSES, STOCHASTIC PROCESS
2010 Mathematics Subject Classification:	60J80, 60J85

Non-parametric Bayesian inference through MCMC method for Y-linked two-sex branching processes with blind choice

ARTICLE HISTORY
Compiled July 9, 2018

ABSTRACT
Dirichlet-process-based non-parametric Bayesian inference is developed for a Y-linked two-sex branching process with blind choice. This stochastic model is suitable for analysing the evolution of the number of carriers of two alleles of a Y-linked gene in a two-sex monogamous population where each female chooses her partner from among the male population without caring about his type (i.e., the allele he carries). The only data assumed to be available are the total number of females and males (regardless of their types) up to some generation and the numbers of each type of male in the last generation. A simulation method which is based on a Dirichlet process and a Gibbs sampler is developed to estimate the posterior distributions of the model's main parameters, i.e., those which play an important role in the long-term behaviour of the number of carriers of the alleles. Finally, the computational efficiency of the algorithm is illustrated with non-trivial example simulations and an application to real data.

KEYWORDS
Y-linked genes, two-sex branching process, blind choice of mates, non-parametric Bayesian inference, Dirichlet process, Gibbs sampler.

1. Introduction

It is well-known that in some animal populations the sex of the individuals is determined by a pair of chromosomes X (or Z) and Y (or W). In the X and Y case, a female has XX chromosomes, while a male has XY chromosomes. Certain characteristics of the individuals are due to genes carried on the X chromosome (X-linked), others to genes carried on the Y chromosome (Y-linked), and still others by genes on both chromosomes (XY-linked). Taking this fact into account, females and males with different genotypes and/or phenotypes are present in any population. Females and males in a generation form mating units (couples) in order to produce offspring. An offspring receives its genetic structure as specified by the inheritance rules associated with the species it belongs to.

A problem with practical relevance is to model and analyse the evolution from generation to generation of sex-linked genes in a two-sex population (see [1], [2], [3], [4]). Recent years have seen the development of two new stochastic models corresponding to the field of branching processes designed to analyse the evolution of characters associated with Y-linked genes (see [5] and [6] and the references therein for a deeper explanation of the motivation). These two models describe the evolution of the number of carriers of two mutually exclusive alleles of a Y-linked gene in a two-sex monogamic population. In [5], what was called the Y-linked two-sex branching process with pref-

erence was introduced, in which it was considered that the characters controlled by the gene may have some influence on the species' mating process, with females having preference for males carrying one of the alleles of the gene. In [6] on the other hand, females were considered to choose their mates without caring about their genotypes since most Y-linked characters are not decisive at the time of mating, thus introducing the Y-linked two-sex branching process with blind choice.

The probabilistic theory underlying these two models has been developed in some depth, determining conditions for the extinction/survival of Y-linked genes and for their long-term behaviour (asymptotic rate of growth) in the population. It was proved (see [5], [6], [7] and [8]) that those conditions depend on certain parameters of the model. In most real situations, however, these parameters are unknown and they need to be estimated.

In order to address this problem, the sample that can be observed must be carefully selected. In this sense, it is reasonable to assume that the total numbers of females and males in each generation can be observed (see [9] for a in-depth discussion of this issue). Indeed, in many animal populations the study of the evolution or extinction of the species (see [10]) or the estimation of the effective population size (see [11] or [12]) is based on population censuses which provide information about the total numbers of females and males over a certain number of generations. The number of generations and the population sizes observed in these censuses are usually not very large. It is therefore also reasonable to consider that only a small population might be available from which to make inferences (see also [13]). Furthermore, it is usually difficult in a genetic context to distinguish the genotype of the males because the trait is not expressed in the phenotype, so that it is reasonable to consider that only the total number of males (regardless of their types) can be observed from the first up to a certain generation.

Based on these ideas, in [9] we studied parametric Bayesian inference for the Y-linked two-sex branching process with blind choice, considering as observed sample the total number of females and males over a number of generations, and adding the information corresponding to the total number of males of each genotype in the last generation (in order to know that the two alleles are present at the population). To solve the problem, we applied a Markov chain Monte Carlo (MCMC) method for estimation from incomplete data.

In the present communication, we continue the previous study, addressing the inferential problem based on the same sample scheme but in a more general Bayesian framework. In particular, there are two main ideas underlying the innovative contributions of this paper. One is that we consider a non-parametric context. This implies a greater level of uncertainty, but it is more realistic and flexible in practice because it is usually hard to determine that the laws of reproduction belong to specific parametric families. While the problem is also taken to be one of incomplete data estimation, the fact of not knowing the families of the reproduction laws implies that more parameters have to be estimated than in the parametric case. The other is that we assume complete ignorance of the reproduction laws. We thus impose no kind of restriction on the cardinality of their supports, introducing a Dirichlet process as the convenient class of prior distributions. These two ideas together constitute a major improvement over the framework presented in [9].

The rest of this paper is organized into five sections. In Section 2 that follows, we present the probabilistic model, and set out the inference problem. In Section 3, we develop a simulation method based on a Dirichlet process and a Gibbs sampler to obtain the posterior distributions of the model's main parameters and the predictive

distributions for as yet unobserved generations. In Section 4, we apply the algorithm to simulated data (with the inclusion of a sensitivity analysis), and verify the robustness of the method by means of a general simulation experiment. In Section 5, we apply the method to the real data set given in [13] corresponding to a pedigree of Y-linked non-syndromic hearing impairment in a Chinese family. Finally, in Section 6, we provide some concluding remarks.

2. The probability model and sample scheme

The branching process considered in this paper was presented in [6] to analyse generation-by-generation the evolution of the number of carriers of two mutually exclusive alleles, labeled R and r, of a Y-linked gene in a two-sex monogamous population in which each female chooses her partner from among the male population without caring about his type. The following is its mathematical definition.

Definition 2.1. Let $\{(FR_{ni}, MR_{ni}) : i = 1, 2, \ldots;\ n = 0, 1, \ldots\}$ and $\{(Fr_{nj},\ Mr_{nj}) : j = 1, 2, \ldots; n = 0, 1, \ldots\}$ be two independent sequences of independent, identically distributed, non-negative and integer-valued bivariate random vectors on the same probability triple $(\Omega, \mathcal{F}, P)$. The sequences $\{(ZR_n, Zr_n)\}_{n\geq 0}$ and $\{(F_{n+1},\ MR_{n+1},\ Mr_{n+1})\}_{n\geq 0}$ are defined recursively, for each $n \geq 0$, as follows:

$$(ZR_0, Zr_0) = (a, b) \in \mathbb{N}_0^2,$$

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

and

$$Mr_{n+1} = \sum_{j=1}^{Zr_n} Mr_{nj},$$

assuming that $\sum_1^0 = 0$; and

- If $F_{n+1} \geq MR_{n+1} + Mr_{n+1}$, then

$$ZR_{n+1} = MR_{n+1} \text{ and } Zr_{n+1} = Mr_{n+1}.$$

- If $F_{n+1} < MR_{n+1} + Mr_{n+1}$, then

$$ZR_{n+1} \sim H(F_{n+1}, MR_{n+1} + Mr_{n+1}, MR_{n+1})$$

and

$$Zr_{n+1} = F_{n+1} - ZR_{n+1},$$

where $H(F_{n+1}, MR_{n+1} + Mr_{n+1}, MR_{n+1})$ denotes the hypergeometric distribution with parameters F_{n+1}, $MR_{n+1} + Mr_{n+1}$, MR_{n+1}.

The two-dimensional process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is called a Y-linked two-sex branching process with blind choice.

The process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is a homogeneous two-type Markov chain. Intuitively, for n fixed, the random vector (ZR_n, Zr_n) represents the total number of couples of types R and r, respectively, at generation n, where the type of a couple is determined by the type of its male. To describe the evolution of the population from this generation onwards, two phases are considered: reproduction and mating.

In the reproduction phase, each couple, independently of the others, generates females and males of its type according to some probability distribution depending on its type. Then, (FR_{ni}, MR_{ni}) and (Fr_{nj}, Mr_{nj}) denote the total number of females and males produced by the ith R-couple and the jth r-couple, respectively, at generation n, and F_{n+1}, and MR_{n+1} and Mr_{n+1} denote the total number of females, and R- and r-males at generation $n+1$. Notice that mutation of the gene is not considered. A more general and complex stochastic process in which that possibility is considered has been presented in [14].

In the mating phase, the total number of individuals in generation $n+1$ is known $(F_{n+1}, MR_{n+1}, Mr_{n+1})$, and the number of couples of each genotype formed in generation $n+1$ is calculated assuming that generations do not overlap, that there is perfect fidelity (monogamy), and that females choose their partners from among the male population without caring about their type. Hence, 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. Equivalently, if the total number of females is less than the total number of males, all females mate. Since females make a blind choice from among the males, the total number of R-couples is given by a hypergeometric distribution with parameters $(F_{n+1}, MR_{n+1} + Mr_{n+1}, MR_{n+1})$, i.e., F_{n+1} males are selected from all males of generation $n+1$, where MR_{n+1} males have R genotype. The rest of the couples will be of type r.

