Estimation problems with the Jelinski-Moranda software reliability model

Francisco J. Samaniego    Simon P. Wilson

We assume that software has an unknown number $N$ of bugs and is being independently tested by $M$ users who work simultaneously. Each bug will be independently discovered by a tester in a random time that is specified by a distribution function $F(t \mid \theta)$. When a bug is discovered it is instantly repaired and testing continues. Thus the time to discovery of the $i$th bug is the minimum of $M(N - i + 1)$ observations from $F$, and so has distribution

$$F_i(t \mid N, \theta) = 1 - (1 - F(t \mid \theta))^{M(N - i + 1)}, \; i = 1, \ldots, N.$$ 

Clearly $F_{N+1}(t \mid N, \theta) = 0, \forall t$. The Jelinski-Moranda model Jelinski and Moranda (1972) is obtained by letting $F(t \mid \theta) = 1 - \exp(-\theta t)$ and $M = 1$.

1 The Random Sampling Scheme

It seems appropriate to begin a detailed discussion of inferential methods in software reliability with a brief review of the origins of the subject and some commentary on the current status of the field. This will serve as appropriate background for the ideas and methods on which the present paper is focused and, more importantly, will provide strong motivation for the need for a new approach. The natural starting point for such a review is the seminal paper by Jelinski and Moranda (1972). That paper restricted attention to data governed by an underlying exponential distribution, and much of the subsequent inferential work in the field has generalized their initial results to other parametric models. Virtually all of the work that related to the basic bug-counting model makes the same fundamental assumption that after any debugging action, the waiting times for further bug discoveries are independent and identically distributed variables from the assumed underlying model. The available data consists of the inter-discovery times $T_1, T_2, \ldots, T_k$ associated with the first $k$ bugs detected. Assuming an exponential distribution for $F$, and assuming that the $k$ inter-discovery times are independent, then

$$T_i \sim \text{EXP}(M(N - i + 1)\theta),$$

where $\text{EXP}(\lambda)$ stands for the exponential distribution with failure rate $\lambda$. The Jelinski-Moranda model treats the sample size "$k$" as a fixed and known number, much like one would consider the process of taking a predetermined sample of a given size in repeated, identical experiments. Virtually all existing literature on bug-counting problems has emulated Jelinski and Moranda in this regard. After describing the standard analysis under such an assumption, we will examine it more closely in the application of interest.

Following Jelinski and Moranda, let us take the maximum likelihood approach to inference concerning the parameter pair $(\theta, N)$. Under the model in 1, the likelihood function, given
the observed data \( t_1, t_2, \ldots, t_k \) is

\[
L(\theta, N \mid t_1, \ldots, t_k) = \prod_{i=1}^{k} M(N - i + 1)\theta \exp(-M(N - i + 1)t_i)
\]

\[
= \frac{M^k\theta^k N!}{(N - k)!} \exp\left(-M\theta \sum_{i=1}^{k} (N - i + 1)t_i\right).
\] (2)

If \( N \) is assumed known, with \( 1 \leq k < N + 1 \), the maximum likelihood estimate of \( \theta \) is

\[
\theta_N = k \left( M \sum_{i=1}^{k} (N - i + 1)t_i \right)^{-1}.
\] (3)

Since \( N \), in reality, is an unknown parameter that must be estimated from data, we will seek to identify the joint MLE of \((\theta, N)\). Toward that end, one is led to study the behavior of the ratio of likelihoods for consecutive values of \( N \), which is easily shown to reduce to:

\[
\frac{L(\theta_{N+1}, N + 1 \mid t_1, \ldots, t_k)}{L(\theta_N, N \mid t_1, \ldots, t_k)} = \left[ \frac{N + 1}{N + 1 - k} \right] \left[ \frac{\sum_{i=1}^{k} (N - i + 1)t_i}{\sum_{i=1}^{k} (N - i + 2)t_i} \right]^k.
\] (4)

Letting \( x = N + 1 \), we may write this ratio as

\[
g(x) = \left[ \frac{x}{x - k} \right] \left[ \frac{x - b}{x - (b - 1)} \right]^k,
\] (5)

where \( b = \sum_{i=1}^{k} it_i / \sum_{i=1}^{k} t_i \) and we note that, by definition, \( 1 \leq b \leq k \).

The derivative of \( g(x) \) is

\[
g'(x) = \frac{(x - b)^{k-1} (x - b + 1)^{k-1} [(2bk - k^2 - k)x - b(k - 1)]}{(x - k)^2(x - b + 1)^{2k}}.
\] (6)

We are interested in the behavior of \( g(x) \) for \( x > k \). From (5), it is clear that \( g(k+) = +\infty \).

