

1991

L.G. Barendregt, M.C.J. van Pul

An analytical solution to the consistency problem
for the Littlewood model in software reliability

Department of Operations Research, Statistics, and System Theory Report BS-R9129 December

CWI, nationaal instituut voor onderzoek op het gebied van wiskunde en informatica

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

An Analytical Solution to the Consistency Problem for the Littlewood Model in Software Reliability

Leo G. Barendregt, Mark C. van Pul

CWI
P.O. Box 4079, 1009 AB Amsterdam
The Netherlands

A very well-known model in software reliability theory is that of Littlewood (1980). The parameters in this model are usually estimated by means of the maximum likelihood (ML) method. The system of likelihood equations can have more than one solution. Only one of them will be consistent, however. We present and compare two different approaches to construct estimators for the model parameters of this particular model and investigate whether they are consistent or not. Our belief is that the ideas and methods developed in this paper could also be of interest for statisticians working outside the field of (software) reliability theory.

1980 Mathematics Subject Classification: 62N05, 62M99.

Keywords and Phrases: Software reliability, Littlewood model, maximum likelihood estimation (MLE), consistency, uniqueness.

1. Introduction

Maximum likelihood estimators are generally used in all fields of statistics. For those MLE's theoretical (asymptotic) properties are often derived, but seldom verified in practice. LeCam (1990) showed how terribly they can behave sometimes. Another thing to keep in mind is the fact that in many cases the likelihood equations will have more than one solution. Classical theorems state that under suitable conditions exactly one of those solutions of the likelihood equations will be consistent and that this consistent solution will be asymptotically normally distributed and efficient. Here the problem arises how to choose from a couple of candidates (solutions of the likelihood equations) the consistent one. More seriously, if one has laboriously numerically determined one solution, how can one be sure there are no others? LeCam (1991) addresses those kind of problems too. His advice is to just apply one-step of the Newton-Raphson method to an initial estimator, which is \sqrt{n} -consistent. It is a well-known result that such an one-step estimator will be asymptotically equivalent to the MLE. Practical results obtained with the one-step Newton-Raphson method, however, often turn out to be very disappointing. Moreover, it is often actually rather difficult to construct a suitable initial estimator.

In this paper we consider this problem for a particular case in the field of software reliability theory. In the next section we introduce Littlewood's parametric software reliability model in more detail. In section 3 we discuss the maximum likelihood estimation method for counting processes and show that for the software reliability model of Littlewood (1980) the likelihood equations can have more than one solution. Hence we face here the problem mentioned earlier. In section 4 a first approach is shown how an initial estimator, that is \sqrt{n} -consistent, can be found. We describe an algorithm that produces an asymptotically efficient estimator. In practice this algorithm works well (and better than the one-step Newton-Raphson method). In the fifth section we describe a second, more analytical approach to the problem, exploiting the mathematical properties of the log-likelihood function of the Littlewood model. Finally, in section 6 we compare the results of the

two methods, when applied on simulated data, generated by the Littlewood model. Our belief is that the ideas and methods developed in this paper could also be of interest for statisticians working outside the field.

2. The Littlewood model

Computer systems have become more and more important in modern society. The problem of estimating the reliability of computer software undergoing debugging has therefore, over the last two decades, received a great deal of attention. For this purpose a considerable number of models has been proposed. We refer to Musa et al. (1987) for a complete overview of the most common software reliability models. Two of the most well-known models in software reliability theory are those of Jelinski-Moranda (1972) and Littlewood (1980). For the Jelinski-Moranda model Moek (1983) gave a criterion on the data, satisfied with probability one when the model is true, under which there exists an unique solution of the maximum likelihood equations. For the model of Littlewood such a criterion, however, is not known and probably will not exist. With use of asymptotic theory it is proved (Van Pul (1990)) that in case of one or more solutions of the likelihood equations exactly one of them will be consistent, if we can make this choice using the data only. Here the earlier mentioned problem arises: how to choose a consistent MLE, when the likelihood equations have more than one solution. Also alternative estimation methods do have the same difficulties with the Littlewood model. In case of M-estimation, nonsingularity of the resulting matrix has unfortunately not been proved yet (see Geurts et al. (1988)).

We consider the following test experiment. A computer program has been executed during a specified exposure period and the interfailure times are observed. The repairing of a fault takes place immediately after it produces a failure and no new faults are introduced with probability one.

Let N be the unknown number of faults initially present in the software. Let the exposure period be $[0, \tau]$ and let $n(t)$, $t \in [0, \tau]$, denote the number of faults detected up to time t . Define $T_0 := 0$ and let T_i , $i=1, 2, \dots, n(\tau)$, the failure time of the i -th occurring failure, while $t_i := T_i - T_{i-1}$, $i=1, 2, \dots, n(\tau)$, denotes the interfailure time, that is the time between the i -th and the $(i-1)$ -th occurring failure. Finally we define $t_{n(\tau)+1} := \tau - T_{n(\tau)}$.

In the Jelinski-Moranda model, introduced in 1972 and a few years later generalized by Musa (1975), the failure rate of the program is at any time proportional to the number of remaining faults and each fault still present makes the same contribution to the failure rate. So if $(i-1)$ faults have already been detected, the failure rate for the i -th occurring failure, λ_i , becomes:

$$\lambda_i = \phi_0 \left[N_0 - (i-1) \right], \quad (2.1)$$

where ϕ_0 is the true failure rate per fault (the occurrence rate) and N_0 is the true number of faults initially present in the software. In terms of counting processes we can write:

$$\lambda^{\text{JM}}(t) = \phi_0 \left[N_0 - n(t-) \right], \quad t \in [0, \tau], \quad (2.2)$$

where $\lambda(t)$, $t \in [0, \tau]$ denotes the failure rate at time t . The interfailure times $t_i, i=1, \dots, n(\tau)$, are independent and exponentially distributed with parameter λ_i given by (2.1).

