ON BRANCHING PROCESSES AND THEIR APPLICATIONS April, 10-13, 2018. Badajoz, Spain.

Approximate Bayesian computation for Y-linked two-sex branching processes with mutations

González, M.; Gutiérrez, C. and Martínez, R.

Department of Mathematics
University of Extremadura
Spain

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

JUNTA DE EXTREMADURA
Consejería de Economía e Infraestructuras

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Contents

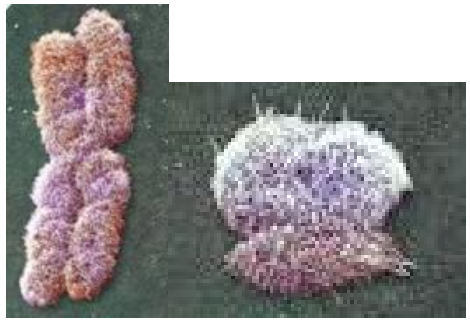
Motivation

Motivation

The sexual chromosomes (X and Y) are directly related with the gender of the individuals:

Females (F)

Males (M)



Y-chromosome:

$\rightarrow$ Two alleles: $\begin{cases} R \text{ original allele} \\ r \text{ mutant allele} \end{cases}$ $\rightarrow$ Individuals: $\begin{cases} F \\ M^R \\ M^r \end{cases}$

r mutant allele:

- Male fertility problems: azoospermia, oligospermia, aspermia.
- Reconstruct the history of paternal lineages.

Population

Mating and Reproduction Phases

$$\left\{\begin{array}{l} F \\ M^R \\ M^r \end{array}\right. \xrightarrow{\text{Mating Phase}} \left\{\begin{array}{l} Z^R \xrightarrow{\text{Reproduction Phase}} \left\{\begin{array}{l} F \\ M^R \\ M^r \text{ (Mutation)} \end{array}\right. \\ Z^r \xrightarrow{\text{Reproduction Phase}} \left\{\begin{array}{l} F \\ M^r \text{ (No backmutation)} \end{array}\right. \end{array}\right.$$

Model assumptions:

- Discrete time model (non-overlapping generations)
- Sexual reproduction
- Two phases: $\left\{\begin{array}{l} \text{Reproduction phase} \\ \text{Mating phase} \end{array}\right.$

Definition of the Model

González M., Gutiérrez C., Martínez R. (2012) *Extinction conditions for Y-linked mutant-allele through two-sex branching processes with blind mating structure.* Journal of Theoretical Biology 307,104-116, 2012.

The reproduction phase

Consider a sequence of i.i.d., non-negative and integer value random vectors

$$\{(F_{nl}^R, M_{nl}^R, M_{nl}^{R\to r}) : l = 1, 2, \ldots, n = 0, 1, \ldots\}$$

Variables

F_{nl}^R : Number of females stemming from the l^{th} R-couple in generation n
M_{nl}^R : Number of males stemming from the l^{th} R-couple in generation n which have preserved the original R-allele
$M_{nl}^{R\to r}$: Number of males stemming from the l^{th} R-couple in generation n whose alleles have mutated and now are of type r

$p_k^R = P(F_{nl}^R + M_{nl}^R + M_{nl}^{R\to r} = k)$, $p^R = \{p_k^R\}_{k\in S^R}$: reproduction law
α : probability for an offspring to be female $(0 < \alpha < 1)$
β : probability of mutation $(0 \leq \beta < 1)$
m_R : reproduction mean

The reproduction phase

Consider a sequence of i.i.d., non-negative and integer value random vectors

$$\{(F_{nl}^R, M_{nl}^R, M_{nl}^{R\to r}) : l = 1, 2, \ldots, n = 0, 1, \ldots\}$$

Variables

F_{nl}^R : Number of females stemming from the l^{th} R-couple in generation n
M_{nl}^R : Number of males stemming from the l^{th} R-couple in generation n which have preserved the original R-allele
$M_{nl}^{R\to r}$: Number of males stemming from the l^{th} R-couple in generation n whose alleles have mutated and now are of type r

