Bayesian estimation for quantification by real-time Polymerase Chain Reaction

Bayesian estimation for quantification by real-time Polymerase Chain Reaction

Nadia Lalam, EURANDOM [1]
Christine Jacob, INRA [2]

Abstract

The aim of the Quantitative Polymerase Chain Reaction is to determine the initial amount X_0 of specific nucleic acids from an observed trajectory of the amplification process, the amplification being achieved through successive replication cycles. This process depends on the efficiency $\{p_n\}_n$ of replication of the molecules, p_n being the probability that a molecule will duplicate at replication cycle n. Assuming $p_n = p$ for all n, we propose to estimate the unknown parameter $\theta = (p, X_0)$ in a Bayesian framework under a Bienaymé-Galton-Watson branching model of the amplification process. The Bayesian approach allows us to take into account some prior information on the parameter. We build and study Bayesian estimators and sets of credibility of the parameter by Markov Chain Monte Carlo methods.

Key words and phrases: Branching processes; Population dynamics; Bayesian inference; Markov Chain Monte Carlo; Quantitative Polymerase Chain Reaction.
2000 Mathematics Subject Classification: 60J85; 62F10; 92B15; 92D25.

1 Introduction

The Polymerase Chain Reaction (PCR) first described by Saiki et al. (1985) is an in vitro enzymatic reaction capable of amplifying the number of copies of a specific DNA fragment. This technique is widely used in molecular biology since it makes it possible to detect very low abundance of DNA. The Reverse Transcription Polymerase Chain Reaction (RT-PCR) is a process in which reverse transcription is done before the amplification (reverse transcription consists in producing a DNA template from an RNA). The RT-PCR procedure allows therefore one to detect low abundance of mRNA. Protocols that not only detect rare nucleic acids but quantitate them as well are increasingly used. The monitoring of DNA molecules as they replicate during PCR is known as real-time or kinetic PCR. The Quantitative PCR (Q-PCR) which aims at determining the initial amount of specific DNA (the target) present in a sample has many applications in disease screening, genes

[1] P.O. Box 513, 5600 MB Eindhoven, The Netherlands
email: lalam@eurandom.tue.nl (corresponding author)

[2] Unit of Applied Mathematics and Informatics, INRA, 78352 Jouy-en-Josas Cedex, France
email: christine.jacob@jouy.inra.fr

expression study or forensic medicine. For more detailed applications, see Ferré (1998).

PCR is formed by the succession of typically 30 to 50 replication cycles. The mechanism of a replication cycle is divided into three steps:
1) The double-stranded DNA is separated into two single strands in a step called denaturation by heating (94 ^oC);
2) The primers, which are short, synthetic sequences of single-stranded DNA complementary to the ends of the target DNA, bind to the target in a step referred to as annealing step (53 ^oC);
3) As the temperature is raised (72 ^oC), a Polymerase enzyme is used to replicate the DNA strands beginning the synthesis process at the region marked by the primers. New double-stranded DNA molecules are synthesized thanks to the activity of the enzyme which facilitates the binding and joining of the complementary nucleotides (deoxy-nucleoside triphosphates or dNTPs) that are free in solutions.

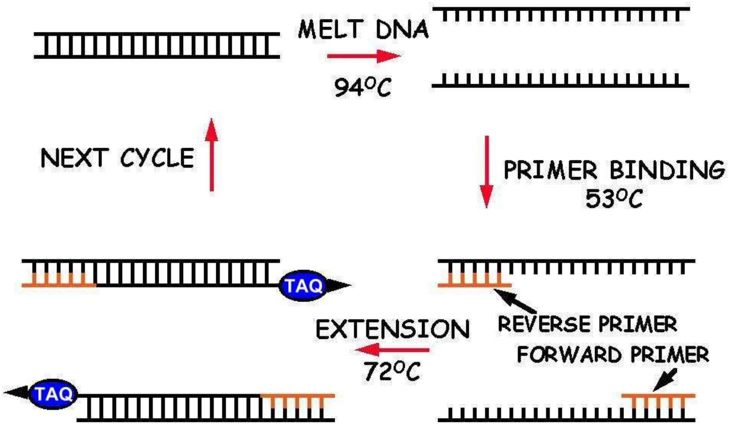

Figure 1: The three steps of a PCR replication cycle (from http://www.surrey-diagnostics.co.uk).

The number of copies of the target DNA is doubled at most at each amplification cycle, but in practice, the probability that a molecule will be successfully duplicated after one cycle, known as the efficiency of the reaction, is less than one. The beginning of PCR is characterized by an exponential increase in target amplification. Then, because of a depletion of reaction components or because of a decline in the Polymerase enzyme activity or because of both (Liu and Saint (2002)), the reaction efficiency slows down and eventually ceases leading to a saturation phase decomposed into a linear phase and a plateau phase.