Since, in nature, R- and r-couples may differ in their reproductive abilities, we consider that, in general, (FR_{01}, MR_{01}) and (Fr_{01}, Mr_{01}) may have different probability distributions. We denote by $p^R = \{p_k^R\}_{k\geq 0}$ and $p^r = \{p_l^r\}_{l\geq 0}$ the reproduction laws of R- and r-couples, respectively. Thus, an R-couple (r-couple) generates $k \geq 0$ ($l \geq 0$) individuals with probability $p_k^R = P(FR_{01} + MR_{01} = k)$ ($p_l^r = P(Fr_{01} + Mr_{01} = l)$). Notice that their supports can have infinite cardinality, or at least they have no known upper bound. We also consider that these reproduction laws have finite means, denoted by m_R and m_r, respectively, representing the average number of offspring generated by a couple of each genotype.

Furthermore, an offspring will be female with probability α, $0 < \alpha < 1$, and male with probability $1-\alpha$. These sex designations are made independently following a binomial scheme among the offspring of any couple, and it is assumed that the genotype has no influence on the sex determination, so that α is the same for both genotypes. As a consequence of this reproduction scheme, one has that the average number of females and males generated by an R-couple are αm_R and $(1-\alpha)m_R$, respectively, while the respective values for an r-couple are αm_r and $(1-\alpha)m_r$.

As mentioned in the Introduction, conditions for the extinction/survival of Y-linked genes and for determining their asymptotic rate of growth were provided in [7] and [6]. These conditions depend on the parameters of the model α, p^R, and p^r.

To estimate these parameters, from this point onwards we shall assume that, for

$N > 0$, the set of vectors

$$\mathcal{FM}_N = \{FM_0, FM_1, \ldots, FM_{N-1}, F_N, MR_N, Mr_N\}$$

is observed, where $FM_n = (F_n, M_n),\ n = 0, \ldots, N-1$, is the vector given by the total numbers of females and males in generation n. Notice that the total number of $R-$ and $r-$ males is assumed to be observed only in the last generation. Moreover, FM_0 is considered to be fixed as the initial generation.

From the sample $\mathcal{FM}_N$, we will draw inferences on (α, p^R, p^r), and consequently also on the reproduction means (m_R, m_r) and on the future population sizes of females, males, and couples of each type, i.e., for any $s \geq 0$, the unobserved vector $(ZR_{N+s}, Zr_{N+s}, F_{N+s+1}, MR_{N+s+1}, Mr_{N+s+1})$.

To this end, as was indicated in the Introduction, we approach the problem from a Bayesian perspective in a non-parametric context, determining the posterior distribution of (α, p^R, p^r) given $\mathcal{FM}_N$, denoted by $(\alpha, p^R, p^r)|\mathcal{FM}_N$, and then deriving the posterior distribution of (m_R, m_r) and, for any $s \geq 0$, $(ZR_{N+s},\ Zr_{N+s},\ F_{N+s+1},\ MR_{N+s+1},\ Mr_{N+s+1})$ denoted by $(m_R, m_r)|\mathcal{FM}_N$ and $(ZR_{N+s},\ Zr_{N+s},\ F_{N+s+1},\ MR_{N+s+1},\ Mr_{N+s+1})|\mathcal{FM}_N$, respectively. Since the branching structure is not derived from $\mathcal{FM}_N$ (i.e., one cannot deduce which individuals are generated by each type of mating unit), there are no closed forms for these posterior distributions. We approximate them by posing the problem as one of incomplete data, and applying a simulation method based on the Gibbs sampler.

3. Implementation of the Gibbs sampling

The posterior distribution $(\alpha, p^R, p^r)|\mathcal{FM}_N$ could be determined if one knew the random sequences $\mathcal{M}Rr_N$ and $\mathcal{Z}Rr_N$. The former provides information about the total number of males of each genotype, i.e.,

$$\mathcal{M}Rr_N = \{MRr_0, \ldots, MRr_{N-1}\},$$

with $MRr_n = (MR_n, Mr_n)$, for $n = 0, \ldots, N-1$. The latter deals with the total offspring of each couple. In particular,

$$\mathcal{Z}Rr_N = \{ZRr_0, \ldots, ZRr_{N-1}\},$$

where $ZRr_n = \{ZR_n(k), k \geq 0,\ Zr_n(l), l \geq 0\}$ for $n = 0, \ldots, N-1$, with $ZR_n(k)$ and $Zr_n(l)$ being the numbers of R- and r-couples in generation n with exactly k and l offspring, respectively, without knowing whether they are males or females, i.e.,

$$ZR_n(k) = \sum_{i=1}^{ZR_n} I_{\{FR_{ni}+MR_{ni}=k\}} \quad \text{and} \quad Zr_n(l) = \sum_{j=1}^{Zr_n} I_{\{Fr_{nj}+Mr_{nj}=l\}},$$

where I_A denotes the indicator function of a set A.

It is easy to verify that $\mathcal{FM}_N$, $\mathcal{M}Rr_N$, and $\mathcal{Z}Rr_N$ are related by the expressions

$$MR_n + Mr_n = M_n, \quad ZR_n = \sum_{k\geq 0} ZR_n(k), \quad Zr_n = \sum_{l\geq 0} Zr_n(l),$$

$$ZR_n + Zr_n = \min\{F_n, MR_n + Mr_n\},$$

$$F_{n+1} + MR_{n+1} + Mr_{n+1} = \sum_{k\geq 0} kZR_n(k) + \sum_{l\geq 0} lZr_n(l),$$

$$\sum_{k\geq 0} kZR_n(k) \geq MR_{n+1} \quad \text{and} \quad \sum_{l\geq 0} lZr_n(l) \geq Mr_{n+1},$$

for every $n = 0, \ldots, N-1$. Although $\mathcal{M}Rr_N$ and $\mathcal{Z}Rr_N$ are unobserved, they can be simulated by considering them to be latent sequences. The posterior distribution

$$(\alpha, p^R, p^r, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N$$

can then be derived by applying the Gibbs sampler. To this end, it is only necessary to determine the conditional posterior distributions

$$(\alpha, p^R, p^r)|(\mathcal{FM}_N, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)$$

and, for $n = 0, \ldots, N-1$,

$$(MRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{M}Rr_{N(-n)}, \mathcal{Z}Rr_{N(-n)}, \alpha, p^R, p^r), \tag{1}$$

where, for $n = 0, \ldots, N-1$, $\mathcal{M}Rr_{N(-n)}$ and $\mathcal{Z}Rr_{N(-n)}$ denote the sets of random sequences given by $\mathcal{M}Rr_N$ and $\mathcal{Z}Rr_N$, respectively, except those variables belonging to generation n.

3.1. Determining $(\alpha, p^R, p^r)|(\mathcal{FM}_N, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)$

The likelihood function of (α, p^R, p^r) based on the sample and latent sequences, $(\mathcal{FM}_N, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)$, can easily be obtained by

$$f((\mathcal{FM}_N, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)|(\alpha, p^R, p^r)) \propto \prod_{n=0}^{N-1} \alpha^{F_{n+1}}(1-\alpha)^{MR_{n+1}+Mr_{n+1}} \prod_{k\geq 0}(p_k^R)^{ZR_n(k)} \prod_{l\geq 0}(p_l^r)^{Zr_n(l)}. \tag{2}$$

Thus, from these multinomial forms and taking into account that no restriction has been imposed on the reproduction laws' cardinality (which is considered to be unknown), the Dirichlet processes (the natural conjugate family) constitute a convenient class of prior distributions for α, p^R, and p^r (see [15] for details of the Dirichlet processes). Therefore one takes α to follow a Dirichlet (beta) distribution with parameters β_1 and β_2, with $\beta_1, \beta_2 > 0$, i.e.,

$$\alpha \sim \text{Be}(\beta_1, \beta_2) \quad \text{and} \quad \pi(\alpha) \propto \alpha^{\beta_1 - 1}(1-\alpha)^{\beta_2 - 1},$$

and

$$p^R \sim \mathrm{DP}(p^R(0), \beta^R) \text{ and } p^r \sim \mathrm{DP}(p^r(0), \beta^r),$$

where DP denotes the Dirichlet process, with $p^R(0) = \{p_k^R(0)\}_{k\geq 0}$ and $p^r(0) = \{p_l^r(0)\}_{l\geq 0}$ the base measures, and β^R and β^r the concentration parameters, $\beta^R, \beta^r > 0$.

Since mating units reproduce independently and the assignment of sex is also independent, one can assume that

$$(\alpha, p^R, p^r) \sim \mathrm{Be}(\beta_1, \beta_2) \bigotimes \mathrm{DP}(p^R(0), \beta^R) \bigotimes \mathrm{DP}(p^r(0), \beta^r),$$

with $\bigotimes$ being the product of independent random processes or vectors.