$p_k^R = P(F_{nl}^R + M_{nl}^R + M_{nl}^{R\to r} = k)$, $p^R = \{p_k^R\}_{k\in S^R}$: reproduction law

α : probability for an offspring to be female $(0 < \alpha < 1)$

β : probability of mutation $(0 \leq \beta < 1)$

m_R : reproduction mean

The reproduction phase

Consider a sequence of i.i.d., non-negative and integer value random vectors

$$\{(F_{nl}^r, M_{nl}^{r\to r}) : l = 1, 2, \ldots, n = 0, 1, \ldots\}$$

Variables

$(F_{nl}^r, M_{nl}^{r\to r})$: Number of females and males generated by the l^{th} r-couple in generation n

$p_k^r = P(F_{nl}^r + M_{nl}^{r\to r} = k)$, $p^r = \{p_k^r\}_{k\in S^r}$: reproduction law

α : probability for an offspring to be female $(0 < \alpha < 1)$

m_r : reproduction mean

The reproduction phase

The Y-linked Two-Sex Branching Process is a bivariate sequence: $\{(Z_n^R, Z_n^r)\}_{n\geq 0}$

Variables

Z_n^R : Total number of R-couples in the n^{th} generation
Z_n^r : Total number of r-couples in the n^{th} generation

For every $n \geq 0$ and provided that the vector (Z_n^R, Z_n^r) is known:

$$F_{n+1}^R = \sum_{i=1}^{Z_n^R} F_{ni}^R, \quad F_{n+1}^r = \sum_{j=1}^{Z_n^r} F_{nj}^r \text{ and } F_{n+1} = F_{n+1}^R + F_{n+1}^r$$

$$M_{n+1}^R = \sum_{i=1}^{Z_n^R} M_{ni}^R, \quad M_{n+1}^{r\to r} = \sum_{j=1}^{Z_n^r} M_{nj}^{r\to r} \text{ and } M_{n+1}^{R\to r} = \sum_{i=1}^{Z_n^R} M_{ni}^{R\to r}$$

$$M_{n+1}^r = M_{n+1}^{r\to r} + M_{n+1}^{R\to r} \text{ and } M_{n+1} = M_{n+1}^R + M_{n+1}^r$$

The mating phase

- From the vector $(F_{n+1}, M^R_{n+1}, M^r_{n+1}) \rightsquigarrow (Z^R_{n+1}, Z^r_{n+1})$

- Mating mechanism: Perfect fidelity mating

$$Z^R_{n+1} + Z^r_{n+1} = \min\{F_{n+1}, M_{n+1}\}$$

- Blind mating structure:

  - ⋆ If $F_{n+1} \geq M_{n+1}$

$$Z^R_{n+1} = M^R_{n+1} \text{ and } Z^r_{n+1} = M^r_{n+1}$$

  - ⋆ If $F_{n+1} < M_{n+1}$

$$Z^R_{n+1} | (F_{n+1}, M^R_{n+1}, M^r_{n+1}) \sim Hyper(F_{n+1}, M_{n+1}, M^R_{n+1})$$

$$Z^r_{n+1} = F_{n+1} - Z^R_{n+1}$$

The mating phase

- From the vector $(F_{n+1}, M^R_{n+1}, M^r_{n+1}) \rightsquigarrow (Z^R_{n+1}, Z^r_{n+1})$
- Mating mechanism: Perfect fidelity mating

$$Z^R_{n+1} + Z^r_{n+1} = \min\{F_{n+1}, M_{n+1}\}$$

- Blind mating structure:
  - ⋆ If $F_{n+1} \geq M_{n+1}$

$$Z^R_{n+1} = M^R_{n+1} \text{ and } Z^r_{n+1} = M^r_{n+1}$$

  - ⋆ If $F_{n+1} < M_{n+1}$

$$Z^R_{n+1} | (F_{n+1}, M^R_{n+1}, M^r_{n+1}) \sim Hyper(F_{n+1}, M_{n+1}, M^R_{n+1})$$

$$Z^r_{n+1} = F_{n+1} - Z^R_{n+1}$$

The mating phase

- From the vector $(F_{n+1}, M^R_{n+1}, M^r_{n+1}) \rightsquigarrow (Z^R_{n+1}, Z^r_{n+1})$