In the literature, under the assumption of constant reaction efficiency, the theory of Bienaymé-Galton-Watson branching processes (Jagers (1975)) in discrete time, the time step being a replication cycle, has been introduced to model the exponential phase of the PCR for estimating replication errors of the DNA Polymerase (Krawczak et al. (1989), Sun (1995), Weiss and Von Haeseler (1995), Piau (2004, 2005)). A simulation analysis using the coalescence theory has been performed by Weiss and Von Haeseler (1997) providing the Maximum Likelihood Estimator of the replication error rate. There also exists an extensive literature involving branching processes when ignoring the replication errors, that is assuming that all the duplicated molecules are identical to the target template. In this setting that we will consider henceforth, Stolovitzky and Cecchi (1996) studied the number of cycles during which the amplification process undergoes an exponential phase and may therefore be modelled by a single-type supercritical Bienaymé-Galton-Watson branching process for which the reaction efficiency at cycle n, denoted by p_n, satisfies $p_n = p$ for all n. Their approach relied on physical characteristics of the reaction. They also proposed a method for inferring the initial DNA molecules number X_0 when considering two sets of samples S_1 and S_2, each with a given number of d identical preparations with unknown initial DNA molecules number X_0. They considered observations of the molecule numbers at replication cycle n_1 (resp. n_2) belonging to the exponential phase in all the d preparations of sample S_1 (resp. S_2) that they denoted by $X_{n_1,i}$ in the sample preparation i (resp. $X_{n_2,i}$). Computing the corresponding average $\nu_1 = \frac{1}{d}\sum_{i=1}^{d} X_{n_1,i}$ (resp. ν_2), they proposed to estimate the initial DNA molecules number X_0 by the quantity $\nu_1^{-n_2/(n_1-n_2)}\nu_2^{n_1/(n_1-n_2)}$ and the reaction efficiency p by $\nu_1^{1/(n_1-n_2)}\nu_2^{-1/(n_1-n_2)} - 1$. Here and in the sequel, let us denote by X_k the DNA molecules number present at cycle k. Relying on a single trajectory of the PCR amplification process in its exponential phase modelled by a Bienaymé-Galton-Watson branching process, Jacob and Peccoud (1998) built Conditional Least Squares Estimators (CLSE) of the reaction efficiency p of the exponential phase based on $n - h$ consecutive observations of $X_{h+1}, \ldots, X_n$ with either h or $n - h$ fixed as n tends to infinity. They also built the Moment Estimator of the initial DNA molecules number X_0 and constructed its asymptotic confidence interval. Olofsson (2003) gave Maximum Likelihood Estimators of the quantities p and X_0 using a censored Bienaymé-Galton-Watson process. Based on the enzymological approach of the PCR carried out by Schnell and Mendoza (1997), Jagers and Klebaner (2003) modelled the amplification process using a near-critical size-dependent branching process with efficiency $p_n = p(X_{n-1}) = K/(K + X_{n-1})$, where K is a Michaelis-Menten constant type of the reaction. The authors explained then theoretically the existence of the linear part of the saturation phase observed by experimentalists on real-time PCR data. Lalam et al. (2004) studied CLSE of $\{p_n\}_n$ in the frame of a size-dependent branching process with a reaction efficiency model generalizing the one proposed by Jagers and Klebaner (2003) and taking into account the saturation phenomena of the amplification in the modelling of $\{p_n\}_n$, as the size of the amplified population increases.

The previous statistical analyses of Q-PCR were made in a frequentist setting. The aim of the present paper is to perform a Bayesian analysis in order to estimate the reaction efficiency of the exponential phase p and the initial DNA molecules number X_0 from a single amplification trajectory. We will use some prior information on the parameter $\theta = (p, X_0)$ and we will rely on the stochastic modelling of the PCR amplification process during the exponential phase. The model will be a supercritical Bienaymé-Galton-Watson branching process for which the reaction efficiency and the initial DNA molecules number are random variables.
We construct Bayesian estimators and sets of credibility of the parameter θ by Markov Chain Monte Carlo (MCMC) methods. MCMC techniques enable one to carry out simulations from a distribution by embedding it as a limiting distribution of a Markov chain and simulating from the Markov chain until it approaches equilibrium (Gamerman, 1997).

Recall that we will not take into account replication errors during the amplification process and assume therefore that, when the duplication of a target DNA molecule is successful, this creates two DNA molecules identical to the target. We will also consider that the DNA molecule numbers are observed without measurement errors.

We introduce our Bayesian approach for real-time Q-PCR in section 2 and define it more precisely in section 3. Simulation results are given in section 4. We conclude with a discussion in section 5.