Taking Equation (2) into account, it follows that the distribution of (α, p^R, p^r) given $(\mathcal{FM}_N, \mathcal{MR}r_N, \mathcal{ZR}r_N)$ is

$$\begin{aligned}(\alpha, p^R, p^r)|(\mathcal{FM}_N, \mathcal{MR}r_N, \mathcal{ZR}r_N) \sim \\ \mathrm{Be}\left(\beta_1 + \sum_{n=1}^{N} F_n, \beta_2 + \sum_{n=1}^{N} M_n\right) \\ \bigotimes \mathrm{DP}\left(\frac{\beta^R}{\beta^R + YR_N} p^R(0) + \frac{1}{\beta^R + YR_N} \sum_{k\geq 0} YR_N(k), \beta^R + YR_N\right) \\ \bigotimes \mathrm{DP}\left(\frac{\beta^r}{\beta^r + Yr_N} p^r(0) + \frac{1}{\beta^r + Yr_N} \sum_{l\geq 0} Yr_N(l), \beta^r + Yr_N\right), \end{aligned} \quad (3)$$

where $YR_N = \sum_{n=0}^{N-1} ZR_n$, $Yr_N = \sum_{n=0}^{N-1} Zr_n$, $YR_N(k) = \sum_{n=0}^{N-1} ZR_n(k)\delta_k$, and $Yr_N(l) = \sum_{n=0}^{N-1} Zr_n(l)\delta_l$, with δ_s being a Dirac delta function at s, $s \geq 0$. Notice that the posterior distribution $\alpha|(\mathcal{FM}_N, \mathcal{MR}r_N, \mathcal{ZR}r_N)$ depends only on $\mathcal{FM}_N$, and the posterior distributions for the reproduction laws of both genotypes depend only on $\mathcal{ZR}r_N$.

Therefore, in order to obtain a closed form for the posterior distribution of (α, p^R, p^r) such as that given in Equation (3), one needs not only the observation of $\mathcal{FM}_N$ but also to know $\mathcal{ZR}r_N$. Notice that we have also considered $\mathcal{MR}r_N$ since $\mathcal{ZR}r_N$ is unobserved and is obtained through $\mathcal{FM}_N$ and $\mathcal{MR}r_N$ by applying the formulas of the model (see Definition 2.1). We shall next determine their joint posterior distribution for each generation n, with $n = 0, \ldots, N-1$.

3.2. Determining $(\boldsymbol{MRr_n, ZRr_n})|(\boldsymbol{\mathcal{FM}_N, \mathcal{MR}r_{N(-n)}, \mathcal{ZR}r_{N(-n)}, \alpha, p^R, p^r})$

Let $fm_n = (f_n, m_n)$ and $mRr_n = (mR_n, mr_n)$, for $n = 0, ..., N$, be non-negative integer vectors with $m_N = mR_N + mr_N$, and then define the sets, for $n = 0, ..., N$,

$$\begin{aligned} A_{fm_n} &= \{F_n = f_n, M_n = m_n\}, \\ A_{mRr_n} &= \{MR_n = mR_n, Mr_n = mr_n\}, \end{aligned}$$

and

$$A_{\overline{fm}_N} = A_{mRr_N} \cap \bigcap_{n=0}^{N} A_{fm_n}.$$

Also, let $zR_n^* = (zR_n(k), k \geq 0)$ and $zr_n^* = (zr_n(l), l \geq 0)$, for $n = 0, ..., N-1$, be two sequences of non-negative integer numbers, and define the sets

$$\begin{aligned} A_{zR_n^*} &= \{ZR_n(k) = zR_n(k), k \geq 0\}, \\ A_{zr_n^*} &= \{Zr_n(l) = zr_n(l), l \geq 0\}. \end{aligned}$$

Moreover, for $n = 0, ..., N-1$, define the sets

$$\begin{aligned} A_{mRr_{N(-n)}} &= A_{mRr_0} \cap \cdots \cap A_{mRr_{n-1}} \cap A_{mRr_{n+1}} \cap \cdots \cap A_{mRr_{N-1}} \\ A_{zRr_{N(-n)}} &= A_{zR_0^*} \cap \cdots \cap A_{zR_{n-1}^*} \cap A_{zR_{n+1}^*} \cap \cdots \cap A_{zR_{N-1}^*} \\ &\quad \cap A_{zr_0^*} \cap \cdots \cap A_{zr_{n-1}^*} \cap A_{zr_{n+1}^*} \cap \cdots \cap A_{zr_{N-1}^*}. \end{aligned}$$

Now, we deal with the posterior distributions for $n = 0, \ldots, N-1$

$$(MRr_n, ZRr_n) | (\mathcal{FM}_N, \mathcal{MRr}_{N(-n)}, \mathcal{ZRr}_{N(-n)}, \alpha, p^R, p^r). \tag{4}$$

We consider first the case $n = 0$, determining the probability distribution of (MRr_0, ZRr_0), assuming that we know (α, p^R, p^r), the sample $\mathcal{FM}_N$, and the future generations, i.e., (MRr_n, ZRr_n), for all $n = 1, ..., N-1$. To simplify the notation, we shall write $P(\cdot)$ to denote the conditional probability with parameters (α, p^R, p^r). One then has to determine the probability

$$P(A_{mRr_0}, A_{zR_0^*}, A_{zr_0^*} | A_{\overline{fm}_N}, A_{mRr_{N(-0)}}, A_{zRr_{N(-0)}}),$$

which is proportional to

$$P(A_{mRr_0} | A_{fm_0}) P(A_{zR_0^*}, A_{zr_0^*} | A_{fm_0}, A_{mRr_0}) P(A_{fm_1}, A_{mRr_1} | A_{zR_0^*}, A_{zr_0^*}), \tag{5}$$

by simple recursive application of the multiplication rule and the Markov property.

In order to calculate the first probability,

$$P(A_{mRr_0} | A_{fm_0}), \tag{6}$$

we assume that the total numbers of females and of males in the initial generation (i.e., F_0 and M_0) are fixed and there is no information on the initial population frequencies of each type of male (i.e., considering the unknown (MR_0, Mr_0) to be nuisance parameters) subject to the constraint $MR_0 + Mr_0 = M_0$ and independent of (α, p^R, p^r). Then, a reasonable choice for the prior distribution of (MR_0, Mr_0) is the uniform distribution on the set $\{(x, y) \in \mathbb{Z}_+^2 : x + y = M_0\}$, with $\mathbb{Z}_+$ denoting the set of non-negative integers.

Given the definition of the model, it is not hard to obtain that the second probability in (5),

$$P(A_{zR_0^*}, A_{zr_0^*} | A_{fm_0}, A_{mRr_0}), \tag{7}$$

is equal to the product of the probabilities

$$P((ZR_0, Zr_0) = (zR_0, zr_0)|A_{fm_0}, A_{mRr_0}) \tag{8}$$

and

$$P(A_{zR_0^*}, A_{zr_0^*}|(ZR_0, Zr_0) = (zR_0, zr_0)), \tag{9}$$

where $zR_0 = \sum_{k\geq 0} zR_0(k)$ and $zr_0 = \sum_{l\geq 0} zr_0(l)$. The probability in (8) is easily obtained from the definition of the model. Furthermore, since the mating units of different types reproduce independently, the expression of Equation (9) is equal to $P(A_{zR_0^*}|ZR_0 = zR_0)P(A_{zr_0^*}|Zr_0 = zr_0)$. Taking into account that p_k^R is the probability that an R-couple generates k offspring, $k \geq 0$, and that there are zR_0 couples, one has that $P(A_{zR_0^*}|ZR_0 = zR_0)$ is equal to $\frac{zR_0!}{\prod_{k\geq 0} zR_0(k)!} \prod_{k\geq 0} (p_k^R)^{zR_0(k)}$. Note that, in this multinomial form, although there are infinite products, only a finite number of $zR_n(k)$ are non-null, actually at most zR_0. The derivation of $P(A_{zr_0^*}|Zr_0 = zr_0)$ is similar.

Finally, considering that the sex-designation follows a binomial scheme and denoting by $tR_1 = \sum_{k\geq 0} kzR_0(k)$ ($tr_1 = \sum_{l\geq 0} lzr_0(l)$) the total progeny of R-couples (r-couples), one obtains that the third probability in (5),

$$P(A_{fm_1}, A_{mRr_1}|A_{zR_0^*}, A_{zr_0^*}), \tag{10}$$

is equal to either

$$\binom{tR_1}{mR_1} \alpha^{tR_1 - mR_1} (1-\alpha)^{mR_1} \binom{tr_1}{mr_1} \alpha^{tr_1 - mr_1} (1-\alpha)^{mr_1},$$

if $tR_1 + tr_1 = f_1 + mR_1 + mr_1$, $tR_1 \geq mR_1$ and $tr_1 \geq mr_1$, or 0 otherwise.

For generations $n = 1, ..., N-1$, the calculation of the distribution in (4) also takes into account past generations. Because of this, such counting is slightly different from the previous case of $n = 0$.