- Mating mechanism: Perfect fidelity mating

$$Z^R_{n+1} + Z^r_{n+1} = \min\{F_{n+1}, M_{n+1}\}$$

- Blind mating structure:

  - ⋆ If $F_{n+1} \geq M_{n+1}$

$$Z^R_{n+1} = M^R_{n+1} \text{ and } Z^r_{n+1} = M^r_{n+1}$$

  - ⋆ If $F_{n+1} < M_{n+1}$

$$Z^R_{n+1} | (F_{n+1}, M^R_{n+1}, M^r_{n+1}) \sim Hyper(F_{n+1}, M_{n+1}, M^R_{n+1})$$

$$Z^r_{n+1} = F_{n+1} - Z^R_{n+1}$$

Bayesian Inference

González M., Gutiérrez C., Martínez R. *Bayesian inference in Y-linked two-sex branching processes with mutations: ABC approach.* arXiv:1801.09064, 2018 (under review).

Bayesian Inference

- Parameters: $\theta = (\alpha, \beta, m_R, m_r)$.
- Sample: $\overline{\mathcal{FM}}_N = \{(F_n, M_n, n = 0, ..., N, M^R_{N-1}, M^r_{N-1}, M^{R\to r}_N, M^{r\to r}_N)\}$.
- Assumption: The coexistence of both genotypes has been observed at least in the last generation ($F_N > 0$, $M^R_N > 0$ and $M^r_N > 0$)
- Methodology: Approximate Bayesian Computation (ABC).
- Objective: Approximate the posterior distribution $\theta|\overline{\mathcal{FM}}_N$

Approximate Bayesian Computation

Tolerance Rejection-ABC Algorithm

For $i = 1$ to m do
repeat
  generate $(\alpha^{\text{sim}}, \gamma, \phi) \sim U(0,1) \times U(0,1) \times U(0,1)$
  generate $\beta^{\text{sim}} = 0$ with probability γ and
    $\beta^{\text{sim}} \sim \pi(\beta)$ with probability $1 - \gamma$
  generate $m_r^{\text{sim}} = 0$ with probability ϕ and
    $m_r^{\text{sim}} \sim \pi(m_r)$ with probability $1 - \phi$
  generate $m_R^{\text{sim}} \sim \pi(m_R)$
  let $\widetilde{\theta} = (\alpha^{\text{sim}}, \beta^{\text{sim}}, m_R^{\text{sim}}, m_r^{\text{sim}})$
  simulate $\mathcal{FM}_N^{\text{sim}}$ from the likelihood $f(\overline{\mathcal{FM}_N}|\widetilde{\theta})$
until $\rho(\mathcal{FM}_N^{\text{sim}}, \overline{\mathcal{FM}_N}) \leq \epsilon$
set $\theta^{(i)} = \widetilde{\theta}$
end for

Illustration of the methodology

- In practice, we have applied the algorithm generating the parameters vector assuming independent non-informative prior distributions for all parameters.
- Then, we simulate Y-BBPs with mutations considering Poisson distributions as reproduction laws of both genotypes taking into account that we know nothing about the true reproduction laws.
- We have simulated a pool of 50 millions of processes considering a tolerance level equal to 0.00002 quantile of the sample of the distances.
- Such pool of processes is valid for all examples independently of the true values of the parameters as for all simulated examples 15 generations have been generated and started with $F_0 = 10$ and $M_0 = 10$