2 Bayesian approach

Denote by X_k the DNA molecules number present at replication cycle k, and p_k the replication probability of a molecule at cycle k. During the exponential phase, the reaction efficiency is assumed to satisfy $p_k = p$, for all k. With probability p, if the duplication has been successful, a DNA molecule gives rise to two DNA molecules at the end of a replication cycle. Otherwise, with probability $1-p$, a DNA molecule remains unchanged. This may be modelled by a branching process

$$X_k = \sum_{i=1}^{X_{k-1}} Y_{k,i},$$

where $Y_{k,i}$ is the number of descendants in cycle k of the *i*th molecule from cycle $k-1$. The random variable $Y_{k,i}$ takes either the value 1 or 2. We assume that $\{Y_{k,i}\}_{k,i}$ are independent and identically distributed (i.i.d.) with $P(Y_{k,i} = 2) = p = 1 - P(Y_{k,i} = 1)$, where $0 < p < 1$. Note that the cases $p = 0$ (the molecules never replicate) and $p = 1$ (all the molecules always replicate) are excluded from the analysis since they never occur in practice in real-time PCR experiments. We therefore consider a supercritical Bienaymé-Galton-Watson branching process $\{X_k\}_k$ modelling the exponential phase of the PCR amplification process defined by

$$\begin{cases} X_0 \\ X_k = X_{k-1} + \text{Bin}(X_{k-1}, p),\ k \geq 1 \end{cases}$$

with unknown parameter $\theta = (p, X_0)$, where p is the reaction efficiency of the exponential phase and X_0 is the initial DNA molecules number. The notation $\text{Bin}(N, p)$ stands for a random variable following a binomial distribution with parameters N and p. Note that, experimentally, the exponential phase lasts a random finite number of cycles and is followed by a saturation phase. But in this study, we restrict our attention to the exponential phase only.

We consider a Bayesian framework and use prior information about the model parameter θ in the inference process. Bayesian inference is drawn by constructing the probability distribution of the parameter θ, based on all that is known about it, given the data. This knowledge incorporates previous information about the phenomena under study and it also relies on values of available observed quantities. The information brought by the data is combined with prior information specified by prior distribution yielding the posterior distribution of the parameter. We will determine the posterior distribution of θ and compute the posterior mean as an estimate of the parameter. Let Θ be the parameter set in which θ takes its values. We will denote by x_k the realization of the random variable X_k. Let $\pi(\theta)$ denote the prior distribution of the parameter θ and let $\pi(x_1, \ldots, x_n|\theta)$ denote the likelihood conditionally to θ based on the observations $(x_1, \ldots, x_n)$. According to Bayes' rule, the posterior distribution of θ is given by

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\pi(\theta)\pi(x_1, \ldots, x_n|\theta)}{\int_{\theta' \in \Theta} \pi(\theta')\pi(x_1, \ldots, x_n|\theta')}. \tag{1}$$

The posterior mean, which is the Bayes estimator that we consider, is solution of the minimization problem

$$\min_{\delta} \int_{\Theta} L(\theta, \delta)\pi(\theta)\pi(x_1, \ldots, x_n|\theta)d\theta,$$

for quadratic loss $L(\theta, \delta) = ||\theta - \delta||^2$. We will construct the credibility set of θ which is the confidence interval of the posterior distribution of θ. Both Bayesians and frequentists compute confidence intervals but their interpretations are very different. The interpretation of the Bayesian confidence interval (credibility interval, or highest density region) is that the probability that the true parameter value θ lies in the interval, given the particular data that are actually observed, is equal to the integrated probability of the posterior distribution over the interval. The interpretation of the frequentist confidence interval is the following: one constructs a $100(1-\alpha)\%$ confidence interval $[\widehat{a}, \widehat{b}]$ for a given parameter value θ, where $\widehat{a}$ and $\widehat{b}$ are functions of the data. Then in the long run, in $100(1-\alpha)\%$ of the samples the interval so constructed will contain θ. The frequentist confidence interval uses therefore the variability of the data, given the parameter θ.

3 Model specification

We compute the posterior distribution of θ defined by (1) after the introduction of the prior distributions and the likelihood of the observations given below.

3.1 Prior distributions

In the exponential phase, the reaction efficiency p is assumed to be independent of X_0. This entails that the prior distribution for $\theta = (p, X_0)$ satisfies $\pi(\theta) = \pi(p)\pi(X_0)$, where $\pi(p)$ (resp. $\pi(X_0)$) is the prior distribution of p (resp. X_0).

3.1.1 Prior for p

We choose a non-informative prior distribution $\pi(p)$ on the reaction efficiency of the exponential phase. We will namely take into consideration the so-called Jeffreys distribution which is based on the Fisher information matrix of the likelihood. Such a distribution is motivated by the requirement of invariance property, that is inference should not depend on how the model is parameterized. By definition, the Jeffreys prior distribution is proportional to the square root of the Fisher information. It is more precisely proportional to the square root of the determinant of the Fisher information matrix, but here this matrix is just a real number since $p \in]0,1[$.