Then, again applying the multiplication rule and the Markov property recursively, one has that the probability

$$P(A_{mRr_n}, A_{zR_n^*}, A_{zr_n^*}|A_{\overline{fm}_N}, A_{mRr_{N(-n)}}, A_{zRr_{N(-n)}})$$

is proportional to the product

$$P(A_{fm_n}, A_{mRr_n}|A_{zR_{n-1}^*}, A_{zr_{n-1}^*}) P(A_{zR_n^*}, A_{zr_n^*}|A_{fm_n}, A_{mRr_n})$$

$$P(A_{fm_{n+1}}, A_{mRr_{n+1}}|A_{zR_n^*}, A_{zr_n^*}).$$

The first and the third probabilities are calculated in the same manner as (10), while the second probability can be calculated analogously to (7).

3.3. Implementation of the method

We now develop the Gibbs sampler based method considering the parameters β_1, β_2, β^R, β^r, $p^{R(0)}$, and $p^{r(0)}$ of the Dirichlet processes and the sample $\mathcal{FM}_N$.

`Initialize` $t = 0$
`Generate` $\alpha^{(0)} \sim \text{Be}(\beta_1, \beta_2)$; $p^{R(0)} \sim \text{DP}(p^R(0), \beta^R)$; $p^{r(0)} \sim \text{DP}(p^r(0), \beta^r)$
  `Fix` $(MRr_n^{(0)}, ZRr_n^{(0)})$, for $n = 0, \ldots, N-1$
  `Iterate` $t = t+1$
    `Generate`, for $n = 0, \ldots, N-1$, $(MRr_n^{(t)}, ZRr_n^{(t)})$ from
    $(MRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{M}Rr_{N(-n)}^{(t)}, \mathcal{Z}Rr_{N(-n)}^{(t)}, \alpha^{(t-1)}, p^{R(t-1)}, p^{r(t-1)})$
    `with`
    $\mathcal{M}Rr_{N(-n)}^{(t)} = (MRr_0^{(t)}, \ldots, MRr_{n-1}^{(t)}, MRr_{n+1}^{(t-1)}, \ldots, MRr_{N-1}^{(t-1)})$
    `and`
    $\mathcal{Z}Rr_{N(-n)}^{(t)} = (ZRr_0^{(t)}, \ldots, ZRr_{n-1}^{(t)}, ZRr_{n+1}^{(t-1)}, \ldots, ZRr_{N-1}^{(t-1)})$
    `Generate` $(\alpha^{(t)}, p^{R(t)}, p^{r(t)}) \sim (\alpha, p^R, p^r)|(\mathcal{FM}_N, \mathcal{M}Rr_N^{(t)}, \mathcal{Z}Rr_N^{(t)})$.

Given the initial observed sample $\mathcal{FM}_N$, the algorithm is initialized by simulating the sequence $(\mathcal{M}Rr_N^{(0)}, \mathcal{Z}Rr_N^{(0)})$ subject to the constraints provided by $\mathcal{FM}_N$. Notice that, although the cardinality of the supports of the reproduction laws may be infinite, once $\mathcal{FM}_N$ is known, only a finite number of the coordinates of $\mathcal{Z}Rr_N$ are non-null. Indeed, $ZR_n(s) = 0$ and $Zr_n(s) = 0$ for all $s > F_{n+1}+M_{n+1}$, for every $n = 0, \ldots, N-1$. Then, given the sample $\mathcal{FM}_N$, the maximum number of coordinates of $p^{R(t)}$ and $p^{r(t)}$, for all $t \geq 0$, which work in the algorithm is given by $\max\{F_{n+1} + M_{n+1}, n = 0, \ldots, N-1\}+1$. Hence, in the last step of the algorithm, taking into account Equation (3) and the properties of the Dirichlet process, one obtains these coordinates from the Dirichlet distribution.

The sequence $\{(\alpha^{(t)}, p^{R(t)}, p^{r(t)}, \mathcal{M}Rr_N^{(t)}, \mathcal{Z}Rr_N^{(t)})\}_{t\geq 0}$ constitutes an ergodic Markov chain, and the stationary distribution of that chain is just the sought-after joint distribution $(\alpha, p^R, p^r, \mathcal{M}Rr_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N$.

For a run of the sequence $\{(\alpha^{(t)}, p^{R(t)}, p^{r(t)}, \mathcal{M}Rr_N^{(t)}, \mathcal{Z}Rr_N^{(t)})\}_{t\geq 0}$, from a practical standpoint, one must choose various elements that can be considered to be independent sample values of the stationary distribution. To this end, one must first choose a burn-in period, L, from which the chain can be considered to have converged. To determine L we will make use of the Gelman-Rubin-Brooks methodological approach (see [16] and [17]). To guarantee the independence of the observations, we shall consider a batch size G on the basis of an autocorrelation diagnostic. Hence, one chooses $Q+1$ vectors in the form

$$\{(\alpha^{(L+kG)}, p^{R(L+kG)}, p^{r(L+kG)}, \mathcal{M}Rr_N^{(L+kG)}, \mathcal{Z}Rr_N^{(L+kG)})\}_{k=0,\ldots,Q}.$$

When G and L are large enough (determined in practice by the above methodological approach), the vectors selected can be considered to be independent samples drawn from (α, p^R, p^r) $|\mathcal{FM}_N$ (see [18]). Since these vectors could be affected by the initial state $(\alpha^{(0)}, p^{R(0)}, p^{r(0)})$, the algorithm is applied T times, yielding a final sample of length $T(Q+1)$. From this sample, we approximate the distribution function of $(\alpha, p^R, p^r)|\mathcal{FM}_N$ by means of kernel density estimators (see [19]).

Finally, we notice that, from $(\alpha, p^R, p^r)|\mathcal{FM}_N$ and taking into account the relation between (p^R, p^r) and (m_R, m_r), one can also obtain a sample from the posterior distribution $(m_R, m_r)|\mathcal{FM}_N$. Moreover, using a Monte-Carlo method, a sample can be

obtained from

$$(ZR_{N+s}, Zr_{N+s}, F_{N+s+1}, MR_{N+s+1}, Mr_{N+s+1})|\mathcal{FM}_N,$$

for any $s \geq 0$, simulating s generations of a Y-linked two-sex branching process with blind choice starting with (F_N, MR_N, Mr_N) individuals and parameters of the model sampled from $(\alpha, p^R, p^r)|\mathcal{FM}_N$.

Remark 1. The algorithm proposed in this paper is readily adaptable to other models with different mating phases, allowing the method to be applied to a wide range of specific real situations. Examples are preference in the choice of males (see [5]) and the situation to be presented in Section 5 to illustrate modeling a Y-linked pedigree. These modifications in the model only involve changes in the calculation of the probability (8) in the proposed algorithm. The difficulties in the explicit calculation of that probability will depend on the distributions considered for each type of couple in the mating phase.

Remark 2. The implementation of the Gibbs sampler given in this section has been for a sampling scheme in which all the generations up to a fixed one could be observed. To obtain the joint posterior distribution of the latent vectors given the parameter vector and the observed sample, i.e., $(\mathcal{M}R_N, \mathcal{Z}Rr_N)|(\alpha, p^R, p^R, \mathcal{FM}_N)$, we determined it generation by generation in a conditional way (see Subsection 3.2). It is worth mentioning that, at least when the number of generations is small, that joint posterior distribution could also be sampled through the forward-backward (FB) algorithm. This is an inference algorithm for Hidden Markov Models (HMM) which computes the smoothed conditional state probability densities for updating HMM parameters according to the Baum-Welch technique (see [20] and [21]). In our case, the FB algorithm could be included in the Gibbs sampler as a block that, given $(\alpha, p^R, p^R, \mathcal{FM}_N)$, calculates the latent vectors $(\mathcal{M}R_N, \mathcal{Z}Rr_N)$ all at once.

4. Application of the method through a simulated example

We set $\alpha = 0.4$ since in most populations the sex-ratio is different from 0.5 (although close to it), and the analysis of the evolution of Y-linked genes is found to be more interesting when $\alpha < 0.5$ (see [6] and [7]). Moreover, in order to illustrate the possible difference between the reproductive abilities of mating units of each type that might exist in nature, we took different reproduction laws with finite support:

$$(p_0^R, p_1^R, p_2^R, p_3^R, p_4^R) = (0.0625, 0.2500, 0.3750, 0.2500, 0.0625)$$

and

$$(p_0^r, p_1^r, p_2^r, p_3^r, p_4^r, p_5^r) = (0.0079, 0.0646, 0.2109, 0.3442, 0.2808, 0.0916).$$

Hence $m_R = 2$ and $m_r = 3.1$. Since $\alpha < 0.5$, $\alpha m_R < 1$ and $1 < \alpha m_r$, using the results given in [6] and [7], one can deduce that the R genotype becomes extinct almost surely, and that the r allele has a positive probability of survival over the course of the generations and that it eventually grows at the asymptotic rate of αm_r.