Illustration of the methodology

We will illustrate, by means of simulated examples, the approximate posterior distribution $\theta|\overline{\mathcal{FM}}_N$ considering different situations observed in the sample:

- Case 1: Observing $M_N^R > 0, M_N^{R\to r} > 0$ and $M_N^{r\to r} > 0$
- Case 2: Observing $M_N^R > 0, M_N^{R\to r} > 0$ and $M_N^{r\to r} = 0$
- Case 3: Observing $M_N^R > 0, M_N^{R\to r} = 0$ and $M_N^{r\to r} > 0$

In any case, the estimation of the posterior distribution of α is always very accurate because, in long term, roughly speaking, such parameter is the quotient between the total number of females and the total number of individuals, which are observed in all generations. We focus then on the estimates of the posterior distributions of the rest of parameters: β, m_R and m_r

Case 1: $M_N^R > 0, M_N^{R\to r} > 0$ and $M_N^{r\to r} > 0$

- In this case, only simulated paths where $\beta^{\text{sim}} > 0$ and $m_r^{\text{sim}} > 0$ have been considered.
- To evaluate the distance between the observed and the simulated data, we have used, re-scaling each coordinate of the vectors,

$$\rho(\mathcal{FM}_N^{\text{sim}}, \overline{\mathcal{FM}}_N) = \left(\sum_{n=1}^{N} \left(\frac{F_n^{\text{sim}}}{F_n} - \frac{F_n}{F_n^{\text{sim}}} \right)^2 + \sum_{n=1}^{N-2} \left(\frac{M_n^{\text{sim}}}{M_n} - \frac{M_n}{M_n^{\text{sim}}} \right)^2 \right.$$
$$+ \left(\frac{M_{N-1}^{R\ \text{sim}}}{M_{N-1}^R} - \frac{M_{N-1}^R}{M_{N-1}^{R\ \text{sim}}} \right)^2 + \left(\frac{M_{N-1}^{r\ \text{sim}}}{M_{N-1}^r} - \frac{M_{N-1}^r}{M_{N-1}^{r\ \text{sim}}} \right)^2 + \left(\frac{M_N^{R\ \text{sim}}}{M_N^R} - \frac{M_N^R}{M_N^{R\ \text{sim}}} \right)^2$$
$$\left. + \left(\frac{M_N^{R\to r\ \text{sim}}}{M_N^{R\to r}} - \frac{M_N^{R\to r}}{M_N^{R\to r\ \text{sim}}} \right)^2 + \left(\frac{M_N^{r\to r\ \text{sim}}}{M_N^{r\to r}} - \frac{M_N^{r\to r}}{M_N^{r\to r\ \text{sim}}} \right)^2 \right)^{1/2}$$

- For a given $\varepsilon > 0$, known as a tolerance level, the proposed algorithm provides samples from $\pi(\theta \mid \rho(\mathcal{FM}_N^{\text{sim}}, \overline{\mathcal{FM}}_N) \leq \varepsilon)$ which is a good approximation to $\pi(\theta \mid \overline{\mathcal{FM}}_N)$ by using a small enough ε

Simulated Example: Case 1

- True Parameters: $m_R = 3.2,\ m_r = 4,\ \ \alpha = 0.46,\ \ \beta = 0.005$
- Initial vector: $(F_0, M_0^R, M_0^r) = (10, 5, 5)$
- Offspring Reproduction Laws of both genotypes: Non-parametric with finite support $\{0, 1, ..., 7\}$
- Observed Sample: $\overline{\mathcal{FM}}_{15}$