In the model, introduced by Littlewood (1980), it is again assumed that at any time the failure rate is proportional to the number of remaining errors. The main difference in the Littlewood model with respect to the Jelinski-Moranda model, is the fact that each fault does not make the same contribution to the failure rate $\lambda(t)$. Littlewood's argument for that is that larger faults will produce failures earlier than smaller ones. He treats ϕ_j , the failure rate of fault j , as a stochastic variable and suggests a Gamma distribution:

$$\phi_j \sim \Gamma(a_0, b_0), \quad j=1, \dots, N.$$

We define the expected occurrence rate of faults not occurred up to time t as

$$\phi(t) := \mathbb{E}\{\phi_j | T_j > t\},$$

with

$$\phi_j \sim \Gamma(a_0, b_0),$$

$$T_j | \phi_j = \phi \sim \exp(\phi).$$

A simple calculation yields that the T_j have exceedance probability

$$\Pr(T_j > t) = \left[\frac{b_0}{b_0 + t} \right]^{a_0},$$

that is, the T_j have a generalized Pareto distribution. Furthermore

$$\phi_j | T_j > t \sim \Gamma(a_0, b_0 + t)$$

and hence:

$$\phi(t) = \frac{a_0}{b_0 + t}.$$

An application of the so called innovation-theorem (Aalen (1978)) now shows, that the failure intensity of the software at time t is given by:

$$\lambda^L(t) = \frac{a_0 [N_0 - n(t-)]}{b_0 + t}. \quad (2.3)$$

By a simple reparametrization, namely:

$$\alpha_0 = \frac{a_0}{b_0}, \quad \varepsilon_0 = \frac{1}{b_0},$$

we get from (2.1):

$$\lambda^{GL}(t) = \alpha_0 \frac{(N_0 - n(t-))}{1 + \varepsilon_0 t}, \quad t \in [0, \tau]. \quad (2.4)$$

Actually formula (2.4) provides an extension of the Littlewood model (2.3), allowing also small values of $\varepsilon_0 \leq 0$. Restricting ourselves to the conventional parameter space

$$\Theta := \{ (N, \alpha, \varepsilon) \in \mathbf{R}^3 \mid N \geq 0, \alpha \geq 0, \varepsilon \geq 0 \},$$

which is not compact, we will investigate the model behaviour at the boundary of the parameterset.

We will see that letting certain combinations of N, α, ε converge to their boundary limits (zero or infinity) at various speeds, this may lead to different limiting models. Apart from the *null-model* (0) (where, as for instance $N=0$, nothing happens) we can roughly distinguish four non-trivial boundary models:

- (1) *The explosion model (E^δ):*
If for instance $0 < N < \infty$, $\alpha = \infty$ and $0 < \varepsilon < \infty$, the failure-intensity at time zero becomes infinite. Thus the expected number of bugs detected makes a jump from zero to δ at time zero and remains constant for $t > 0$. The failure-intensity drops to zero for $t > 0$. Special cases are E^0 (the null-model), E^N and E^∞ . E^Δ represents the class of all explosion models $\{E^\delta \mid 0 \leq \delta \leq \infty\}$.
- (2) *The Jelinski-Moranda model (JM):*
If $0 < N < \infty$, $0 < \alpha < \infty$ and $\varepsilon = 0$, we are dealing with the model we discussed earlier, namely the *Jelinski-Moranda model* (2.2) with occurrence rate parameter ϕ equal to α . We can therefore treat the Jelinski-Moranda model as a special (limit-)case of the Littlewood model.
- (3) *The inhomogeneous Poisson model (IP):*
If $N \rightarrow \infty$, $\alpha \rightarrow 0$ and $0 < \varepsilon < \infty$ such that $N\alpha \rightarrow \delta$, the influence of the past of the counting-process $n(t-)$ is eliminated from expression (2.4) and the general model reduces to an *inhomogeneous Poisson model* with intensity function $\lambda(t) = \delta/(1+\varepsilon t)$, $t \in [0, \tau]$.
- (4) *The homogeneous Poisson model (HP):*
If $N \rightarrow \infty$, $\alpha \rightarrow 0$ and $\varepsilon \rightarrow 0$ such that $N\alpha \rightarrow \delta$, all time-dependence is eliminated from the expression (2.4) and the general model reduces to a *homogeneous Poisson model* with constant failure intensity δ .

Table 1 shows which of the above mentioned limiting models occur for which (N, α, ε) .