Let us determine the Jeffreys prior for p in the setting of the branching process $\{X_k\}_k$ modelling the PCR exponential phase, where X_0 is given. The Fisher information for p equals

$$I_n(p) = E_\theta([\frac{\partial \log \pi(X_1, \ldots, X_n|\theta)}{\partial p}]^2). \tag{2}$$

Let $Z_k = X_k - X_{k-1}$ be distributed as $\mathrm{Bin}(X_{k-1}, p)$ and let $\mathcal{F}_k$ be the sigma-algebra generated by X_0, ..., X_k. For given p and X_0, the log-likelihood of the sample $(X_1, \ldots, X_n)$ is $\sum_{k=1}^n \log f_p(Z_k|\mathcal{F}_{k-1})$, where $f_p(Z_k|\mathcal{F}_{k-1}) = C_{X_{k-1}}^{Z_k} p^{Z_k}(1-p)^{X_{k-1}-Z_k}$ (see subsection 3.2 for the detailed expression of the likelihood). Consequently the Fisher information for p has the following expression, the prime being the derivative with respect to p:

$$\begin{aligned} I_n(p) &= E_\theta([\sum_{k=1}^n \frac{f_p'(Z_k|\mathcal{F}_{k-1})}{f_p(Z_k|\mathcal{F}_{k-1})}]^2) \\ &= \sum_{k=1}^n E_\theta([\frac{f_p'(Z_k|\mathcal{F}_{k-1})}{f_p(Z_k|\mathcal{F}_{k-1})}]^2) + 2\sum_{k<l} E_\theta(\frac{f_p'(Z_k|\mathcal{F}_{k-1})}{f_p(Z_k|\mathcal{F}_{k-1})}\frac{f_p'(Z_l|\mathcal{F}_{l-1})}{f_p(Z_l|\mathcal{F}_{l-1})}), \end{aligned}$$

where

$$\frac{f_p'(Z_k|\mathcal{F}_{k-1})}{f_p(Z_k|\mathcal{F}_{k-1})} = \frac{X_k - (1+p)X_{k-1}}{p(1-p)}.$$

First,

$$\begin{aligned} E_\theta([X_k - (1+p)X_{k-1}]^2) &= E_\theta([\sum_{i=1}^{X_{k-1}} \{Y_{k,i} - (1+p)\}]^2) \\ &= E_\theta(E_\theta([\sum_{i=1}^{X_{k-1}} \{Y_{k,i} - (1+p)\}]^2 |\mathcal{F}_{k-1})) \\ &= E_\theta(\text{var}(Y_{1,1})X_{k-1}) \\ &= p(1-p)(1+p)^{k-1}X_0, \end{aligned}$$

since $E_\theta(X_{k-1}) = E_\theta(E_\theta(X_{k-1}|\mathcal{F}_{k-2})) = (1+p)E_\theta(X_{k-2})$ and by iteration, one gets $E_\theta(X_{k-1}) = (1+p)^{k-1}X_0$.
Second, for $k < l$,

$$\begin{aligned} & E_\theta((X_k - (1+p)X_{k-1})(X_l - (1+p)X_{l-1})) \\ = \; & E_\theta((X_k - (1+p)X_{k-1})E_\theta(X_l - (1+p)X_{l-1}|\mathcal{F}_{l-1})) \\ = \; & 0 \text{ since } E_\theta(X_l - (1+p)X_{l-1}|\mathcal{F}_{l-1}) = 0. \end{aligned}$$

Therefore

$$I_n(p) = \frac{\sum_{k=1}^n (1+p)^{k-1}X_0}{p(1-p)} = \frac{(1+p)^n - 1}{p^2(1-p)}X_0. \tag{3}$$

Consequently, the Jeffreys prior for p is proportional to

$$\sqrt{\frac{I_n(p)}{X_0}} = \sqrt{\frac{(1+p)^n - 1}{p^2(1-p)}}.$$

It would be interesting to determine the Jeffreys prior for p when relying on $(X_h, \dots, X_n)$, where $h \geq 2$ is a replication cycle such that, from this cycle on, the noise inherent to the observations $(x_h, \dots, x_n)$ is negligible. It is indeed well-known that the early observations of real-time PCR trajectories are extremely noisy (Peirson et al. (2003)) and therefore unreliable for the inference. For $h \geq 2$, the likelihood reads

$$\pi(x_h, \dots, x_n|\theta) = [\prod_{k=h+1}^{n} C_{x_{k-1}}^{x_k - x_{k-1}} p^{x_k - x_{k-1}} (1-p)^{2x_{k-1} - x_k}] \pi(x_h|\theta),$$

$$\text{with } \pi(x_h|\theta) = \sum_{x_1, \dots, x_{h-1}} \prod_{k=2}^{h} C_{x_{k-1}}^{x_k - x_{k-1}} p^{x_k - x_{k-1}} (1-p)^{2x_{k-1} - x_k}.$$

This entails that the Fisher information for p based on $(X_h, \dots, X_n)$ equals

$$\begin{aligned} I_{h,n}(p) &= E_\theta([\frac{\partial \log \pi(X_h, \dots, X_n|\theta)}{\partial p}]^2) \\ &= E_\theta(\{\frac{\partial \log \pi(X_h|\theta)}{\partial p}\}^2 + \{\frac{X_n - X_h}{p(1-p)}\}^2 + \{\frac{\sum_{k=h}^{n-1} X_k}{1-p}\}^2 \\ &\quad + \frac{2}{p(1-p)} \frac{\partial \log \pi(X_h|\theta)}{\partial p} (X_n - X_h) \\ &\quad - \frac{2}{1-p} \frac{\partial \log \pi(X_h|\theta)}{\partial p} \sum_{k=h}^{n-1} X_k \\ &\quad - \frac{2}{p(1-p)^2} (X_n - X_n) \sum_{k=h}^{n-1} X_k). \end{aligned}$$

Due to the complex expression of $\frac{\partial \log \pi(X_h|\theta)}{\partial p}$, the computation of $I_{h,n}(p)$ for $h \geq 2$ is difficult. As a consequence, the Jeffreys prior for p based on $\sqrt{I_{h,n}(p)}$ is not straightforwardly obtainable.