For this model, we fixed the values $(F_0, MR_0, Mr_0) = (3, 2, 2)$ and simulated 7 generations of a Y-linked two-sex branching process with blind choice. Notice that the initial frequencies $(MR_0, Mr_0) = (2, 2)$ are considered unknown and balanced for both

n	0	1	2	3	4	5	6	7
F_n	3	5	3	6	6	4	4	5
M_n	4	5	11	3	6	9	4	8

Table 1. Simulated data.

genotypes. Table 1 presents the total number of females and males obtained over the course of the generations, with the split of the $M_7 = 8$ males into $(MR_7, Mr_7) = (2, 6)$ also being obtained from the simulation process.

Notice also that it would be difficult to determine on the basis of simple observation anything about the future behaviour of the Y-linked character. Assuming that there was no prior information available for α, we considered a beta prior distribution with parameters $\beta_1 = 1$ and $\beta_2 = 1$, i.e., a uniform distribution. Also, for p^R and p^r we considered prior Dirichlet processes with concentration parameter 1 and Poisson distribution base measure, since this type of distribution has been found to be appropriate as a reproduction law (see for example [22] or [23]) and the supports of reproduction laws are not known in principle. We assumed that the two base measures have the same mean, since we consider a priori the same reproductive ability for the two genotypes as we have no prior information on which to make any other choice. With this assumption, the total number of couples in each generation would follow a two-sex branching process (see [6]), and the theory developed in [24] could be applied. Using the simulated data and the squared-error loss estimate given in the latter of the aforementioned two papers, our estimate of the prior common mean reproductive ability was 2.8214.

Finally, the probability in (6) is calculated by means of a uniform distribution on the set $\{(x, y) \in \mathbb{Z}_+^2 : x + y = 4; x, y > 0\}$, as no prior information on the initial frequencies of M_0 is considered.

We then applied the algorithm of the previous section, simulating 20 chains ($T = 20$) formed by 20 000 iterations of the method. As a test of the convergence of the resulting probabilities to the stationary distribution, Table 2 gives the estimated potential scale reduction factor together with an upper confidence limit for α, and the first coordinates of p^R and p^r. Because in all the cases the estimated scale reduction factors are close to unity, they suggest that further simulations will not improve the values of the scalar estimators listed. Moreover, Gelman-Rubin-Brooks diagnostic tools indicate that a burn-in period of 5000 is enough ($L = 5000$). Table 2 also gives the autocorrelation values for iterations $5000 - 20\,000$ at lags 1, 50, and 100. The lag-1 and lag-50 autocorrelation values indicate that a batch procedure is necessary, and the lag-100 values that the batch size $G = 100$ is sufficient. We thus obtain a sample of size 3020 ($Q = 150$) from $(\alpha, p^R, p^r)|\mathcal{FM}_7$.

To evaluate the algorithm's efficiency, Table 3 is an extract of some summary statistics for the posterior distribution of α, the reproduction means m_R and m_r, and the unobserved variables ZR_7, Zr_7, F_8, MR_8, and Mr_8 (which we shall deal with later). Notice that the number of observations can be considered to be a reasonable choice since, in all cases, the time-series standard errors (TSSE) and the Monte-Carlo standard errors (MCSE) are less than 5% of the posterior standard deviation (SD). Moreover, those errors are very close to each other owing to the batch procedure.

Figures 1 and 2 show the estimated posterior density for α, m_R, and m_r, given $\mathcal{FM}_7$, with 95% high posterior density (HPD) sets. The contour plot of the estimated joint posterior distribution $(m_R, m_r)|\mathcal{FM}_7$ is also shown in Figure 1. Moreover $p^R|\mathcal{FM}_7$ and $p^r|\mathcal{FM}_7$ were also obtained, but are not presented for the sake of simplicity. The

	Potential Scale Reduction		Autocorrelations		
	Est.	97.5%	lag-1	lag-50	lag-100
α	1.00	1.00	-0.00386	0.00086	-0.00237
p_0^R	1.00	1.00	0.71626	-0.00573	-0.00417
p_1^R	1.00	1.00	0.81615	0.07985	0.00942
p_2^R	1.00	1.00	0.86957	0.07170	0.00428
p_3^R	1.00	1.00	0.88165	0.06646	0.01421
p_4^R	1.00	1.00	0.81865	0.05702	0.02754
p_0^r	1.00	1.00	0.75616	0.04110	0.00317
p_1^r	1.00	1.00	0.81709	0.07472	0.02502
p_2^r	1.00	1.00	0.87558	0.05526	0.00672
p_3^r	1.00	1.00	0.92610	0.17157	0.05989
p_4^r	1.00	1.00	0.87570	0.12401	0.01779

Table 2. Potential scale reduction factors and autocorrelations for α, and the first values of p^R and p^r.

	α	m_R	m_r	ZR_7	Zr_7	F_8	MR_8	Mr_8
MEAN	0.4230	2.4959	2.9744	1.2617	3.7382	6.2006	2.0057	6.3898
SD	0.0541	0.4994	0.4240	0.6321	0.6321	2.4707	1.6526	2.6969
MCSE	0.0010	0.0089	0.0075	0.0112	0.0112	0.0439	0.0293	0.0479
TSSE	0.0010	0.0093	0.0088	0.0116	0.0116	0.0474	0.0349	0.0455

Table 3. Summary statistics for the posterior distributions of α, m_R, m_r, ZR_7, Zr_7, F_8, MR_8, and Mr_8, given $\mathcal{FM}_7$.

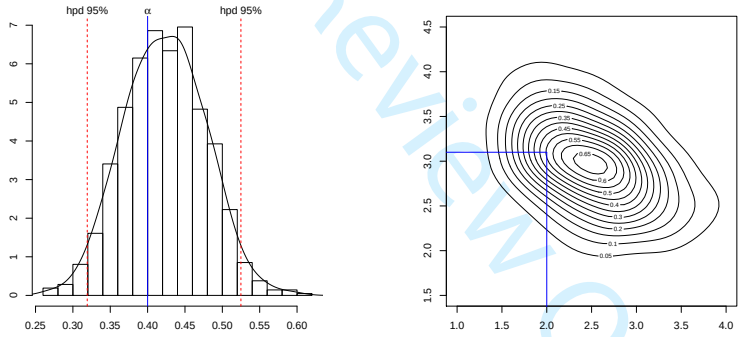

Figure 1. Estimated density for $\alpha|\mathcal{FM}_7$ with 95% HPD set (left), and contour plot of $(m_R, m_r)|\mathcal{FM}_7$ (right).

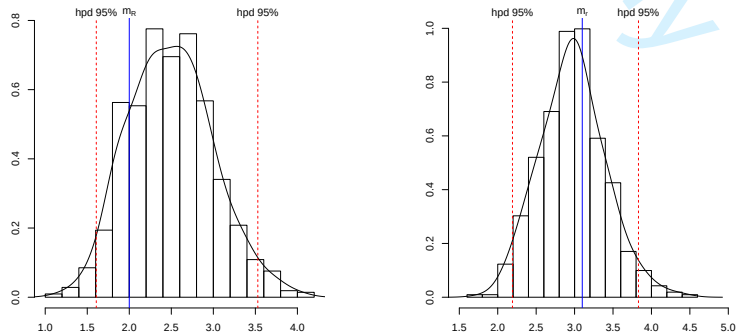

Figure 2. Estimated density for $m_R|\mathcal{FM}_7$ (left) and $m_r|\mathcal{FM}_7$ (right), with 95% HPD set.

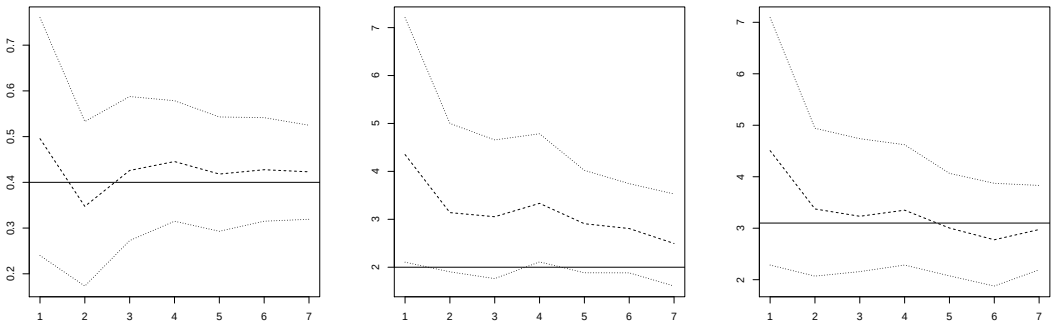

Figure 3. Evolution of the squared-error loss estimates of α (left), m_R (middle), and m_r (right), with 95% HPD sets.