Regarding the behavior of \( g(x) \) for \( x > k \), there are two cases to consider.

**Case 1** \( 1 \leq b < (k+1)/2 \). In this case, \( g'(x) < 0 \) for all \( x > k \), the function \( g(x) \) is decreasing for \( x \in (k, +\infty) \), and approaches its asymptote \( y = 1 \) from above. As a consequence, \( g(x) > 1 \) for all \( x > k \), and there is no solution to the equation \( g(x) = 1 \). This implies that the ratio of likelihoods is an increasing function of \( N \) and that the MLE of \( N \) is \( \hat{N} = +\infty \).

**Case 2** \( (k+1)/2 \leq b \leq k \). In this case, the equation \( g'(x) = 0 \) has a unique solution for \( x \in (k, +\infty) \), that solution being given by

\[
x^* = \frac{b(b-1)}{2b-k-1}.
\] (7)

The value of \( x^* \) is \( k \) when \( b = k \), and is otherwise larger than \( k \). The function \( g(x) \) is decreasing in the interval \((k, x^*)\) and then increases for \( x > x^* \) and approaches the asymptote \( y = 1 \) from below. It follows that the equation \( g(x) = 1 \) has a unique solution \( X^* \) in the interval \((k, x^*)\).
Since the variable \( t \) represents \( N + 1 \), we have that the ratio of likelihoods in 4 is increasing for \( N < \lceil X^* \rceil \) and decreasing for \( N > \lceil X^* \rceil \), so that the MLE of \( N \) is \( \hat{N} = \lceil X^* \rceil \), where \( [w] \) represents the greatest integer less than or equal to \( w \). The MLE cannot be represented here in closed form, since it is a root of a polynomial of high degree, but since there is but one such root in the interval \((k, +\infty)\), this root is easily found numerically.

The search for the joint MLE of the parameter pair \((\theta, N)\) thus comes to a somewhat unsatisfactory conclusion. The peculiarity that \( \hat{N} = \infty \) is possible is mentioned specifically in Singpurwalla and Wilson (1999). The solution sketched above is not entirely counterintuitive however, since \( \hat{N} = \infty \) only when the ratio \( b \) is small. If all the observed \( t_i \)'s were the same value, say \( t_i = t_0 \), \( \forall i \), then \( b = (k + 1)/2 \), the boundary value of case 1. A value of \( b \) smaller than this will tend to arise only when some of the later \( t_i \)'s are substantially smaller than some of the early \( t_i \)'s, an event to which the model assigns rather small probability. Indeed, under the model in 1, the later \( t_i \)'s correspond to observed minima based on sample sizes that are smaller than those in the early \( t_i \)'s. Thus, the \( k \) random observations are stochastically ordered, that is, \( T_1 <_{st} T_2 <_{st} \cdots <_{st} T_k \). When the \( T_i \)'s are very close to each other in value, it is an indication that \( N \), the number of bugs in the software, is large. Thus, when the \( T_i \)'s behave in a manner contrary to the expected ordering, a large \( N \) is indicated, and, in extreme circumstances, the likelihood cannot be maximized in \( N \). In a given application, one would be at a loss in using the MLE if the condition in case 1, that is, the inequality \( 1 \leq b < (k + 1)/2 \) is obtained. In that event, \( N \) cannot be estimated from data, and since the MLE of \( \theta \) depends of \( N \), that estimate is likewise unusable. The MLE would, on the other hand, be a viable estimator of the pair \((\theta, N)\) under the condition in Case 2, provided the various assumptions made were considered to be appropriate.

Let us now return to the assumptions underlying the Jelinski-Moranda model. While the parametric assumption might be suspect in certain applications, and the independence assumption might require some scrutiny as well, it is the “fixed \( k \)” assumption that should cause the greatest pause. Consider testing a new piece of software containing an unknown number of bugs \( N \). A fixed and predetermined choice of \( k \), the number of bugs to be sampled, automatically implies that one has assumed that \( N \geq k \). The latter assumption is, in general, both risky and impossible to justify. Further, the assumption runs counter to the experimenter’s natural inclination to sample as many bugs as possible, both because so doing improves the software, through the removal/repair of the discovered bugs, but also because the precision of one’s estimates of model parameters tends to improve as the sample size (a role played here by \( k \)) increases. Choosing \( k \) to be large in advance of the sampling experiment increases the risk that the chosen \( k \) exceeds the true value of \( N \), a situation that leads to the untenable situation of having designed a never-ending experiment. It is clear that the “fixed \( k \)” assumption of the Jelinski-Moranda model, and of all software reliability models that have followed suit in using that assumption, is seriously flawed. The confidence one might have in the implicit assumption that \( N \geq k \) depends on how much smaller one believes \( k \) is than \( N \), a matter that is basically unknown and involves considerable uncertainty.