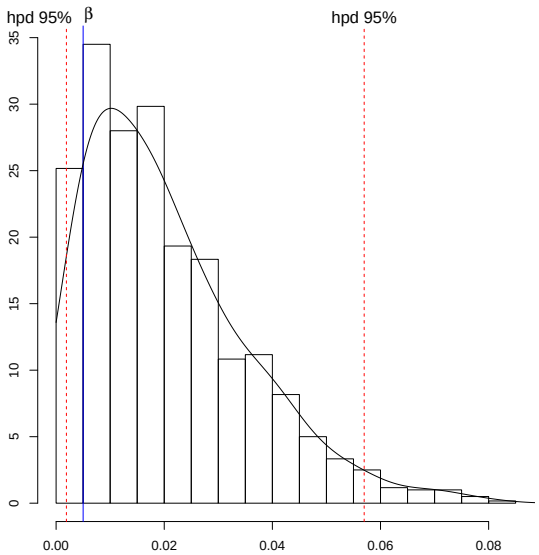
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F_n	16	21	33	53	112	188	342	609	1112	1985	3563	6547	11980	21904	40101
M_n	23	36	46	75	103	215	397	731	1275	2340	4233	7716	13983	25441	46893

with $M_{14}^R = 754, M_{14}^r = 24687, M_{15}^R = 1043, M_{15}^{R\to r} = 6, M^{r\to r} = 45844$

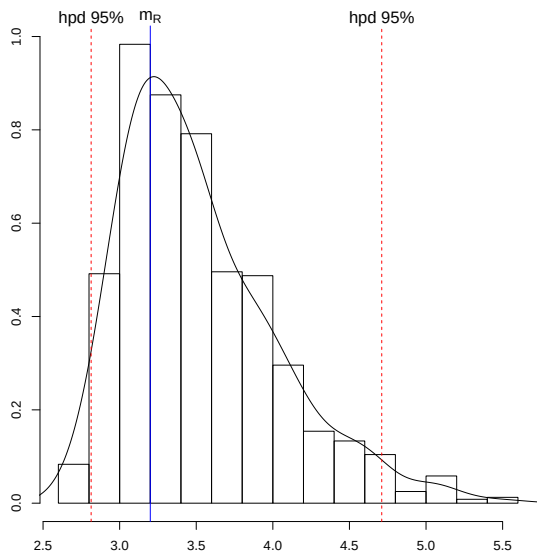
Simulated Example: Case 1

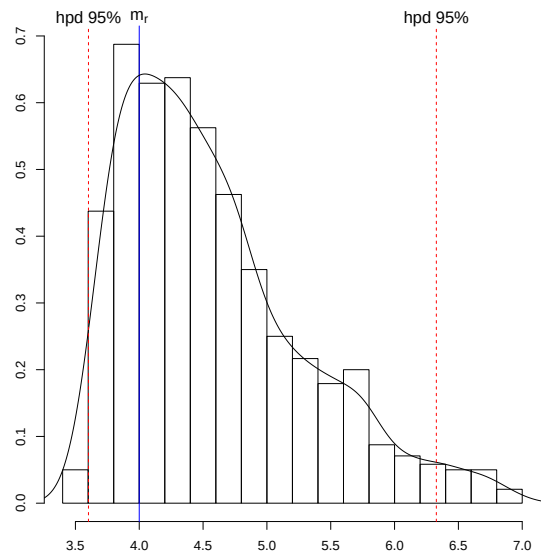
Approximate posterior densities:

$$m_R \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$



$$\beta \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$





$$m_r \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$

Case 2: $M_N^R > 0, M_N^{R\to r} > 0$ and $M_N^{r\to r} = 0$

- In this case, $m_r|\overline{\mathcal{FM}}_N$ should present an atom at zero with positive probability because $M_N^{r\to r} = 0$ could be observed in models where $m_r = 0$ or $m_r > 0$. So the estimation of m_r is difficult in this case.

- Now, only simulated paths $\mathcal{FM}_N^{\mathsf{sim}}$ such that $M_N^{r\to r\mathsf{sim}} = 0$ have been considered.