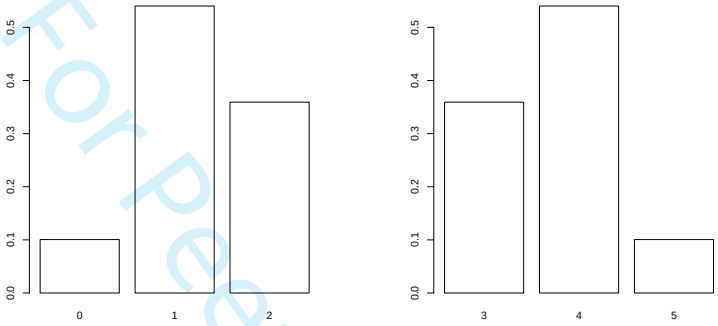

Figure 4. Estimated predictive distributions $ZR_7|\mathcal{FM}_7$ (left) and $Zr_7|\mathcal{FM}_7$ (right).

HPD sets contain the true values of these parameters. Figure 3 shows the consistency of the squared-error loss estimates of the three parameters. Notice that we obtained a better result for the r genotype, since the R genotype becomes extinct almost surely.

Figures 4 and 5 illustrate the predictive posterior distributions of the total number of females, and of males and mating units of each type in the next generation. One observes from Table 3 that the sample size is sufficient for accurate estimates to be made. The predicted behaviour in this generation is in keeping with the fact that the R genotype becomes extinct almost surely and there is a positive probability of the r genotype's survival.

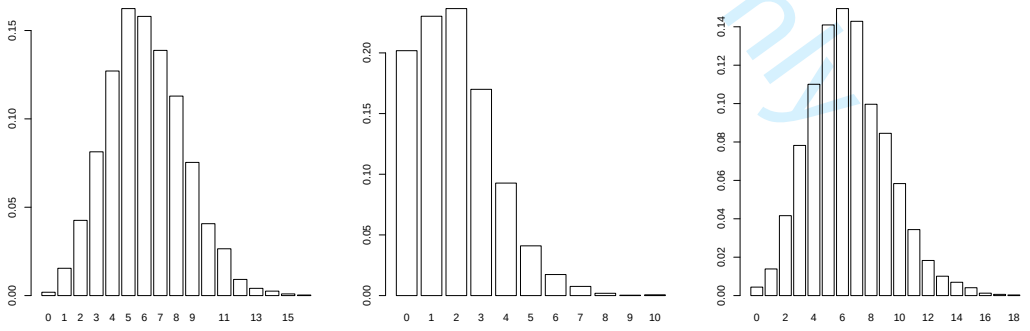

Figure 5. Estimated predictive distributions $F_8|\mathcal{FM}_7$ (left), $MR_8|\mathcal{FM}_7$ (middle), and $Mr_8|\mathcal{FM}_7$ (right).

Concentration parameter	Poisson			Geometric		
	MEAN	HPD 95%		MEAN	HPD 95%	
0.50	0.42046	0.31491	0.52405	0.42109	0.31407	0.53500
0.75	0.42087	0.31168	0.52872	0.42023	0.31258	0.53264
1	0.42300	0.31920	0.52487	0.42156	0.31157	0.52776
5	0.42073	0.31392	0.53252	0.42024	0.31228	0.52856
10	0.41694	0.30804	0.52225	0.41891	0.31220	0.52103
20	0.42088	0.32049	0.52820	0.41943	0.31321	0.52349

Table 4. Sensitivity analysis for $\alpha|\mathcal{FM}_7$.

Concentration parameter	Poisson			Geometric		
	MEAN	HPD 95%		MEAN	HPD 95%	
0.50	2.48592	1.68230	3.48328	2.43697	1.46135	3.59668
0.75	2.48394	1.50657	3.54474	2.41127	1.42340	3.55337
1	2.49595	1.60745	3.32928	2.42343	1.36314	3.56445
5	2.52246	1.79568	3.37029	2.48820	1.51501	3.72488
10	2.54912	1.88674	3.27969	2.51921	1.62203	3.69884
20	2.58385	2.02126	3.20585	2.53486	1.64409	3.64686

Table 5. Sensitivity analysis for $m_R|\mathcal{FM}_7$.

4.1. Sensitivity and robustness analysis

In the previous section, as prior distributions for p^R and p^r we assumed Dirichlet processes with concentration parameter 1 and a Poisson distribution base measure. We now describe a discrete sensitivity analysis carried out to study the influence of the prior parameters. For that, we considered Poisson and geometric distributions as base measure, and different values for the concentration parameter. In Tables 4-6 we present the estimates of α, m_R, and m_r under squared-error loss as well as their 95% HPD sets, obtained in the sensitivity analysis. For α, we obtained similar figures in all cases, since the posterior distribution depends only on $\mathcal{FM}_7$ (see Equation (3)). For m_R and m_r, the squared-error loss estimates tended to be close to the mean of the base measures as the concentration parameter increased (as usual). Anyway, one can conclude that those initial values do not significantly influence the estimation of the parameters. Moreover, in all cases the HPD sets contained the true values of m_R and m_r, although their ranges were slightly greater for the geometric base than for the Poisson base.

Finally, we performed a robustness analysis by means of a series of simulated examples taking $\alpha = 0.45$ and considering different values of m_R and m_r. In this way, we analysed all forms of the asymptotic behaviour of the alleles: coexistence, fixation, and extinction (see [6] and [7] for a detailed description of such behaviour). To run the

Concentration parameter	Poisson			Geometric		
	MEAN	HPD 95%		MEAN	HPD 95%	
0.50	2.97716	2.18279	3.87841	2.99914	2.03458	4.20865
0.75	2.94969	2.07192	3.83351	2.99217	2.00224	4.00436
1	2.97439	2.19181	3.83126	3.01474	1.94133	4.22338
5	2.92762	2.18572	3.70971	2.97231	1.95779	4.14504
10	2.88905	2.31871	3.51036	2.90738	1.96482	3.94306
20	2.85999	2.29291	3.42091	2.89028	2.03489	3.92406

Table 6. Sensitivity analysis for $m_r|\mathcal{FM}_7$.

PARAMETERS			MEAN		95% HPD				PATH
					m_R		m_r		TYPE
α	m_R	m_r	m_R	m_r					
0.45	2.5	3.5	2.8694	3.1680	2.0054	3.9306	2.2497	4.1054	(1)
	2.5	1.5	2.4391	2.5810	1.7398	3.2455	1.9967	3.2252	(2)
	1.5	2.5	2.2921	2.3216	1.5568	3.1686	1.7515	3.0627	(3)
	2	1.5	2.2664	2.2558	1.3891	3.2645	1.4150	3.2761	(4)

Table 7. Application of the method for different values of m_R and m_r, with $\alpha = 0.45$, based on samples corresponding to paths for which (Type 1) both genotypes have survived, (Type 2) the R genotype has survived and r has become extinct, (Type 3) the R genotype has become extinct and r has survived, and (Type 4) both genotypes have become extinct.

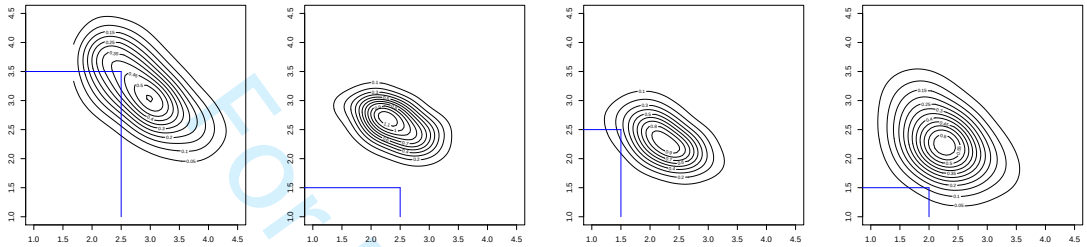

Figure 6. Contour plots of $(m_R, m_r)|\mathcal{FM}_7$ obtained from application of the Gibbs algorithm to samples with the different values of (α, m_R, m_r) given in Table 7. First graphic corresponds with sample for which both genotypes survive; second plot with that for which R genotype survives and r becomes extinct; third one with that for which R genotype becomes extinct and r survives and fourth plot with that in which both genotypes become extinct.

experiment, we assumed the observation of the first 7 generations considering there to be a positive number of males of both genotypes in the last generation. We then applied the method studied here, observing that the 95% HPD sets contained the true value of the parameters in all cases except the fixation cases for the genotype that will become extinct. Table 7 presents a summary of the different parameters used in the examples together with their squared-error loss estimates and 95% HPD sets. Figure 6 shows the $(m_R, m_r)|\mathcal{FM}_7$ contour plots for each sample. One observes that the posterior distributions related to the surviving genotype are very accurate. However, as is usual in a branching process context (see [25]), the results in some fixation or extinction cases were less accurate.

5. Application to real data

In [13], an 84-member pedigree (of whom 76 were still alive) of a Chinese family of Tujia ethnicity with non-syndromic hearing impairment was studied. All the affected individuals are patrilineal males, indicating that their hearing loss is a hereditary Y-linked pattern. As females of this family do not present the disorder and mate with unrelated males (who present normal hearing), two types of Y-linked alleles are present: one which transmits hearing loss (the allele of interest, termed R) and the other which transmits normal hearing (allele r). This situation can thus be modeled using a Y-linked two-sex branching process with blind choice (the data shows that the hearing impairment does not affect mate selection).