In all that follows, we will restrict our analysis to the case $h = 1$ and we will use all the available information from replication cycles 1 to n for inferring the parameter.

3.1.2 Prior for X_0

The initial DNA molecules number X_0 is obtained by extraction of DNA from a biological sample. This can be accounted for by a Poisson distribution (Nedelman et al. (1992)). We therefore propose a Poisson distribution with parameter λ, denoted by Poisson(λ), for the prior distribution $\pi(X_0)$. The Jeffreys principle would lead to put a prior on λ proportional to $1/\sqrt{\lambda}$, but this prior is improper. We will rather assume here that λ is a random variable with uniform distribution of fixed support $[a,b]$. The prior distribution $\pi(X_0)$ thus defined is called a two-stage or hierarchical prior, that is a prior for λ (known as hyper-prior) is put on the parameter of the prior Poisson(λ). The hyper-parameter λ represents the mean of X_0. The choice of the support $[a,b]$, where a and b are constants, has to be selected by the experimenter based on some biological information, e.g. a preliminary approximate range in which X_0 is susceptible to lie.

3.2 Likelihood

We will consider successive observations from the exponential phase ranging from replication cycles 1 to n. As already indicated, we assume that $\{X_k\}_{1\leq k\leq n}$ is observed with no measurement error.

Let us recall that $X_k = X_{k-1} + \text{Bin}(X_{k-1},p)$. Then, for $k \geq 2$,

$$\begin{aligned} P(X_k = x_k|X_{k-1} = x_{k-1},\theta) &= P(X_k = x_k|X_{k-1} = x_{k-1},p) \\ &= P(X_{k-1} + \text{Bin}(X_{k-1},p) = x_k|X_{k-1} = x_{k-1},p) \\ &= P(\text{Bin}(X_{k-1},p) = x_k - X_{k-1}|X_{k-1} = x_{k-1},p) \\ &= C_{x_{k-1}}^{x_k - x_{k-1}} p^{x_k - x_{k-1}} (1-p)^{2x_{k-1}-x_k}. \end{aligned}$$

Hence the likelihood is equal to

$$\begin{aligned} \pi(x_1,\ldots,x_n|\theta) &= [\prod_{k=2}^{n} C_{x_{k-1}}^{x_k - x_{k-1}} p^{x_k - x_{k-1}} (1-p)^{2x_{k-1}-x_k}] \\ &\quad .C_{X_0}^{x_1 - X_0} p^{x_1 - X_0} (1-p)^{2X_0 - x_1} \\ &= (\prod_{k=2}^{n} C_{x_{k-1}}^{x_k - x_{k-1}})[\frac{p}{1-p}]^{x_n - x_1} (1-p)^{s_{n-1}} \\ &\quad .C_{X_0}^{x_1 - X_0} p^{x_1 - X_0} (1-p)^{2X_0 - x_1}, \end{aligned}$$

where $s_{n-1} = \sum_{k=1}^{n-1} x_k$.

3.3 Posterior distribution

We deduce from subsections 3.1 and 3.2 the expression of the posterior distribution of θ denoted by $\pi(\theta|x_1,\ldots,x_n)$. This quantity combines information from the priors and the sample. Recall that we consider the exponential phase with p following a non-informative Jeffreys distribution based on $\sqrt{I_n(p)/X_0}$ and X_0

following a Poisson distribution of parameter λ, with λ uniformly distributed over $[a, b]$. In view of (1),

$$\begin{aligned}
\pi(\theta|x_1,\ldots,x_n) &\propto \pi(\theta)\pi(x_1,\ldots,x_n|\theta)\\
&= \pi(p)\pi(X_0)\pi(x_1,\ldots,x_n|\theta)\\
&\propto \sqrt{\frac{I_n(p)}{X_0}}\int 1_{a\leq\lambda\leq b}\frac{\lambda^{X_0}}{X_0!}e^{-\lambda}d\lambda.[\frac{p}{1-p}]^{x_n-x_1}(1-p)^{s_{n-1}}\\
&\quad .C_{X_0}^{x_1-X_0}p^{x_1-X_0}(1-p)^{2X_0-x_1}.
\end{aligned}$$

Let $J(X_0) = \int_a^b \lambda^{X_0}e^{-\lambda}d\lambda$. Integration by parts yields the relationship $J(X_0) = F(X_0)+X_0J(X_0-1)$, where $F(X_0) = a^{X_0}e^{-a}-b^{X_0}e^{-b}$. By iteration, we deduce that

$$J(X_0) = F(X_0) + X_0F(X_0-1) + X_0(X_0-1)F(X_0-2) + \ldots + X_0!F(1) + X_0!F(0).$$

