A moment estimator of the initial number of individuals from a population-size-dependent branching process used to model Polymerase Chain Reactions

NADIA LALAM

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2007
A moment estimator of the initial number of individuals from a population-size-dependent branching process used to model Polymerase Chain Reactions

Nadia Lalam

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, July 2007
A moment estimator of the initial number of individuals from a population-size-dependent branching process used to model Polymerase Chain Reactions

NADIA LALAM
Chalmers University of Technology
Department of Mathematical Statistics
SE-412 96 Göteborg, Sweden
E-mail: lalam@math.chalmers.se

Abstract
Consider a binary-splitting population-size-dependent branching process for which the offspring from generation \( n + 1 \), conditionally on the past of the process up to generation \( n \), follows a Bernoulli distribution with probability of success depending on the population size at generation \( n \) according to a Michaelis-Menten type model. We aim at determining the initial population size of the process based on consecutive observations of the process. An estimator based on the method of moments is proposed and its behavior when considering a finite number of observations is given. The model presented here may be used to represent DNA amplification by PCR. In this molecular biology setting, the determination of the starting amount of DNA fragments presents important practical applications.

Key words: Branching process; Moment estimator; Polymerase Chain Reaction.

1 Introduction
Consider a binary splitting population-size-dependent branching process with reproduction rate of Michaelis-Menten type defined as follows. Denote by \( X_n \) the number of individuals present at generation \( n \), and by \( Y_{n+1,i} \) the number of offspring of individual \( i \) belonging to generation \( n \). The number of individuals present at generation \( n + 1 \) is defined by the relationship

\[
X_{n+1} = \sum_{i=1}^{X_n} Y_{n+1,i}.
\]  

(1)
The assumptions that we make on model (1) are:

A1: \( \{Y_{n,i}\}_i \) are independent and identically distributed conditionally on the \( \sigma \)-algebra generated by \( X_0, X_1, \ldots, X_{n-1} \). We will henceforth denote this \( \sigma \)-algebra by \( \mathcal{F}_{n-1} \).

A2: \( E(Y_{n,i}|\mathcal{F}_{n-1}) = m_K(X_{n-1}) \), where

\[
m_K(x) = 1 + \frac{K}{K + x},
\]

where \( K \) is much larger than \( X_0 \), with \( K > 0 \).

A3: \( P(Y_{n,i} = 2) = 1 - P(Y_{n,i} = 1) \).

Assumption A1 entails that the process \( \{X_n\} \) defined by (1) is a branching process. Assumption A2 means that the offspring reproduction rate has a Michaelis-Menten type decreasing shape with an unknown parameter \( K \). The assumption that \( X_0 \ll K \) is useful for defining the moment estimator of \( X_0 \) (see next section). Assumption A3 accounts for the fact that the offspring distribution takes it values in the binary set \( \{1, 2\} \).

Under A1 to A3, \( \{X_n\} \) is a binary-splitting population-size-dependent branching process. We aim at estimation for its initial population size \( X_0 \) based on the observation of \( (X_1, \ldots, X_n) \).


Within the framework of a binary-splitting population-size-dependent branching process satisfying A1-A3, we will define in Section 2 the moment estimator of \( X_0 \) relying on an estimator of the parameter \( K \) from the amplification rate model (2). A Monte Carlo study will be performed to illustrate the behavior of the estimator in Section 3. The branching process investigated here has been proposed by Jagers and Klebaner (2003) to model Polymerase Chain Reaction (PCR). This technique of molecular biology will be described in Section 4 and the potential applications of our study will be discussed.
2 Moment estimator

In order to define the moment estimator of the starting population size of the process \( \{X_n\} \), we will rely on the following martingale \( \{W_n\} \) with respect to the filtration \( \{\mathcal{F}_n\} \) given by

\[
W_n = \frac{X_n}{\prod_{\ell=0}^{n-1} m_K(X_{\ell})}.
\]

Because the parameter \( K \) arising in (3) is unknown, we will replace it either by the conditional least squares estimator \( \hat{K}_n \) based on \( X_h, \ldots, X_n \), with \( h \) fixed, studied by Lalam et al. (2004), or by the estimator \( \hat{K}_{n} \) equal to \( X_n/n \) defined by Jagers and Klebaner (2003). The strong consistency and the rate of convergence of the conditional least squares estimator of \( K \), which minimizes the quantity

\[
\sum_{\ell=h+1}^{n} \left( X_\ell - m_\theta(X_{\ell-1})X_{\ell-1} \right)^2,
\]

with respect to \( \theta \), was established in Lalam et al. (2004). Jagers and Klebaner (2003) proved the strong consistency of their estimator and gave a 95% confidence interval for \( K \).