In this section, we apply our proposed method to the real data provided in the aforementioned study [13] in which only an R family line is observed. This pedigree can be considered to be part of a branch of an entire population tree. However, since not all the generation sizes of the process are observed but only those of one branch,

the mating phase of our model needs to be modified in order to fit the data.

Prior to explaining this modification, we must first indicate that we shall use the notation (α, m_R, m_r) for the parameters of interest and $(F_n, MR_n, Mr_n, ZR_n, Zr_n)$ for the variables, although in this latter case they correspond only to the observed pedigree, not to the whole population (which is unknown). To facilitate reading the rest of the section, we shall not introduce any new notation.

From analysing the data (see [13]), we were able to conclude that the characteristics of the evolution of the family tree were:

(1) All males mate, as is usual in patriarchal societies, with perfect fidelity.
(2) Not all females mate, and the proportion of females who do depends on the ratio between the two types of males at each generation (seeking to balance the number of couples of each type in that generation). This behaviour has also been found in other databases (see [26]).
(3) At each generation, females generated in the family tree only form couples with external r-males (assuming that the possibility of a female mating with an external R-male is negligible).
(4) Mating with relatives is not allowed.

Given these characteristics, our family tree can be considered to lie within an entire population tree which includes a sufficient number of females at each generation for all the males in the family to mate with, so that $ZR_n = MR_n$, and $Zr_n = Mr_n + X_n$, with X_n being the number of females generated in the family tree who mate. These females mate with males who do not belong to the pedigree, and, in accordance with characteristic (3), are always r-males. Considering characteristic (2), it is reasonable to model X_n by a binomial distribution with size F_n and probability MR_n/M_n. Indeed, when the number of r-couples formed by r-males generated in the family tree is small in comparison with the number of R-couples, many females will tend to mate with external r-males in order to balance the population, and then MR_n/M_n should be large (i.e., close to unity). And vice versa, when the number of those r-couples is large in comparison with the number of R-couples, few females will tend to mate, and MR_n/M_n should be small.

This adaptation of the mating phase of our general model involves only a slight change in the calculation of the probability in Equation (8) when implementing the method, using now a binomial instead of a hypergeometric distribution. Before applying the method, let us first describe the sample. The data given in [13] start with an R-male and contain the complete family tree generated by this male over 4 generations (including couples and offspring). Both the number of couples in the third generation and the offspring forming the last generation can be considered open because in the first case some couples may give birth to further offspring, and in both cases some individuals may be too young to mate.

In view of this consideration, although the complete family tree is available, in order to apply the method we only consider as observed sample the total numbers of female and male offspring in the first three generations who have a direct patrilineal relationship with the first couple observed in this family. This information is given in Table 8. For the males, the type is only distinguished in the last (third) generation, with $M_3 = 10$ being split into $(MR_3, Mr_3) = (7, 3)$. These data form the vector $\mathcal{FM}_3$. Since the third generation couples and the fourth generation individuals are still open, they are not included in the sample, although they will be useful in evaluating the goodness of the predictions given by the method.

n	0	1	2	3
F_n	0	1	3	9
M_n	1	4	6	10

Table 8. Real data.

Since the pedigree starts with an R-male, we consider $M_0 = MR_0 = 1$, and then $(ZR_0, Zr_0) = (1, 0)$, which allows the probabilities of Expressions (6) and (8) to be obtained trivially. Moreover, since $(F_1, M_1) = (1, 4)$, that initial couple generates 5 offspring. This information could be useful in estimating the initial prior common reproductive ability, so that for both offspring reproduction laws we assume a Dirichlet process with concentration parameter 1 and Poisson distribution base measure of mean 5. This choice will not affect the final estimate since, as was shown in the previous section, the method is insensitive to the initial values of the base distribution. In estimating the parameter α, we assume no prior information is available so that we again consider (as in the simulated example) a beta prior distribution with parameters $\beta_1 = \beta_2 = 1$.

Now we apply the algorithm proposed in this paper, and simulate $T = 20$ chains formed by 20 000 iterations of the method. The Gelman-Rubin-Brooks diagnostic tools indicate that a sufficient burn-in period is $L = 5000$ with a batch size of $G = 300$. We thus obtain a sample of size 1020 ($Q = 50$) from the conditional distribution of the parameter vector $(\alpha, p^R, p^r)|\mathcal{FM}_3$.

Table 9 presents some summary statistics for the posterior distributions of α and the reproduction means m_R and m_r, and predictive posterior distributions of the variables ZR_3, Zr_3, F_4, MR_4, and Mr_4.

Since we have considered that all males of type R mate, $\{ZR_n\}_{n\geq 0}$ behaves as a Galton-Watson process with mean growth rate $(1-\alpha)m_R$. It is therefore important to estimate this parameter in order to determine the fate of the R-allele in the population. The corresponding results are given in Table 9, and Figure 7 (left) shows the parameter's estimated posterior density. From this estimated posterior density, we obtain that the probability of the parameter's being greater than 1 is $P((1-\alpha)m_R > 1|\mathcal{FM}_3) \simeq 0.9775$, with a Bayes factor equal to 43.3478, clearly greater than unity. Hence, one can conclude that the mean growth rate is greater than 1, and therefore, applying branching process theory, that there exists a positive probability for hearing impairment not to disappear from this family in the following generations. This is also reflected in the prediction of ZR_3 (notice that this is not random once MR_3 has been observed) and in the predictive posterior distribution of MR_4 shown in Figure 7 (right). The data set has 5 R-males in the fourth generation, although 2 male offspring were still too young to be diagnosed at the time of examination and some couples of the third generation may still give birth to new offspring. Our prediction, with its apparent overestimate of a value of 5 could therefore still be considered adequate. Besides the study of the fate of the R-allele, the results allow the reproductive ability of the r-allele to be quantified and compared with that of the R-allele. Figure 8 (left) shows the contour plot of the estimated joint posterior distribution of (m_R, m_r) given $\mathcal{FM}_3$. We estimated that $P(m_R > m_r|\mathcal{FM}_3) \simeq 0.6520$ with a Bayes factor equal to 1.8732, also greater than unity. Hence, one can conclude that the mean number of offspring per R-couple is greater than the mean number of offspring per r-couple. Nonetheless, we also predict that the r-allele will persist in the population even though it has a lower reproductive ability than the R allele (see Figure 8 (right)).

	α	m_R	m_r	$(1-\alpha)m_R$	ZR_3	Zr_3	F_4	MR_4	Mr_4
MEAN	0.3995	2.7538	2.4230	1.6518	7	9.3216	17.0284	11.5588	13.6324
SD	0.0799	0.5623	0.9296	0.3994	0	1.3556	6.7140	4.4510	7.1394
MCSE	0.0025	0.0176	0.0291	0.0125	0	0.0425	0.2102	0.1394	0.2235
TSSE	0.0025	0.0176	0.0291	0.0125	0	0.0445	0.2102	0.1394	0.2235

Table 9. Summary statistics for the posterior distributions of α, m_R, m_r, $(1-\alpha)m_R$,ZR_3, Zr_3, F_4, MR_4, and Mr_4, given $\mathcal{FM}_3$.

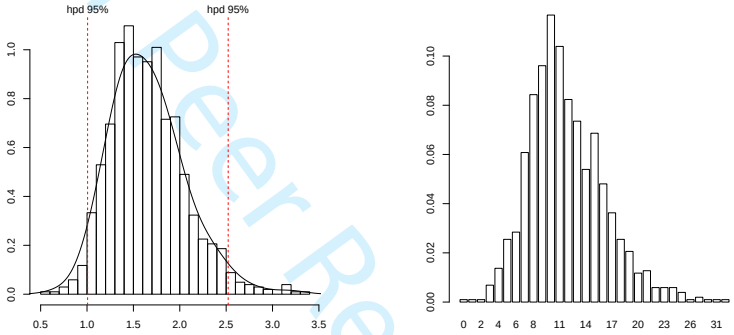

Figure 7. Estimated density for $(1-\alpha)m_R|\mathcal{FM}_3$ with 95% HPD set (left) and estimated predictive distribution $MR_4|\mathcal{FM}_3$ (right).

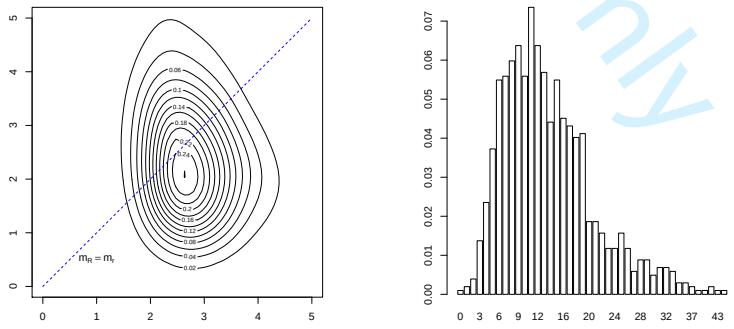

Figure 8. Contour plot of $(m_R, m_r)|\mathcal{FM}_3$ (left) and estimated predictive distribution $Mr_4|\mathcal{FM}_3$ (right).