Therefore, the posterior distribution of θ satisfies

$$\begin{aligned}
\pi(\theta|x_1,\ldots,x_n) &\propto \sqrt{\frac{I_n(p)}{X_0}}\frac{J(X_0)}{X_0!}[\frac{p}{1-p}]^{x_n-x_1}(1-p)^{s_{n-1}}\\
&\quad .C_{X_0}^{x_1-X_0}p^{x_1-X_0}(1-p)^{2X_0-x_1}. \qquad (4)
\end{aligned}$$

The posterior distribution does not have a form belonging to some known distributions family. Due to the analytical intractability of the posterior distribution, one needs to perform simulations (Chen et al. (2000)). This will allow one to sample θ from its posterior distribution and to determine the corresponding Bayesian estimator based on the posterior distribution together with credibility intervals.
We will use the software WinBUGS[3] in order to implement our simulation study. WinBUGS approximates the posterior distribution of θ by Markov Chain Monte Carlo (MCMC) techniques which amount to simulate a Markov chain whose stationary distribution is the joint posterior probability distribution of the parameters of the model. The parameters are first assigned arbitrary initial values, and the chain is simulated until it converges to the stationary distribution. Observations from the chain at stationarity are subsequently used to estimate the joint posterior

[3] The software WinBUGS is publicly available at http://www.mrc-bsu.cam.ac.uk/bugs. Note that the Jeffreys prior density defined in our study is not a classical density present in the reference list of the density distributions available from the software WinBUGS. Therefore, we will follow the ”zeros trick” indicated in the WinBUGS user manual in the section ”Tricks: advanced use of the BUGS language” in order to specify this prior. We will normalize $\sqrt{I_n(p)/X_0}$ so that Jeffreys prior density integrates to 1. The normalizing constant C_n such that $\frac{1}{C_n}\int_0^1 \sqrt{I_n(p)/X_0}dp = 1$ will be obtained using the software Mathematica.

probability of the parameters. This allows one to compute credibility intervals of θ. The MCMC numerical integration technique is widely used for implementation of the Bayes procedure (Gilks et al. (1996)). We will take as Bayesian estimators of θ the mean of the posterior distribution.

As emphasized by Gamerman (1997), one has to keep in mind that no matter how large the MCMC sample is, it only provides a partial substitute for the information contained in the posterior density, that is an approximation to the posterior is constructed via the MCMC technique used when it is not possible to extract information from the posterior analytically. The software WinBUGS uses the Gibbs sampling method as a means for stochastic simulation using Markov chains. Gibbs sampling is a MCMC scheme where the transition kernel is constituted by the full conditional distributions. Denote the distribution of interest by $\mathcal{L}(\theta)$, where $\theta = (\theta_1, \ldots, \theta_d)^T$. Consider that the full conditional distributions $\mathcal{L}_i(\theta_i) = \mathcal{L}(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ are available. Gibbs sampling aims at approximating $\mathcal{L}$ when direct generation schemes are complicated or unavailable but when generations from the $\mathcal{L}_i$ are possible. It provides an alternative generation scheme based on successive generations from the full conditional distributions as follows:
Step 1. Set initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})^T$.
Step 2. Obtain a new value $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_d^{(j)})^T$ from $\theta^{(j-1)}$ through successive generation of values

$$\begin{aligned}
&\theta_1^{(j)} \sim \mathcal{L}(\theta_1 | \theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
&\theta_2^{(j)} \sim \mathcal{L}(\theta_2 | \theta_1^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
&\vdots \\
&\theta_d^{(j)} \sim \mathcal{L}(\theta_d | \theta_1^{(j-1)}, \ldots, \theta_{d-1}^{(j-1)}).
\end{aligned}$$

Step 3. Return to step 2 until convergence is reached.
This scheme defines a Markov chain since the probabilistic change at iteration j depends only on chain values at iteration $j-1$. When convergence is reached, the resulting value $\theta^{(j)}$ is a draw from the stationary distribution $\mathcal{L}$. As the number of iterations increases, the chain approaches its equilibrium. Convergence is then assumed to hold approximately. See chapter 5 of Gamerman (1997) for more detail on Gibbs sampling.

4 Simulation results

We run 20 000 MCMC cycles after a burn-in period of 30 000 cycles, the burn-in cycles being discarded from the analysis. We consider a PCR trajectory consisting in 30 replication cycles for which the true values of the parameters are $p = 0.7$ and $X_0 = 50$.