	$N=0$			$0 < N < \infty$			$N=\infty$		
	$\varepsilon=0$	$0 < \varepsilon < \infty$	$\varepsilon=\infty$	$\varepsilon=0$	$0 < \varepsilon < \infty$	$\varepsilon=\infty$	$\varepsilon=0$	$0 < \varepsilon < \infty$	$\varepsilon=\infty$
$\alpha=0$	0	0	0	0	0	0	HP	IP	E^Δ
$0 < \alpha < \infty$	0	0	0	JM	LW	0	E^∞	E^∞	E^Δ
$\alpha=\infty$	0	0	0	E^N	E^N	0 or E^N	E^∞	E^∞	E^Δ

Table 1: Boundary cases of the Littlewood model (LW)

Note that for instance for $0 < N < \infty$, $\alpha \rightarrow \infty$ and $\varepsilon \rightarrow \infty$ the limiting model heavily depends on the way we let α and ε increase. In figure 1 we plotted expected number of faults detected versus time for various choices of parameter triples (N, α, ε) approaching the boundary of the parameter set Θ . The bold curves represent the limit models.

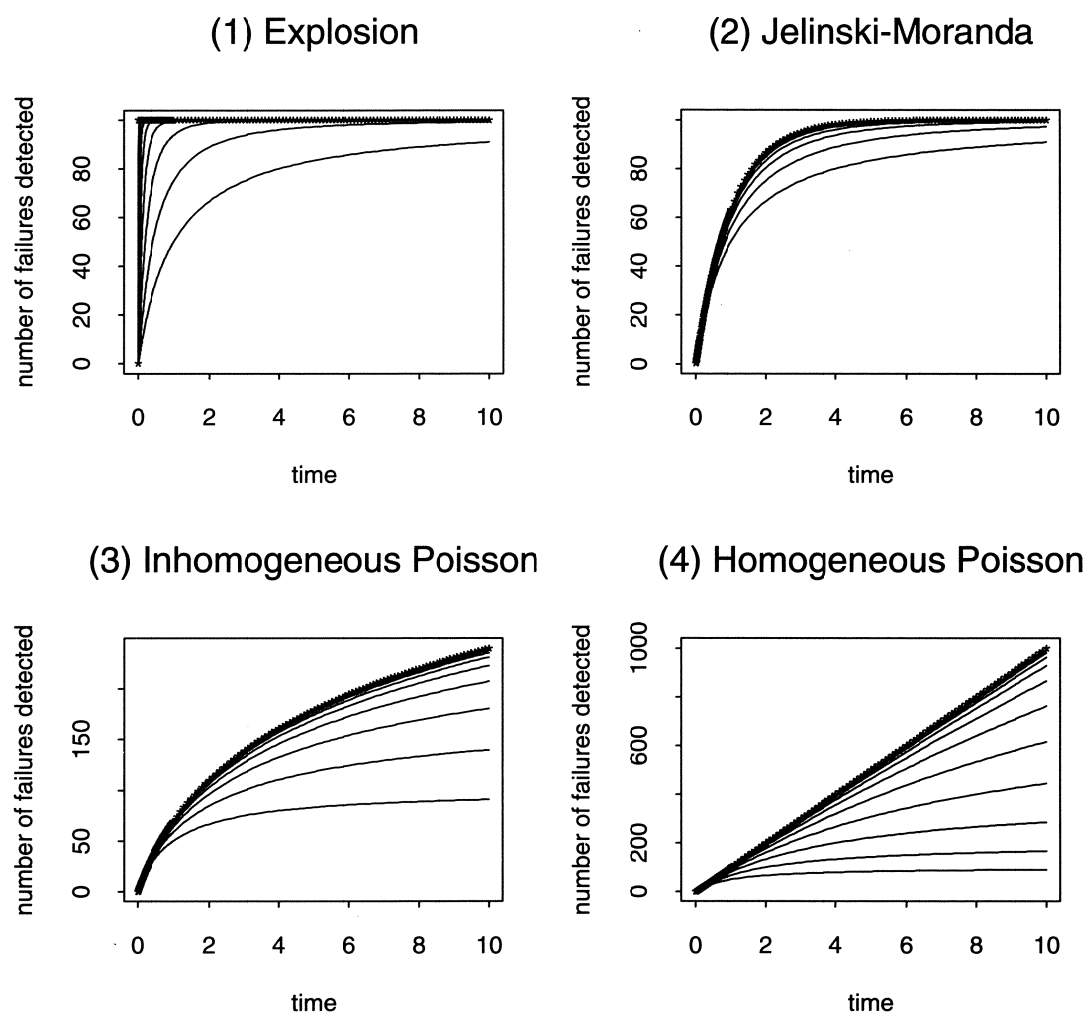


Figure 1: Boundary cases of the Littlewood model (LW)

3. Maximum likelihood estimation for counting processes

By using the information obtained from the test experiment one can estimate the the parameters of the underlying model. Mostly maximum likelihood estimation is used for this purpose. Aalen (1978) showed using theory of Jacod (1975) that the likelihood function for the vector-parameter θ of the stochastic intensity $\lambda^F(t)$ of a counting process $n(t)$ observed on $[0, t]$ and conditioned on the past F_{t-} , is given by:

$$L(\theta; t | F_{t-}) := \exp \left[\int_0^t \log \lambda^F(s; \theta) dn(s) - \int_0^t \lambda^F(s; \theta) ds \right].$$