- To evaluate the distance between the observed and the simulated data, we have considered the metric previously described in Case 1 $\rho(\mathcal{FM}_N^{\mathsf{sim}}, \overline{\mathcal{FM}}_N)$ but deleting the last sum term.

- For a given $\varepsilon > 0$, known as a tolerance level, the proposed algorithm provides samples from $\pi(\theta \mid \rho(\mathcal{FM}_N^{\mathsf{sim}}, \overline{\mathcal{FM}}_N) \leq \varepsilon)$ which is a good approximation to $\pi(\theta \mid \overline{\mathcal{FM}}_N)$ by using a small enough ε

Simulated Example: Case 2

- True Parameters: $m_R = 3,\ m_r = 0,\ \ \alpha = 0.45,\ \ \beta = 0.1$
- Initial vector: $(F_0, M_0^R, M_0^r) = (10, 5, 5)$
- Offspring Reproduction Laws of both genotypes: Non-parametric with finite support $\{0, 1, ..., 7\}$
- Observed Sample: $\overline{\mathcal{FM}}_{15}$

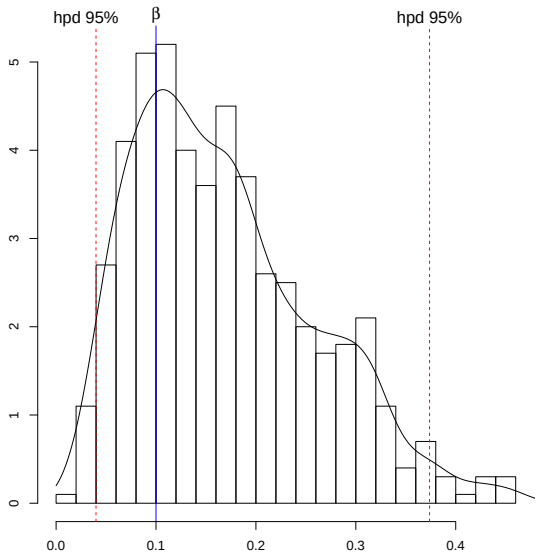
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F_n	6	7	13	8	9	11	15	23	27	34	52	56	70	81	97
M_n	7	7	9	13	7	8	20	22	34	48	48	73	79	108	115

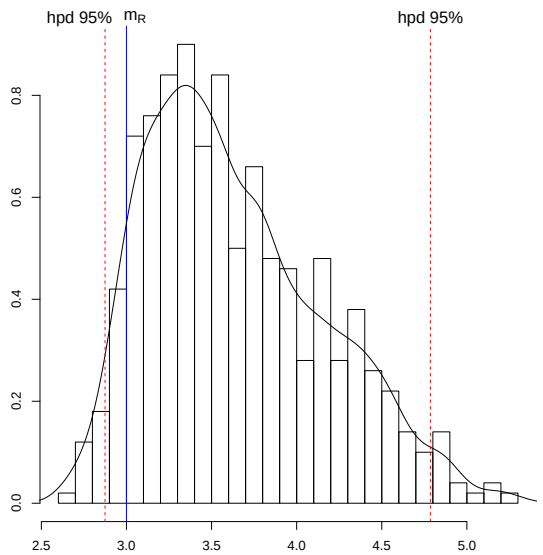
with $M_{14}^R = 96,\ M_{14}^r = 12,\ M_{15}^R = 99,\ M_{15}^{R\to r} = 16,\ M^{r\to r} = 0$