The estimation summary of the posterior distribution that will be provided consists in the marginal posterior means and credibility intervals of p and X_0 based on the observations from cycles 1 to n and based on the priors

$$\begin{aligned} \pi(p) &\sim \text{Jeffreys' prior depending on } n \\ \pi(X_0) &\sim \text{Poisson}(\lambda) \text{ with } \lambda \sim \text{Uniform}(30, 70). \end{aligned}$$

For different values of n, we present the marginal posterior means, standard deviations and 95 $\%$ credibility intervals for p (resp. X_0) in table 1 (resp. table 2).

n	Mean	Standard deviation	2.5$\%$	97.5$\%$
5	0.7111	0.01477	0.6814	0.7403
10	0.6973	0.003742	0.6899	0.7047
15	0.6997	9.931 10^{-4}	0.6975	0.7014
20	0.7002	2.633 10^{-4}	0.6996	0.7007
25	0.7001	6.91 10^{-5}	0.7	0.7002
30	0.7	1.817 10^{-5}	0.6999	0.7

Table 1: Summary of the results for the parameter p according to n.

n	Mean	Standard deviation	2.5$\%$	97.5$\%$
5	49.99	13.55	27	77
10	49.99	13.55	27	77
15	49.99	13.55	27	77
20	49.99	13.55	27	77
25	49.99	13.55	27	77
30	49.99	13.55	27	77

Table 2: Summary of the results for the parameter X_0 according to n.

As expected, the more observations we consider, that is the greater n, the better the estimate of p since its standard deviation decreases and its 95 $\%$ credibility interval becomes narrower around the true value of p. This suggests a consistency property of the Bayesian estimator of p as n increases analogous to the strong consistency of the CLSE of p proved in the frequentist setting by Jacob and Peccoud (1998). The information for estimating p is brought by the amplification process $\{X_k\}_{1 \leq k \leq n}$. But n has no influence on the estimate of X_0 as can be viewed from formula (4) in which the marginal posterior density of X_0 does not depend on the observations from cycles $[2, n]$: the term $p^{x_n - x_1}(1-p)^{x_1 - x_n + s_{n-1}}$ with $s_{n-1} = \sum_{k=1}^{n-1} x_k$ from (4) does not depend on X_0 and therefore disappears when computing the marginal posterior distribution of X_0. This remark is in accordance with the fact indicated by Jacob and Peccoud (1998) that there is no consistent estimator of X_0 as the number of observations $n - h$ tends to infinity when considering observations of $X_{h+1}, \ldots, X_n$ in the frequentist setting. This can also be noticed from the study of Olofsson (2003) who indicated that the Maximum Likelihood Estimator of X_0 based on a censored process $(X_c, \ldots, X_n)$ is of the order of $X_c/(1+p)^c$. This entails that increasing n does not have an impact on the behavior of this estimator.

5 Concluding remarks

We used the classical modelling of the evolution in time of DNA molecule numbers undergoing the PCR exponential phase by a supercritical Bienaymé-Galton-Watson branching process $\{X_k\}_{0 \le k \le n}$. Relying on this modelling, we performed a Bayesian statistical analysis providing the construction of Bayesian estimators and credibility sets for the parameter $\theta = (p, X_0)$. Our simulation study suggests that the Bayesian estimator of p is consistent. This asymptotic behavior would be similar to the strong consistency of the frequentist CLSE of p proved by Jacob and Peccoud (1998). Another remark coming from the study of table 2 is that the Bayesian estimator of X_0 is not consistent since the credibility set does not improve as n increases. One can deduce this also from the marginal posterior distribution of X_0 (see formula (4) for the joint posterior distribution of θ) which does not depend on n. This is also analogous to the remark of Jacob and Peccoud (1998) made in the frequentist approach that there is no consistent estimator of X_0.

The aim of Q-PCR is to determine the initial DNA molecules quantity. In a frequentist framework, Jacob and Peccoud (1998) constructed an asymptotic confidence interval of X_0, as the replication cycle n tends to infinity. Simulations with finite n were performed in Peccoud and Jacob (1996). Since $\lim_{n\to\infty} X_n(1+p)^{-n} \overset{a.s.}{=} W_{X_0,p}$, where $E(W_{X_0,p}) = X_0$ and $\sigma^2(W_{X_0,p}) = X_0(1-p)/(1+p)$, they considered the Moment Estimator $\widehat{X}_{0,\widehat{n}_s} = X_{\widehat{n}_s}/(1+\widehat{p}_{\widehat{n}_s})^{\widehat{n}_s}$, where the cycle $\widehat{n}_s$ is an estimator of the end of the exponential phase. Defining $f_{n,p}(X_0) = (\widehat{X}_{0,n} - X_0)/\widehat{\sigma}(W_{X_0,p})$ with $\lim_{n\to\infty} f_{n,p}(X_0) \overset{a.s.}{=} \overset{o}{W}_{X_0,p}$, where $\overset{o}{W}_{X_0,p} = (W_{X_0,p} - X_0)/\sigma(W_{X_0,p})$, they used the property $\lim_{n\to\infty} P(f_{n,p}(X_0) \in \overset{o}{W}_{X_0,p,\alpha}) = 1 - \alpha$ for building an asymptotic confidence interval of X_0: $\lim_{n\to\infty} P(X_0 \in f_{n,\widehat{p}}^{-1}(\cup_{x_0} \overset{o}{W}_{x_0,\widehat{p},\alpha})) \ge 1 - \alpha$, where $\widehat{p}$ is the CLSE of p based on the exponential phase. The asymptotic confidence interval at level α is $f^{-1}_{n,\widehat{p}}(\cup_{x_0} \overset{o}{W}_{x_0,\widehat{p},\alpha})$. It would be of interest to compare this asymptotic confidence interval of X_0, when estimating X_0 by the frequentist Moment Estimator, with the credibility intervals when using our Bayesian approach.