The past here only consists of the minimal (or self-exciting) history of the counting process $n(t)$:

$$F_{t-} := \sigma \{ n(s) : 0 \leq s < t \}.$$

For the Littlewood model with intensity (2.4), parameter $\theta=(N, \alpha, \varepsilon)$ and exposure period $t=\tau$, the log-likelihood function becomes:

$$\begin{aligned} \log L(N, \alpha, \varepsilon; \tau) &:= \sum_{i=1}^{n(\tau)} \log \left[\frac{\alpha(N-i+1)}{1+\varepsilon T_i} \right] - \sum_{i=1}^{n(\tau)+1} \frac{\alpha}{\varepsilon} (N-i+1) \log \left[\frac{1+\varepsilon T_i}{1+\varepsilon T_{i-1}} \right] \\ &= n(\tau) \log(\alpha) - \alpha(N-n(\tau)) \frac{\log(1+\varepsilon\tau)}{\varepsilon} \\ &\quad + \sum_{i=1}^{n(\tau)} \log(N-i+1) - (\alpha+\varepsilon) \sum_{i=1}^{n(\tau)} \frac{\log(1+\varepsilon T_i)}{\varepsilon}, \end{aligned} \quad (3.1)$$

where $T_0 := 0$ and $T_{n(\tau)+1} := \tau$. Hence the log-likelihood equations are:

$$\frac{\partial}{\partial N} \log L(N, \alpha, \varepsilon; \tau) = \sum_{i=1}^{n(\tau)} \frac{1}{N-i+1} - \alpha \frac{\log(1+\varepsilon\tau)}{\varepsilon} = 0; \quad (3.2)$$

$$\frac{\partial}{\partial \alpha} \log L(N, \alpha, \varepsilon; \tau) = \frac{n(\tau)}{\alpha} - (N-n(\tau)) \frac{\log(1+\varepsilon\tau)}{\varepsilon} - \sum_{i=1}^{n(\tau)} \frac{\log(1+\varepsilon T_i)}{\varepsilon} = 0; \quad (3.3)$$

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} \log L(N, \alpha, \varepsilon; \tau) &= \frac{\alpha(N-n(\tau))}{\varepsilon^2} \left[\log(1+\varepsilon\tau) - \frac{\varepsilon\tau}{1+\varepsilon\tau} \right] + \frac{\alpha}{\varepsilon^2} \sum_{i=1}^{n(\tau)} \left[\log(1+\varepsilon T_i) - \frac{\varepsilon T_i}{1+\varepsilon T_i} \right] \\ &\quad - \sum_{i=1}^{n(\tau)} \frac{T_i}{1+\varepsilon T_i} = 0. \end{aligned} \quad (3.4)$$

Remark 3.1:

Natural questions to ask are whether $\log L(N, \alpha, \varepsilon; \tau)$, as defined by (3.1), might have more than one (local) maximum, and whether the system of likelihood equations (3.2)-(3.4) might have more than one solution, which is a different problem. The answer to both questions is affirmative. Consider the following dataset:

$$\begin{aligned} n(\tau) &= 3; \\ T_1 &= 1, T_2 = 399.9, T_3 = 400.1; \\ \tau &= 709.5. \end{aligned}$$

We will see later (in remark 5.2) that $\log L(N, \alpha, \varepsilon; \tau)$ has a global maximum at the boundary ($N=n(\tau)$ and $\varepsilon=0$) and both a local maximum and a saddlepoint in the interior of the parameterset Θ . The dataset constructed here is of course in bad agreement with the Littlewood-model. In section 5 we will consider this dataset in more detail. (*End of remark 3.1*)

As the system of the three highly non-linear likelihood equations (3.2)-(3.4) obviously can not be solved analytically Moek (1984) and Geurts et al. (1988) use the parametrization (2.3) and suggest to simplify the problem by fixing one of the parameters $a=1$. Furthermore, Moek (1984) was able to express both N and a as functions of b and used this to derive an equation $f(b)=0$ from which b can be solved numerically relatively easy. Criteria were formulated which ensure the existence of at least one solution. The problem of the possibility of multiple solutions is ignored. Geurts et al. (1988) signal this problem but do not provide a solution. They state that in case of multiple solutions exactly one of them will be consistent; moreover this consistent solution is asymptotically normal distributed and efficient. In sections 4 and 5 two approaches are presented to construct estimates that are indeed consistent. Method I in section 4 describes how an initial estimator can be obtained that is \sqrt{n} -consistent. Starting with this initial estimator several (iterative) methods will then provide us with estimators that are asymptotically equivalent with the MLE. Another, more analytical approach is presented by Method II in section 5. Here we reduce the problem of maximizing the log likelihood function $\log L(N, \alpha, \varepsilon; \tau)$, given in (3.1), to a one-dimensional one by eliminating first the parameter α explicitly and then the parameter N implicitly. We show that the resulting estimator has desired statistical properties.

4. Method I

4.1. Constructing an initial estimator $\tilde{\theta}$; applying an (iterative) optimization procedure on it.

We define $m(t)$ as the expected number of failures detected up to time t . From (2.4) it follows that:

$$m(t) := \mathbf{E} \{ n(t) \} = N \left[1 - \left(\frac{1}{1+\varepsilon t} \right)^{\alpha/\varepsilon} \right].$$

In theorem 1 we will prove that given an arbitrary triplet $(s_1, s_2, s_3) \in [0, \tau]^3$ solving

$$m(s_i) = n(s_i)$$

for N , α and ε , will have one solution at most. This solution $\tilde{\theta}$, which can be viewed as a generalized moment estimator, will turn out to be a good starting point for estimating θ_0 .