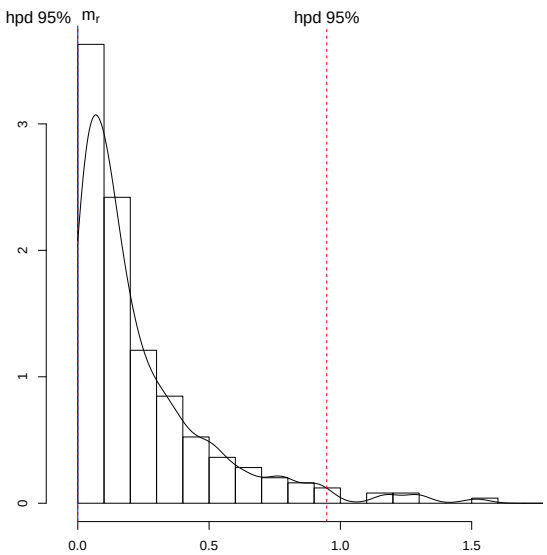
Simulated Example: Case 2

Approximate posterior densities:

$$m_R \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$







$$\beta \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$

$$m_r \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$

$$P(m_r = 0 \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon) = 0.504$$

Case 3: $M_N^R > 0, M_N^{R\to r} = 0$ and $M_N^{r\to r} > 0$

- In this case, $\beta|\overline{\mathcal{FM}}_N$ should present an atom at zero with positive probability because $M_N^{R\to r} = 0$ could be observed in models where $\beta = 0$ or $\beta > 0$. So the estimation of β is difficult in this case.

- Now, only simulated paths $\mathcal{FM}_N^{\mathsf{sim}}$ such that $M_N^{R\to r\mathsf{sim}} = 0$ have been considered.

- To evaluate the distance between the observed and the simulated data, we have considered the metric previously described in Case 1 $\rho(\mathcal{FM}_N^{\mathsf{sim}}, \overline{\mathcal{FM}}_N)$ but deleting the penultimate sum term.

- For a given $\varepsilon > 0$, known as a tolerance level, the proposed algorithm provides samples from $\pi(\theta \mid \rho(\mathcal{FM}_N^{\mathsf{sim}}, \overline{\mathcal{FM}}_N) \leq \varepsilon)$ which is a good approximation to $\pi(\theta \mid \overline{\mathcal{FM}}_N)$ by using a small enough ε

Simulated Example: Case 3

- True Parameters: $m_R = 3,\ m_r = 3.5,\ \ \alpha = 0.65,\ \ \beta = 0.01$
- Initial vector: $(F_0, M_0^R, M_0^r) = (10, 5, 5)$
- Offspring Reproduction Laws of both genotypes: Non-parametric with finite support $\{0, 1, ..., 7\}$
- Observed Sample: $\overline{\mathcal{FM}}_{15}$

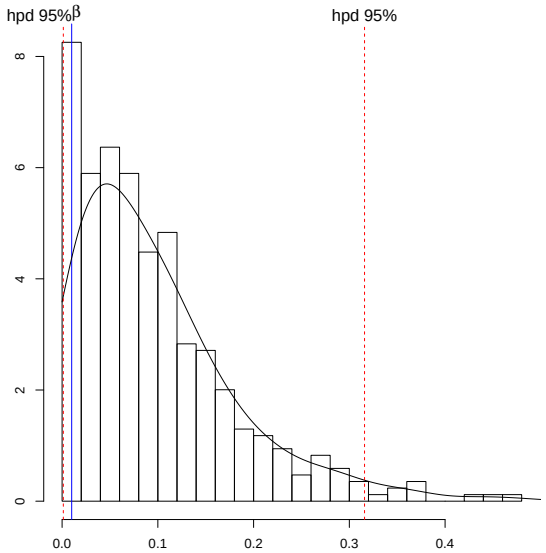
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F_n	24	18	32	23	28	25	45	76	90	112	135	157	185	202	204
M_n	10	14	11	14	16	30	35	41	50	62	73	78	92	88	100

with $M_{14}^R = 11, M_{14}^r = 77, M_{15}^R = 10, M_{15}^{R\to r} = 0, M^{r\to r} = 90$