Remark 3. The software environment for statistical computing and graphics **R** ("GNU S", see [27]) was used to perform the simulation study by means of parallel computing. We used the CODA package (see [28]) to analyse the convergence of the method, and the GenKern package (see [29]) for the two-dimensional kernel density estimation.

6. Concluding remarks

A procedure for drawing inferences on the reproductive abilities of a Y-linked gene has been developed by using a two-sex branching process with blind choice. This stochastic model is suitable for analysing the evolution of the number of carriers of two alleles of a Y-linked gene in a two-sex monogamous population in which each female chooses her partner from among the male population without caring about his type, since either it is not expressed in his phenotype or, if it is expressed, it is not decisive at mating time. This assumption led us to consider a sampling scheme based on real situations in nature in which the total numbers of females and males (with unrecognized genotypes) are observed in each generation up to some given generation. Based on this sample, and adding the information corresponding to the total number of males of each genotype in the last generation, we took a Bayesian approach to the inference problem in a non-parametric framework, without assuming any knowledge of the reproduction laws. We then used a Dirichlet process to deal with the problem of ignorance of the cardinality of the reproduction laws' supports. The problem was considered to be an incomplete data estimation problem with a latent sequence structure, and solved by applying the Gibbs sampler (a Markov chain Monte Carlo method). There have been other approaches to the problem of estimating parameters from incomplete sample data in the context of branching processes (e.g., [30] and [31] based on a Bayesian perspective, and [32],[33], and [34] based on the EM algorithm). An essential difference with the present case is that in those other studies it was possible to construct the latent sequences independently generation by generation. Since the sample considered in the present case does not have a Markovian structure, this was impossible, and therefore the distribution of latent sequences depends on past and future observations (see Equation (1)). Applying the proposed method, we approximated the posterior distributions of the main parameters of the model and the predictive distributions for as yet unobserved generations. In a simulation study, we confirmed the accuracy, efficiency, and robustness of the algorithm in making inferences about the main parameters of the model, in spite of how small an amount of information the sample represents. This kind of small sample in terms of population size and number of observed generations is usual in many population studies (see the Introduction). In this sense, it is also important to point out that the results were satisfactory despite the smallness of the data set (in our simulated example, we observed 7 generations with fewer than 15 individuals in each generation, and in the real example, there were only 3 generations with 9 individuals at most). Obviously the method is scalable, but the greater the dimensionality, the more computationally expensive its application will be. The results are more accurate for both genotypes when they have both survived, or for the surviving genotype in the case of fixation. The method is readily adaptable to mating schemes that differ from the one initially considered in the present study. By way of illustration of this fact, we presented the analysis of a real data set corresponding to hereditary Y-linked hearing impairment in a Chinese family of Tujia ethnicity.

In conclusion, to obtain informative posterior distributions of the reproduction laws, it is enough to observe the total numbers of females and males in each generation together with the number of males of each genotype in just the last generation. Moreover, note that it is unnecessary either to observe the total number of couples of each type in any generation or to have any prior information about the reproductive abilities (the method is robust to the choice of the base distribution of the Dirichlet process). The results that have been presented show that our procedure provides a useful framework in which to model real genetic problems, and illustrates the power of our non-parametric Bayesian approach.

Funding

The research was supported by grant MTM2015-70522-P (MINECO/FEDER, UE) and grant IB16103 (Junta de Extremadura/European Regional Development Fund).

Acknowledgements

The authors would like to thank the Associate Editor and the Reviewer for providing invaluable comments and suggestions which have significantly improved this paper.

References

[1] Chen G, Chen S. A stochastic model relating the number of cells at X-inactivation to the allelic ratio in normal heterozygous adult females. Biometrical Journal. 2003;6:758–771.
[2] Mode CJ, Sleeman CK. Stochastic processes in genetics and evolution: Computer experiments in the quantification of mutation and selection. World Scientific; 2012.
[3] Mode CJ, Sleeman CK, Raj T. On the inclusion of self regulating branching processes in the working paradigm of evolutionary and population genetics. Front Gene. 2013;4:11:doi: 10.3389/fgene.2013.00011.
[4] Neves A, Moreira C. Applications of the Galton-Watson process to human DNA evolution and demography. Physica A. 2006;368:132–146.
[5] González M, Hull DM, Martínez R, et al. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. Math Biosci. 2006;202:227–247.
[6] González M, Martínez R, Mota M. Bisexual branching processes to model extinction conditions for Y-linked genes. J Theor Biol. 2009;258:478–488.
[7] Alsmeyer G, Gutiérrez C, Martínez R. Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice. J Theor Biol. 2011;275:42–51.
[8] González M, Martínez R, Mota M. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. Math Biosci. 2008;215:167–176.
[9] González M, Gutiérrez C, Martínez R. Parametric Bayesian inference for Y-linked two-sex branching models. Stat and Comput. 2013;23:727–741.
[10] Lucena-Perez M, Soriano L, López-Bao J, et al. Reproductive biology and genealogy in the endangered Iberian lynx: Implications for conservation. Mamm Biol. 2018;89:7–13.
[11] Oliveira L, Fraga L, Majluf P. Effective population size for South American sea lions along the Peruvian coast: the survivors of the strongest El Niño event in history. J Mar Biol Assoc UK. 2012;92(8):1835–1841.
[12] Kokko H, Lindström J, Ranta E, et al. Estimating the demographic effective population size of the Saimaa ringed seal (*Phoca hispida saimensis* Nordq.). Animal Conservation. 1998;1:47–54.

[13] Fu S, Yan J, Wang X, et al. The audiological characteristics of a hereditary Y-linked hearing loss in a Chinese ethnic Tujia pedigree. Int J Pediatr Otorhinolaryngol. 2011; 75(2):202–206.
[14] González M, Gutiérrez C, Martínez R. Extinction conditions for Y-linked mutant-alleles through two-sex branching processes with blind-mating structure. J Theor Biol. 2012; 307:104–116.
[15] Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Statist. 1973;1 (2):209–230.
[16] Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Statist. 1998;7:434–455.
[17] Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. Statist Sci. 1992;7 (4):457–511.
[18] Tierney L. Markov chain for exploring posterior distribution (with discussion). Ann Statist. 1994;22:1701–1762.
[19] Browman A, Azzalini A. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Oxford University Press; 1997.
[20] Baum L, Petrie G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annal Math Statist. 1970;41(1):164–171.
[21] Baum L. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Proc 3rd Symposium Inequalities III, Univ Calif. 1972;:1–8.
[22] Bertoin J, Fontbona J, Martínez S. On prolific individuals in a supercritical continuous-state branching process. J Appl Probab. 2008;3:714–726.
[23] Yanev G. Empirical Bayes estimators for the reproduction parameter of Borel-Tanner distribution. Applied Statistics Research Progress. 2008;:27–33, Nova Sci. Publ., New–York.
[24] Molina M, González M, Mota M. Bayesian inference for bisexual Galton-Watson processes. Comm Statist Theory Methods. 1998;27:1055–1070.
[25] Ahsanullah M, Yanev GPE. Records and branching processes. Nova Sci. Publ. Inc.; 2008.
[26] Wang Q, Lu C, Rao S, et al. Y-linked inheritance of non-syndromic hearing impairment in a large Chinese family. J Med Genet. 2004;41(6):e80.
[27] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/.
[28] Plummer M, Best N, Cowles K, et al. coda: Output analysis and diagnostics for mcmc; 2016. R package version 0.19-1; Available from: http://CRAN.R-project.org/package=coda.
[29] Lucy D, Aykroyd R. Genkern: Functions for generating and manipulating binned kernel density estimates; 2013. R package version 1.2-60; Available from: http://CRAN.R-project.org/package=GenKern.
[30] González M, Martín J, Martínez R, et al. Non-parametric Bayesian estimation for multitype branching processes through simulation-based methods. Comp Stat and Data Anal. 2008;52:1281–1291.
[31] González M, Gutiérrez C, Martínez R, et al. Bayesian inference for controlled branching processes through MCMC and ABC methodologies. RACSAM. 2013;107:459–473.
[32] Daskalova N. Nonlinear dynamics of electronic systems. (Communications in Computer and Information Science; Vol. 438.) Springer; 2014. Chapter EM Algorithm for Estimation of the Offspring Probabilities in Some Branching Models; pp. 181–188.
[33] González M, Gutiérrez C, Martínez R. Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes. J Comput Biol. 2012;19(9):1015–1026.
[34] Hautphenne S, Fackrell M. An EM algorithm for the model fitting of Markovian binary trees. Comp Stat and Data Anal. 2014;70:19–34.