In view of the relationship \( E(W_n) = X_0 \), we could define the moment estimator of the initial population size \( X_0 \) by

\[
\tilde{X}_{0,n} = \frac{X_n}{\prod_{\ell=0}^{n-1} m_{\hat{K}}(X_{\ell})}.
\]

Because the unknown quantity \( X_0 \) appears in the denominator, we will rather approximate \( m_{\hat{K}}(X_0) \) by 2 by using the fact that, from assumption A2, the value of \( X_0 \) is much smaller than the true value of \( K \). We therefore define the moment estimator by

\[
\hat{X}_{0,n} = \frac{X_n}{2\prod_{\ell=1}^{n-1} m_{\hat{K}}(X_{\ell})}.
\]

3 Numerical simulation

The random variables \( \{Y_{n+1,i-1}\} \) are independent and they follow a Bernoulli distribution with parameter \( m_K(X_n) - 1 \) conditionally to \( X_n \). Using the fact that a sum of \( N \) independent random variables which follow a Bernoulli distribution with parameter \( p \) has a Binomial distribution with parameters \( N \) and \( p \), one can write the population size of the branching process at generation \( n + 1 \) as

\[
X_{n+1} = X_n + \text{Binomial}(X_n, m_K(X_n) - 1).
\]
In the PCR setting and under a Galton-Watson branching process model for \( \{ X_n \} \), the use of the property that \( \{ Y_{n+1,i} - 1 \} \) are independent Bernoulli random variables was already noted in Stolowitzky and Cecchi (1996) in order to represent \( X_{n+1} \) as the sum of \( X_n \) and a binomial random variable.

This rewriting of the model followed by \( \{ X_n \} \) according to (5) enables easy simulations of realizations of the branching process \( \{ X_n \} \). When \( X_n \) is not large, we will use formula (5) to generate a realization of \( X_{n+1} \). But when \( X_n \) is large, say \( X_n \) is greater than \( 10^4 \), then we will rather use an approximation provided by the Poisson theorem. The process \( \{ X_n \} \) satisfies \( \lim_{n \to \infty} X_n \overset{a.s.}{\to} \infty \). Because \( \lim_{N \to \infty} (m_K(N) - 1)N \overset{a.s.}{\to} K \), Poisson’s theorem entails that the Binomial random variables \( \text{Bin}(N, m_K(N) - 1) \) tends in distribution to a Poisson distribution with parameter \( K \) as \( N \) tends to infinity. Therefore, for \( X_n \) large enough (which is realized for \( n \) of the order of tens), say \( X_n \geq 10^4 \), we might use the approximation

\[
X_{n+1} = X_n + \text{Poisson}(K)
\]

to generate a realization of \( X_{n+1} \) from the realization of \( X_n \) and from the value of \( K \).

Simulations of \( \{ X_k \} \), for \( k \) ranging from 1 to 40, are performed with the parameter values \( X_0 = 100 \) and \( K = 10^4 \). Figure 1 represents the plot of the moment estimator \( \hat{X}_{0,n} \) versus the replication cycle \( n \) when \( K \) is estimated by \( X_n/n \). Figure 2 represents the plot of the moment estimator \( \hat{X}_{0,n} \) versus \( n \) when \( K \) is estimated by the conditional least squares estimator based on \( X_h, \ldots, X_n \) with \( h = 20 \) and \( n \geq h + 1 \).

![Figure 1: Simulation 1. On the x-axis: replication cycle \( n \); on the y-axis: moment estimator \( \hat{X}_{0,n} \) with \( K \) estimated by \( X_n/n \). The true value of \( X_0 \) is 100.](image)

As expected, it appears that the moment estimator based on consecutive observations for estimating \( K \) is more precise that the one based on a single observation.
Figure 2: Simulation 1. On the x-axis: replication cycle \( n \); on the y-axis: moment estimator \( \hat{X}_{0,n} \) with \( K \) estimated by the conditional least squares estimator based on \( X_h, \ldots, X_n \) with \( h = 20 \) and \( n \geq h + 1 \). The true value of \( X_0 \) is 100.

For estimating \( K \), that is, the more information is used, the better the moment estimator is. For other simulations with the same parameter values, one obtains similar results drawn in Figures 3 and 4 for simulation 2, and in Figures 5 and 6 for simulation 3.

Jagers and Klebaner (2003) proved that \( \{X_n\} \) is asymptotically linear under A1-A3. This might explain the fact that, when estimating \( K \) by \( X_n/n \), there is a systematic over-estimation of \( X_0 \) (see Figures 1, 3, and 5): the moment estimator increases while \( n \) is small and the branching process undergoes an exponential increase, whereas the moment estimator decreases for larger \( n \) for which the process approaches its linear phase.

Figure 3: Simulation 2. On the x-axis: replication cycle \( n \); on the y-axis: moment estimator \( \hat{X}_{0,n} \) with \( K \) estimated by \( X_n/n \). The true value of \( X_0 \) is 100.
Figure 4: Simulation 2. On the x-axis: replication cycle $n$; on the y-axis: moment estimator $\hat{X}_{0,n}$ with $K$ estimated by the conditional least squares estimator based on $X_h, \ldots, X_n$ with $h = 20$ and $n \geq h + 1$. The true value of $X_0$ is 100.