Lemma 4.1:

Consider the function:

$$x(t) = N \left[1 - \left(\frac{1}{1+\varepsilon t} \right)^{\alpha/\varepsilon} \right], \quad t \geq 0, \quad N > 0, \quad \alpha > 0, \quad \varepsilon > 0. \quad (4.1)$$

Given three points in time s_1, s_2, s_3 , with $0 < s_1 < s_2 < s_3$, the determinant of the Jacobian

$$D := \begin{vmatrix} \frac{\partial}{\partial N}x(s_1) & \frac{\partial}{\partial \alpha}x(s_1) & \frac{\partial}{\partial \varepsilon}x(s_1) \\ \frac{\partial}{\partial N}x(s_2) & \frac{\partial}{\partial \alpha}x(s_2) & \frac{\partial}{\partial \varepsilon}x(s_2) \\ \frac{\partial}{\partial N}x(s_3) & \frac{\partial}{\partial \alpha}x(s_3) & \frac{\partial}{\partial \varepsilon}x(s_3) \end{vmatrix} \quad (4.2)$$

is non-zero.

Proof of lemma 4.1: (Duistermaat)

Fix s_2 and s_3 . Consider $D(s_1, s_2, s_3)$ as a function of s_1 only. It is easy to verify that

$$\frac{\partial^2}{\partial s_1^2} D(s_1)$$

has only one zero, hence only 3 zeros for $D(s_1)$. Obviously, however, $D(s_1)=0$ for s_1 equal to 0, s_2 and s_3 . As it is given that $0 < s_1 < s_2 < s_3$ it follows immediately, that D is non-zero. \square

Theorem 4.1:

We again consider the function $x(t)$ defined by (4.1). Let be given three points in time s_1, s_2, s_3 , with $0 < s_1 < s_2 < s_3$ and three positive numbers x_1, x_2, x_3 . Then there is at most one parameter-triple (N, α, ε) , such that:

$$x(s_i) = x_i, \quad i=1,2,3. \quad (4.3)$$

Remark 4.1:

It is obvious that there are situations in which there exists no solution (N, α, ε) to (4.3). This is the case for instance when monotonicity

$$0 < x_1 < x_2 < x_3 \quad (4.4)$$

or convexity

$$\frac{x_1}{s_1} > \frac{x_2 - x_1}{s_2 - s_1} > \frac{x_3 - x_2}{s_3 - s_2} \quad (4.5)$$

are violated. Notice that (4.4) and (4.5) are necessary, but not sufficient conditions for the existence of a solution of (3.1.3). See also remark 4.3. (End of remark 4.1)

Remark 4.2:

Obviously the result of theorem 4.1, would follow directly from the non-zerosness of the determinant D , given in (4.2), if we could prove that the boundary of the image of

$$\phi: \mathbf{R}_+^3 \rightarrow \mathbf{R}^3: (N, \alpha, \varepsilon) \rightarrow (x(s_1), x(s_2), x(s_3))$$

is simply connected. These two facts together, namely, would imply that the mapping ϕ is 1-1 by standard covering arguments (see e.g. Greenberg (1966)). The first result, i.e. the non-zerosness of the determinant, follows from lemma 4.1 and will be used later on. The connectivity of the boundary, however, is much harder to show. Therefore we have chosen for an alternative, more direct analytical proof, using an appropriate transformation of the parameters N, α and ε . (*End of remark 4.2*)

Proof of theorem 4.1:

Let (N, α, ε) a solution of (4.3). Define

$$u_i := \frac{\log(1 + \varepsilon s_i)}{\varepsilon}, \quad i=1,2,3. \quad (4.6)$$

Then we have

$$x_i = N \left[1 - \exp(-\alpha u_i) \right], \quad i=1,2,3;$$

and hence

$$u_i = -\frac{1}{\alpha} \log\left(1 - \frac{x_i}{N}\right), \quad i=1,2,3. \quad (4.7)$$

The symmetry in expressions (4.6) and (4.7) inspires us to define for $i=1,2,3$:

$$v_i(\lambda) := \begin{cases} \frac{\log(s+t_i\lambda)}{\lambda} & \text{if } \lambda > 0; \\ s_i & \text{if } \lambda = 0; \end{cases} \quad (4.8)$$

$$w_i(\mu) := \begin{cases} \frac{-\log(1-x_i\mu)}{\mu} & \text{if } 0 < \mu < \frac{1}{x_3}; \\ x_i & \text{if } \mu = 0. \end{cases} \quad (4.9)$$

We now define $V_1 := v_2/v_1$, $V_2 := v_3/v_1$, $W_1 := w_2/w_1$, $W_2 := w_3/w_1$ and consider the two curves in \mathbf{R}^2 :

$$V(\lambda) := (V_1(\lambda), V_2(\lambda)), \quad 0 \leq \lambda < \infty; \quad (4.10)$$

$$W(\mu) := (W_1(\mu), W_2(\mu)), \quad 0 \leq \mu < \frac{1}{x_3}. \quad (4.11)$$

It is easy to check that every intersection of the curves V and W corresponds (in a 1-1 way) with a solution (N, α, ε) of (4.3). We will show now that there exist at most one such crossing, by proving:

Lemma 4.2:

For a crossing, say (λ_0, μ_0) of V and W , given by (4.8)–(4.11), we have:

$$\frac{dV_1}{dV_2}(\lambda_0) < \frac{dW_1}{dW_2}(\mu_0). \quad (4.12)$$

The proof of lemma 4.2 will be given in appendix A. From (4.12) it follows directly that the curves $V(\lambda)$ and $W(\mu)$ can have one intersection at most. This corresponds with at most one parameter-triple (N, α, ε) as a solution to (4.3). Theorem 4.1 is now completely proved. \square