Because of the tenuous nature of the “fixed \( k \)” assumption, we have taken a different approach in the present study. Our inference will assume that the number of sampled bugs is random rather than fixed. This departure from the Jelinski-Moranda type modeling is not a huge conceptual leap. Indeed, it is very common in reliability studies generally, and it is thus somewhat surprising that it does not appear to have been employed in the present context. We propose to study inference problems concerning the underlying model \( F \) governing bug generation and for \( N \), the unknown number of bugs, via type I censored samples. In other
words, we will assume that the software of interest is observed for a fixed predetermined amount of time $T^\ast$. The protocol for bug discovery and removal is the same as that discussed above. However, in this alternative model, the number $k$ of bugs discovered in the time interval $(0, T^\ast)$ is a nonnegative random variable. While $k$ can take on the value zero with positive probability, this probability is negligibly small in most applications of interest. In the subsection that follows, we develop inference for the exponential version of the bug-counting model under the assumption of type I censoring. We close the section with a discussion of robustness issues, motivating the need to go beyond the parametric models and methods that dominate the software reliability literature and leading us to the main developments of the paper which utilize the type I censoring framework mentioned above but adopts a non-parametric approach to the estimation problems of interest.

2 Type I Censored Sampling

As argued above, a type I censoring scheme is a more appropriate and realistic framework for sampling software bugs than the fixed sample-size approach. This alternative framework assumes that testing is conducted for a fixed time $T^\ast$, from which $k$ bugs are encountered at inter-arrival times $t_1, t_2, \ldots, t_k$ with distribution functions $F_1, F_2, \ldots, F_k$, followed by a right-censored observation $t_{k+1} = T^\ast - \sum_{i=1}^{k} t_i$ with distribution $F_{k+1}$. The time $t_{k+1}$ is the amount of time elapsed between the last observed bug and the end of the observation period.

We consider the model with the type I censoring scheme and $F(t \mid \theta) = 1 - \exp(-\theta t)$. Thus, the time between the discovery of the $i$-th and $i$th bug is the minimum of $M(N - i + 1)$ independent exponentially distributed variables, hence exponential with rate $M(N - i + 1)\theta$. This yields a likelihood function

$$L(\theta, N \mid t_1, \ldots, t_{k+1}) = \begin{cases} e^{-MN\theta T^\ast}, & \text{if } k = 0, \\ \frac{M^k \theta^k N!}{(N-k)!} \exp \left( -M\theta \sum_{i=1}^{k} (N - i + 1) t_i - M\theta(N-k) t_{k+1} \right), & k \geq 1. \end{cases}$$

(8)

While $N$ is unknown it is generally true that $MN\theta T^\ast$ can be assumed to be suitably large and as a consequence $P(T_1 > T^\ast)$ will be close to 0. We are therefore interested in maximising the likelihood

$$L(\theta, N \mid t_1, \ldots, t_{k+1}) = \frac{M^k \theta^k N!}{(N-k)!} \exp \left( -M\theta \sum_{i=1}^{k} (N - i + 1) t_i - M\theta(N-k) t_{k+1} \right), k \geq N,$$

(9)

for a fixed $k \geq 1$ over the parameters $(\theta, N) \in (0, \infty) \times \{k, k+1, \ldots\}$.

First let us assume that $N$ is known. Then (9) is clearly maximized by

$$\hat{\theta}_N = \frac{k}{M \sum_{i=1}^{k} (k - i + 1) t_i + (N-k)T^\ast}.$$

(10)

Denoting the mean of the underlying exponential distribution by $\mu = 1/\theta$, we have that the MLE for $\mu$, for $k \geq 1$, is

$$\hat{\mu}_N = \frac{M \left[ \sum_{i=1}^{k} (N - i + 1) t_i + (N-k) t_{k+1} \right]}{k} = \frac{M \left[ \sum_{i=1}^{k} (k - i + 1) t_i + (N-k) T^\ast \right]}{k},$$