The Bayesian estimators of p and X_0 constructed in this simulation study were based on $\{X_k\}_{1\leq k\leq n}$ assumed to be observed without measurement errors. In practice, the initial real-time PCR data are very noisy so that the statistical analysis of the process using real data should not include the first observations which are not reliable. For real-time PCR data whose noisy observations are expressed in fluorescence units, the observed fluorescence F_k at replication cycle k may be modelled by

$$F_k = \alpha X_k + \varepsilon_k \tag{5}$$

with unknown proportionality constant α and disturbance ε_k, as proposed by Peccoud and Jacob (1998) assuming normality of the noise $\{\varepsilon_k\}_k$. Future work consists in using a Bayesian approach to treat real-time PCR data $\{F_k\}$ using the model defined by (5).

Another interesting axis of research concerns the extension of the present Bayesian approach to observations from the saturation phase also. This would be relevant since these data are relatively less noisy. When considering observations belonging to the saturation phase, the efficiency p_n decreases as the replication cycle n increases. Therefore, the unknown parameter would be $\theta = (\{p_n\}_n, X_0)$, where $p_n = p$ for cycle n from the exponential phase, and p_n is a decreasing function of n for cycle n from the saturation phase.

References

[1] Chen M.-H., Shao Q.-M., Ibrahim J. G. (2000) *Monte Carlo methods in Bayesian computation.* Springer-Verlag, New York.

[2] Ferré F. (1998) *Gene quantification.* Ed. Ferré F., Birkhauser, New-York.

[3] Gamerman D. (1997) *Markov Chain Monte Carlo. Stochastic simulation for Bayesian inference.* Chapman and Hall, London.

[4] Gilks W. R., Richardson S., Spiegelhalter D. J. (1996) *Markov Chain Monte Carlo in practice.* Chapman and Hall, London.

[5] Jacob C., Peccoud J. (1998) Estimation of the parameters of a branching process from migrating binomial observations. *Adv. Appl. Prob., 30(4), 948–967.*

[6] JAGERS P. (1975) *Branching Processes with Biological Applications.* John Wiley and Sons, London.

[7] JAGERS P., KLEBANER F. (2003) Random variation and concentration effects in PCR. *J. Theoret. Biol., 224, 299–304.*

[8] KRAWCZAK M., REISS J., SCHMIDTKE J., ROSLER U. (1989) Polymerase chain reaction : replication errors and reliability of gene diagnosis. *Nucleic Acids Res., 17,2197–2201.*

[9] LALAM N., JACOB C., JAGERS P. (2004) Modelling of the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Adv. Appl. Prob., 36(2), 602–615.*

[10] LIU W., SAINT D. A. (2002) Validation of a quantitative method for real time PCR kinetics. *Biochemical and Biophysical Research Communications, 294, 347–353.*

[11] NEDELMAN J., HEAGERTY P., LAWRENCE C. (1992) Quantitative PCR: procedures and precisions. *Bull. Math. Biol., 54, 477–502.*

[12] OLOFSSON U. (2003) Branching processes: Polymerase Chain Reaction and mutation age estimation. *Thesis, Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden.*

[13] PECCOUD J., JACOB C. (1996) Theoretical uncertainty of measurements using quantitative Polymerase Chain Reaction. *Biophysical Journal, 71, 101–108.*

[14] PECCOUD J., JACOB C. (1998) Statistical estimations of PCR amplification rates. In *Gene Quantification.* Ed. Ferré F., Birkhauser, New-York, pp. 111–128.

[15] PEIRSON S. N., BUTLER J. N., FOSTER R. G. (2003) Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. *Nucleic Acids Res., 31(14), e73.*

[16] PIAU D. (2004) Immortal branching Markov processes: averaging properties and PCR applications. *Ann. Probab., 32(1), 337–364.*

[17] PIAU D. (2005) Confidence intervals for nonhomogeneous branching processes and polymerase chain reactions. *Ann. Probab., 33(2), 674–702.*

[18] Saiki R., Scharf S., Faloona F., Mullis K., Horn G. T., Ehrlich H. A., Arnheim M. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science, 239, 487–491.*

[19] Schnell S., Mendoza C. (1997) Enzymological considerations for a theoretical description of the Quantitative Competitive Polymerase Chain Reaction. *J. Theor. Biol., 184, 433–440.*

[20] Stolovitzky G., Cecchi G. (1996) Efficiency of DNA replication in the polymerase chain reaction. *Biophysics, 93, 12947–12952.*

[21] Sun F. (1995) The PCR and branching processes. *J. of Computational Biology, 2(1), 63-86.*

[22] Weiss G., Von Haeseler A. (1995) Modeling the PCR. *J. of Computational Biology, 2(1), 49–61.*

[23] Weiss G., von Haeseler A. (1997) A coalescent approach to the Polymerase Chain Reaction. *Nucleic Acids Res., 25(15), 3082–3087.*