Remark 4.3:

Define

$$\lambda := \frac{1+\sqrt{5}}{2} \quad (4.13)$$

Set $(s_1, s_2, s_3) = (1, 2, 3)$ and $(x_1, x_2, x_3) = (2, 3, y)$, where

$$3.672276 \sim \frac{2\log(1+3\lambda)}{\log(1+\lambda)} < y < 4. \quad (4.14)$$

Then one can easily verify the monotonicity and convexity conditions (4.4)–(4.5). Despite of this, the curves $V(\lambda)$ and $W(\mu)$, defined by (4.8)–(4.11), have no intersection. This example counters the idea that (4.4)–(4.5) would also be sufficient for the existence of such a crossing. We need an extra condition, perhaps embodying the fact that $x'''(t) > 0$ for all t . (*End of remark 4.2*)

It is obvious that maximizing $\log L(N, \alpha, \varepsilon; \tau)$, as defined in (3.1), can be very hard, but is not impossible. In Van Pul (1990) we have used a standard optimization program written in Pascal, called *Amoeba* and described in Vetterling et al. (1985), which carries out a fast down-hill simplex method. This method does not treat the problem as a series of one-dimensional maximizations and only function evaluations are involved, not derivatives. As input it requires the four edges of a non-degenerate simplex S in \mathbf{R}^3 , a (function) tolerance ε and the function f , which is to be maximized. An appropriate sequence of reflections, expansions and contractions of the initial simplex S *should always converge* to the maximum of the function f , not necessarily lying in the initial simplex S . Guarantees, however, that Amoeba finds the absolute maximum in the presence of local maxima, cannot be given. The existence of more than one (local) maximum, however, is very well possible, although so far multiple solutions have never been found in practice, as we know (see also Moek (1983)). The dataset constructed in remark 3.1 shows, however, that it is theoretically possible. Furthermore it is possible that Amoeba crashes, because the maximum of the function was taken at the boundary of the simplex. In such a case we had to restart the Amoeba procedure with a different initial simplex S' . In the sequel of this paper, we will need an optimization procedure with nice mathematical properties. Therefore we introduce the following assumption:

Assumption 4.1:

There exists an optimization program, say *Max*, which, when given a continuous, concave function $f: \mathbf{R}^3 \rightarrow \mathbf{R}$ and a compact subset S of \mathbf{R}^3 , *will always return* an $x_{\max} \in S$ with the property that for all $x \in S$:

$$f(x) \leq f(x_{\max}).$$

The algorithm to find a consistent MLE consists of the following four steps:

(step1) Choose three time points $0 < s_1 < s_2 < s_3 < \tau$, independent of the the past of the counting process. For instance: $s_i := i\tau/3$. Suppose the number of faults detected up to those time points are n_1, n_2, n_3 . We assume all conditions necessary for the existence of a crossing of $V(\lambda)$ and $W(\mu)$ are satisfied.

(step2) One can determine numerically estimators $\tilde{\lambda}$ for λ_0 and $\tilde{\mu}$ for μ_0 , such that $V(\tilde{\lambda}) = W(\tilde{\mu})$.

(step3) Compute \tilde{N} , $\tilde{\alpha}$ and $\tilde{\epsilon}$ by:

$$\begin{aligned}\tilde{N} &:= \frac{1}{\tilde{\mu}}, \\ \tilde{\alpha} &:= -\tilde{\lambda} \frac{\log(1-x_i\tilde{\mu})}{\log(1+s_i\tilde{\lambda})}\end{aligned}\tag{4.14}$$

and

$$\tilde{\epsilon} := \tilde{\lambda}.$$

Notice that $\tilde{\alpha}$ in (4.14) is independent of i . Define the initial estimator $\tilde{\theta} := (\tilde{N}, \tilde{\alpha}, \tilde{\epsilon})$.

(step4) Construct a regular tetrahedron around $\tilde{\theta}$ with vertex length equal to $Cn(\tau)^{-1/3}$ and apply the optimization program *Max* (hypothesized in Assumption 4.1) to the (locally concave!) log likelihood function (3.1) on this tetrahedron. With a probability tending to one, the consistent maximum likelihood estimator $\hat{\theta}$ will be in this tetrahedron too and if so *Max* will find it.

Remark 4.4:

If such a crossing as mentioned in step 1 does not exist, we could repeat this step with new s_i , but the developed theory could then not be applied; for the theory it is namely essential that the choice of the s_i is independent of the data. Note that if it is very hard to find a triple (s_1, s_2, s_3) , satisfying the necessary conditions, then one might suspect that the data reject the Littlewood model decisively. It is then, of course, of no use searching for accurate estimators. (*End of remark 4.4*)

4.2. Statistical properties of method I

In order to discuss the statistical properties of the estimator, we have to give some more background. Important will be the way in which we will treat asymptotics. It does not make sense to let τ , the stopping time, grow to infinity. In the long run the estimate of the total number of faults will trivially be equal to the true number of faults. It makes more sense to (conceptually) increase the number of faults in the program. The idea is that then asymptotics should be relevant to the practical situation in which N_0 is large and $n(\tau)/N_0$ not close to zero or one.