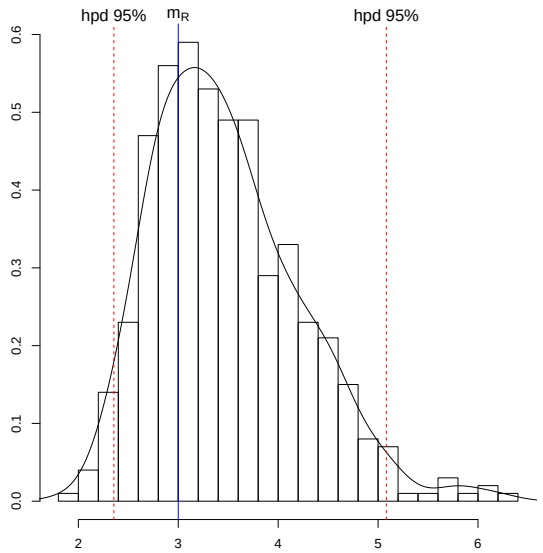
Simulated Example: Case 3

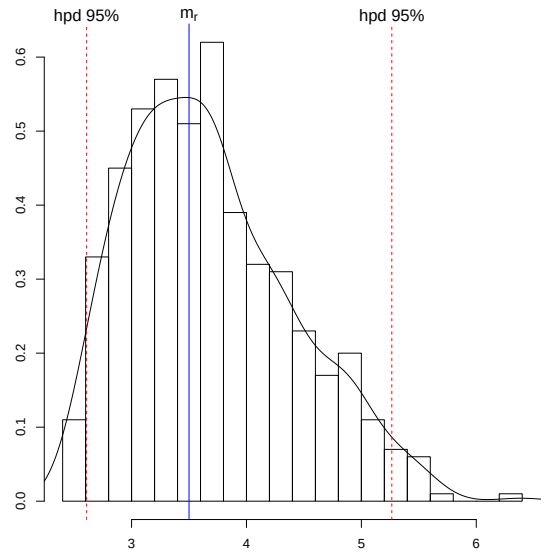
Approximate posterior densities:

$$m_R \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$



$$\beta \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$





$$m_r \mid \rho(\mathcal{FM}_{15}^{\text{sim}}, \overline{\mathcal{FM}}_{15}) \leq \varepsilon$$

Conclusions

1. A two-sex branching process has been presented. It is a suitable model for describing the evolution of a Y-linked gene and its mutations.
2. The aim of the work has been to make inference about the parameters of such model.
3. Bayesian inference is easily made using Approximate Bayesian Computation.
4. Introducing information about the different types of males, only in the last generation, it is possible to obtain accurate approximations to the posterior distributions of the parameters of the model.

Conclusions

1. A two-sex branching process has been presented. It is a suitable model for describing the evolution of a Y-linked gene and its mutations.

2. The aim of the work has been to make inference about the parameters of such model.

3. Bayesian inference is easily made using Approximate Bayesian Computation.

4. Introducing information about the different types of males, only in the last generation, it is possible to obtain accurate approximations to the posterior distributions of the parameters of the model.

Conclusions

1. A two-sex branching process has been presented. It is a suitable model for describing the evolution of a Y-linked gene and its mutations.
2. The aim of the work has been to make inference about the parameters of such model.
3. Bayesian inference is easily made using Approximate Bayesian Computation.
4. Introducing information about the different types of males, only in the last generation, it is possible to obtain accurate approximations to the posterior distributions of the parameters of the model.

Conclusions

1. A two-sex branching process has been presented. It is a suitable model for describing the evolution of a Y-linked gene and its mutations.

2. The aim of the work has been to make inference about the parameters of such model.

3. Bayesian inference is easily made using Approximate Bayesian Computation.

4. Introducing information about the different types of males, only in the last generation, it is possible to obtain accurate approximations to the posterior distributions of the parameters of the model.

Thank you very much!

Acknowledgements:

This research has been supported by the Ministerio de Economía y Competitividad of Spain (grant MTM2015-70522-P), the Junta de Extremadura (grant IB16103) and the Fondo Europeo de Desarrollo Regional (FEDER).