(11)
from which we see that the numerator is the traditional “total time on test” (TTT) statistic. In this latter expression, one can view the process of observing exactly \( k \) failures by time \( T^* \) as consisting of \( (k+1) \) epochs, within each of which one has a specific number of opportunities to observe a failure. In the first epoch there are \( MN \) opportunities (i.e., independent waiting times distributed according to an exponential with rate \( \theta \)) to find a bug, and these conditions persist until time \( t_1 \), contributing \( MN t_1 \) to the TTT. In the second epoch there are \( M(N-1) \) opportunities, that contribute \( M(N-1) t_2 \) to the TTT. In the last epoch \( M(N-k) \) opportunities remain and the contribution to the TTT consists of this number times the remaining length of time \( t_{k+1} \) that these conditions persist. The above discussion demonstrates that, for known \( N \), the MLE of the exponential parameter \( \theta \) or its reciprocal \( \mu \) takes the standard form from the familiar exponential based estimation problem in reliability, that is TTT divided by the number of observed failures.

As in Section 1, we now consider \( N \) to be unknown and identify the joint MLE of \((\theta, N)\) by studying the behavior of the ratio of likelihoods for consecutive values of \( N \), which is easily shown to reduce to

\[
\frac{L(\theta_{N+1}, N+1 | t_1, \ldots, t_{k+1})}{L(\theta_N, N | t_1, \ldots, t_{k+1})} = \frac{N + 1}{N + 1 - k} \left[ \frac{\sum_{i=1}^{k} (k - i + 1) t_i + (N - k) T^*}{\sum_{i=1}^{k} (k - i + 1) t_i + (N + 1 - k) T^*} \right]^k. \tag{12}
\]

Letting \( x = N + 1 \), we may write this ratio as

\[
h(x) = \frac{x}{x - k} \left[ \frac{b + (x - k - 1) T^*}{b + (x - k) T^*} \right]^k, \tag{13}
\]

where

\[
b = \sum_{i=1}^{k} (k - i + 1) t_i. \tag{14}
\]

We see that the numerator of (12) has zeroes at \( N = -1 \) and \( N = k - b/T^* \), while the denominator has zeroes at \( N = k - 1 \) and \( N = k - b/T^* - 1 \). The only valid values of \( N \) are \( N \geq k \).

We now investigate this function. A set of data from this model with \( M = 1 \), \( N = 250 \) and \( \theta = 0.01 \) is simulated and used to construct the likelihood as a function of \( N \) for various \( k \). The left of Fig. 1 is (12) for a simulation of \( k = 2 \) values from the model with \( T^* = 0.502 \); \( b \) is calculated to be 0.1832. The function therefore has zeros at \( N = -1, 1.635 \) and asymptotes at \( N = 0.635, 1 \). We see that for \( N \geq 2 \) the ratio is always smaller than 1, hence the likelihood is decreasing for all valid \( N \). The MLE is therefore at \( N = k = 2 \).

The right of Fig. 1 is (12) for a simulation of \( k = 195 \) values from the model with \( T^* = 175.0 \); \( b \) is calculated to be 12032. The function therefore has zeros at \( N = -1, 126.25 \) and asymptotes at \( N = 125.25, 194 \). We see a repeat of the behaviour of the second example for \( N > k - 1 \), with the ratio being larger than 1 initially, but now it is larger than 1 for \( N \leq 228 \) and smaller than 1 for \( N \geq 229 \), thus the MLE is at \( N = 228 \).

3 Robustness

To illustrate the lack of robustness of the parametric bug-counting model, we fit the exponential model to a simulated data set of \( k = 80 \) discovery times, with \( M = 1 \), \( T^* = 95 \) (\( t_{81} = 1.645 \)
being the final right censored observation, $N = 100$ and Weibull inter-discovery distribution $F(t) = 1 - \exp(-0.01t^{1.5})$; thus $F_i(t) = 1 - \exp(-0.01(101 - i)t^{1.5})$. Eighty discovery times are generated. The data are displayed on the left of Fig. 2. The maximum likelihood estimate for the exponential model is ($\hat{N}, \hat{\theta}$) = (80, 0.026). This is unsatisfactory for two reasons. First, we have estimated $\hat{N} = k$ so no new bugs are predicted to appear. We have found that this is a common feature of maximum likelihood estimation for parametric models. Secondly the estimated $F(t) = 1 - \exp(-0.026t)$ is plotted on the right of Fig. 2 along with the true $F$, and we see that the true $F$ has been very badly estimated, even though this particular Weibull is only modestly different from an exponential.

References


Figure 2: On the left, the simulated Weibull data set and, on the right, the estimated $F$ using MLE for the exponential case (dashed line) with the true Weibull $F$ used to generate the data (solid line).