Let a counting process $n(t)$ be given. Only during a specific time interval $[0, \tau]$, are jumps of the counting process $n(t)$ observed. In this paper we will assume that the intensity function associated with the counting process exists and is a member of some specified parametric family, that is:

$$\lambda(t) := \lambda(t; N, \psi), \quad t \in [0, \tau], \quad N \in \mathbf{N}, \quad \psi \in \Psi, \quad \Psi \subset \mathbf{R}^{p-1}$$

for an integer p . Let N_0 and ψ_0 be the true parameter values. Typically the parameter N_0 represents the scale or size of the problem (sometimes $N_0 = n(\infty)$), while ψ_0 is a nuisance vector parameter. We will be

interested in estimation of N_0 and ψ_0 as $N_0 \rightarrow \infty$. We assume that the model is also meaningful for non-integer N . For instance the intensity function (2.2) of the Jelinski-Moranda model can be generalized to

$$\lambda^{\text{JM}}(t) = \phi \left[N - n(t-) \right] I_{n(t) < N}, \quad t \in [0, \tau],$$

where I_{\cdot} denotes the indicator function. As we are particularly interested in the parameter estimation when N_0 is large, we will introduce a series of counting processes $n_\nu(t)$, $t \in [0, \tau]$, $\nu=1, 2, \dots$ and let N conceptually increase. Let $N = N_\nu \rightarrow \infty$ for $\nu \rightarrow \infty$. By the reparametrization

$$N_\nu = \nu \gamma_\nu$$

with a dummy variable γ_ν , we can denote the associated intensity functions by

$$\lambda_\nu(t; \gamma, \psi) := \lambda(t; \nu \gamma, \psi), \quad t \in [0, \tau], \quad \gamma \in \mathbf{R}^+, \quad \psi \in \Psi, \quad \nu=1, 2, \dots$$

Now we consider the estimation of γ and ψ as $\nu \rightarrow \infty$. If the real-life situation has $\nu=N_0$, then $\gamma=\gamma_0=1$ and $\psi=\psi_0$. It is rather unorthodox to increase a model parameter itself, in this case N . This complication is solved by estimating γ . We will assume that the maximum likelihood estimators $(\hat{\gamma}_\nu, \hat{\psi}_\nu)$ for (γ_0, ψ_0) exist. Typically, $(\hat{\gamma}_\nu, \hat{\psi}_\nu)$ is a root of the likelihood equations

$$\frac{\partial}{\partial(\gamma, \psi)} \log L_\nu(\gamma, \psi; \tau) = 0, \quad \nu=1, 2, \dots, \quad (4.15)$$

where the likelihood function at time t $L_\nu(\gamma, \psi; t)$ is given by (see Aalen (1978)):

$$L_\nu(\gamma, \psi; t) := \exp \left[\int_0^t \log \lambda_\nu(s; \gamma, \psi) dn_\nu(s) - \int_0^t \lambda_\nu(s; \gamma, \psi) ds \right]. \quad (4.16)$$

We define for $\nu=1, 2, \dots$ the stochastic process $x_\nu(t)$ by:

$$x_\nu(t) := \nu^{-1} n_\nu(t), \quad t \in [0, \tau].$$

In most practical situations, this sequence of stochastic processes converges uniformly on $[0, \tau]$ in probability to a deterministic function $x_0(t)$ as $\nu \rightarrow \infty$ (Kurtz (1983)). We assume that the counting processes n_ν are generated by associated intensity functions $\lambda_\nu(t)$, satisfying:

$$\lambda_\nu(t) = \nu \beta(t; \theta; x_\nu(t-)),$$

for an arbitrary non-negative and non-anticipating function $\beta: [0, \tau] \times \Theta \times K \rightarrow \mathbf{R}^+$ where the model parameter $\theta=(\gamma, \psi)$ consists of the parameter of most interest γ and a nuisance parameter vector ψ . Furthermore, we define for $\theta \in \Theta \subseteq \mathbf{R}^p$, $t \in [0, \tau]$, $i, j \in 1, 2, \dots, p$ and $\nu=1, 2, \dots$ the log-likelihood:

$$C_\nu(\theta, t) := \log L_\nu(\theta, t);$$

the score function:

$$U_{vi}(\theta, t) := \frac{\partial}{\partial \theta_i} C_v(\theta, t) \quad (4.17)$$

and the minus information matrix:

$$I_{vij}(\theta, t) := \frac{\partial^2}{\partial \theta_i \partial \theta_j} C_v(\theta, t). \quad (4.18)$$

Under classical smoothness and boundedness conditions on the function β (see for instance Borgan (1984) or Van Pul (1990)), we have the following result:

Theorem 4.2:

- (i) *Consistency of ML-estimators: With a probability tending to 1, the likelihood equations (4.15) have exactly one consistent solution $\hat{\theta}_v$. Moreover this solution provides a local maximum of the likelihood function (4.16).*
- (ii) *Asymptotic normality of the ML-estimators: Let $\hat{\theta}_v$ be the consistent solution of the maximum likelihood equations (4.15), then*

$$\sqrt{v}(\hat{\theta}_v - \theta_0) \xrightarrow{D} N(0, \Sigma^{-1}), \quad v \rightarrow \infty,$$

where the matrix $\Sigma = \{\sigma_{ij}(\theta_0)\}$ with for $i, j \in \{1, 2, \dots, p\}$, $\theta \in \Theta_0$: is given by

$$\sigma_{ij}(\theta) = \int_0^\tau \frac{\frac{\partial}{\partial \theta_i} \beta(s, \theta, x_0) \frac{\partial}{\partial \theta_j} \beta(s, \theta, x_0)}{\beta(s, \theta, x_0)} ds \quad (4.19)$$

and can be estimated consistently from the observed information matrix I_v , given in (4.18).