Figure 5: Simulation 3. On the x-axis: replication cycle $n$; on the y-axis: moment estimator $\hat{X}_{0,n}$ with $K$ estimated by $X_n/n$. The true value of $X_0$ is 100.
Figure 6: Simulation 3. On the x-axis: replication cycle \( n \); on the y-axis: moment estimator \( \hat{X}_{0,n} \) with \( K \) estimated by the conditional least squares estimator based on \( X_h, \ldots, X_n \) with \( h = 20 \) and \( n \geq h + 1 \). The true value of \( X_0 \) is 100.

4 Application to Polymerase Chain Reactions

The PCR technique has become the method of choice in molecular biology to replicate DNA fragments (Edwards et al., 2004). It is especially used when the initial number of nucleic acid fragments is small so that the number of DNA molecules has to be amplified for further analysis and for quantification purposes.

The stochastic model described in Section 1 has been developed by Jagers and Klebaner (2003) in order to model DNA amplification by PCR. They derived the parametric form of the offspring amplification rate (2) from a Michaelis-Menten kinetics approximation proposed by Schnell and Mendoza (1997a,b). This model accounts for the fact that the amplification rate decreases as more and more DNA molecules accumulate (Raeymaekers, 2000). \( K \) is a parameter that embodies the thermodynamic environment of the reaction (Schnell and Mendoza, 1997a).

Occasionally, nucleotides are deleted, added, or substituted while synthesizing a new DNA molecule from a DNA template. Within the model of Jagers and Klebaner (2003) that we adopt, these copying errors are neglected and one assumes that the newly synthesized molecules are identical to their templates. Discarding the copying errors is a common assumption in quantitative methods for PCR.

Assuming that one can observe numbers of DNA molecules replicated at consecutive PCR replication cycles, and that these observations are not corrupted by some noise, the moment estimator \( \hat{X}_{0,n} \) may be used to estimate the initial population size of the DNA fragments amplified by PCR based on the observation of \( (X_1, \ldots, X_n) \). In practice, when considering experimental data expressed in numbers of DNA molecules, one should take into account the noise inherent to the measuring device. The main current approach in quantitative PCR consists in
relying on observations of the amplification process such that the measurement error may be neglected. In our setting, this would amount to adapting our estimator so that it relies on the observation of \((X_\nu, \ldots, X_n)\), where \(\nu\) is some random cycle such that, from this cycle on, the considered observations are significantly above the background noise level. This adapted estimation approach is a current line of research.

5 Discussion

This study investigated the problem of estimation of the initial size of a binary-splitting population-size-dependent branching process. The moment estimator presented here is easy to compute. It might serve as an initial estimator for an algorithm computing an other more efficient estimator of the initial population size (Bickle and Doksum, 2001).

In the framework of PCR, there are typically a few dozens of observations. In a more general setting, it would be of interest to investigate the asymptotic behavior of the moment estimator \(\hat{X}_{0,n}\), as \(n\) tends to infinity, which may be rewritten as

\[
\hat{X}_{0,n} = W_n \frac{m_{R_n}(X_0)}{2} \prod_{\ell=1}^{n-1} \frac{m_K(X_\ell)}{m_{R_n}(X_\ell)}.
\]

By the martingale convergence theorem (Hall and Heyde, 1980), there exists a random variable \(W\) such that \(\lim_{n \to \infty} W_n \overset{a.s.}{=} W\), with \(E(W) < \infty\). Under A1-A3, \(\lim_{n \to \infty} X_n \overset{a.s.}{=} \infty\), and, according to Pierre-Loti-Viaud (1994), \(\{W > 0\} \overset{a.s.}{=} \{X_n \to \infty\} \text{ as } n \to \infty\}. As a consequence, one would be led to study the quantity \(\lim_{n \to \infty} \prod_{\ell=1}^{n-1} \frac{m_K(X_\ell)}{m_{R_n}(X_\ell)}\). This could be done by noting that

\[
\log \prod_{\ell=1}^{n-1} \frac{m_K(X_\ell)}{m_{R_n}(X_\ell)} = \sum_{\ell=1}^{n-1} \log \left(1 + \frac{m_K(X_\ell) - m_{R_n}(X_\ell)}{m_{R_n}(X_\ell)}\right).
\]

In view of (2), one could use

\[
\log \left(1 + \frac{m_K(X_\ell) - m_{R_n}(X_\ell)}{m_{R_n}(X_\ell)}\right) = \log \left(1 + \frac{(K - \hat{K}_n)X_\ell}{(K + X_\ell)(2\hat{K}_n + X_\ell)}\right).
\]

A current line of investigation for the asymptotic behavior of (4) consists in the analysis of the previous relationship as \(n\) tends to infinity.

Acknowledgement: The author is grateful to professor Peter Jagers for helpful discussions. This research was funded by the Gothenburg Mathematical Modelling Centre.
References