- (iii) *Local asymptotic normality of the model: With $U_v, v=1, 2, \dots$ given in (4.17), we have for all $h \in \mathbf{R}^p$:*

$$\log \frac{dP_{\theta_v}}{dP_{\theta_0}} - h^T U_v + \frac{1}{2} h^T \Sigma h \xrightarrow{P_{\theta_0}} 0, \quad v \rightarrow \infty,$$

where $\theta_v = \theta_0 + v^{-\frac{1}{2}} h$ and $U_v \xrightarrow{D} N(0, \Sigma)$, Σ given by (4.19).

- (iv) *Asymptotic efficiency of the ML-estimators: $\hat{\theta}_v$ is asymptotically efficient in the sense that the limit distribution for any other regular estimator $\tilde{\theta}_v$ for θ_0 satisfies:*

$$\sqrt{v}(\tilde{\theta}_v - \theta_0) \xrightarrow{D(\theta_0)} Z + Y,$$

where $Z \sim_d N(0, \Sigma^{-1})$, Z and Y independent.

This theorem and its proof can be found in Van Pul (1990). We are now able to formulate the following results:

Theorem 4.3:

$\tilde{\theta}$ is \sqrt{n} -consistent.

Proof of theorem 4.3:

We first prove consistency of $\tilde{\theta}$. From Lemma 4.1 and Theorem 4.1 it follows directly, that we can write

$$\tilde{\theta} = g(x),$$

for some function $g: \mathbf{R}^3 \rightarrow \Theta$, with continuous partial derivatives. Let $\tilde{\theta}_0 := g(x_0)$ and $\tilde{\theta}_v := g(x_v)$, where we use the asymptotics developed in the beginning of this section. Then by the theorem 8.1 of Kurtz (1981) we have:

$$x_v \rightarrow x$$

in probability uniformly in t on $[0, \tau]$. Hence by the continuity of g we deduce the consistency:

$$\tilde{\theta}_v - \tilde{\theta}_0 = g(x_v) - g(x_0) \rightarrow_p 0,$$

as $v \rightarrow \infty$. The \sqrt{n} -consistency now follows immediately from theorem 8.2 of Kurtz (1981) and an application of the delta method. This proves theorem 4.3. \square

Corollary 4.1:

The algorithm described in section 4.1 works, that is, it yields an efficient estimator for θ_0 .

Corollary 4.2:

The maximum likelihood estimator $\hat{\theta}$ with the smallest Euclidean distance to $\tilde{\theta}$ is consistent.

Corollary 4.3:

The k -step Newton-Raphson modification of $\tilde{\theta}$ is asymptotically equivalent with the MLE $\hat{\theta}$.

As we do not know of the existence of an optimization algorithm, like *Max*, satisfying assumption 4.1, to apply corollary 4.1 we make use of the program *Amoeba* instead. In practice, this yields nice results (see Van Pul (1990)). Corollary 4.2 is a pure theoretical result. In general there are no numerical recipes available that deliver all(!) solutions of a system of non-linear equations. Corollary 4.3 is frequently used in practical situations, but results often turn out to be very disappointing, especially for $k=1$. For $k=4$ or $k=5$ this method should however work (see Anderson et al. (1992)).

5. Method II**5.1. Exploiting the mathematical properties of the log-likelihood function**

It will turn out to be of great advantage to apply the parameter-transformation:

$$M := N - n(\tau),$$

The log-likelihood, as expressed in the new parameters, is:

$$\log L(\alpha, M, \varepsilon) = n(\tau) \log \alpha - (\varepsilon + \alpha) \sum_{i=1}^{n(\tau)} \frac{\log(1 + \varepsilon T_i)}{\varepsilon} - \alpha M \frac{\log(1 + \varepsilon \tau)}{\varepsilon} + \sum_{i=1}^{n(\tau)} \log(M + i) \quad (5.1)$$

with parameter-domain $\{(\alpha, M, \epsilon): \alpha > 0, M \geq 0, \epsilon \geq 0\}$. When M and ϵ are kept fixed, it is quite easy to maximize $\log L(\alpha, M, \epsilon)$; it can be done analytically. If

$$R_2(M, \epsilon) := \sup_{\alpha > 0} \log L(\alpha, M, \epsilon)$$

then

$$\sup_{\alpha > 0, M \geq 0, \epsilon \geq 0} \log L(\alpha, M, \epsilon) = \sup_{M \geq 0, \epsilon \geq 0} R_2(M, \epsilon).$$

Now

$$\frac{\partial \log L}{\partial \alpha} = \frac{n(\tau)}{\alpha} - \sum_{i=1}^{n(\tau)} \frac{\log(1+\epsilon T_i)}{\epsilon} - M \frac{\log(1+\epsilon\tau)}{\epsilon}$$

and it is easy to check that $\log L(\alpha, M, \epsilon)$ for M, ϵ fixed, is maximal for

$$\alpha = n(\tau) \left[\sum_{i=1}^{n(\tau)} \frac{\log(1+\epsilon T_i)}{\epsilon} + M \frac{\log(1+\epsilon\tau)}{\epsilon} \right]^{-1}; \quad (5.2)$$

so

$$\begin{aligned} R_2(M, \epsilon) &= n(\tau) \log n(\tau) - n(\tau) \log \left[\sum_{i=1}^{n(\tau)} \frac{\log(1+\epsilon T_i)}{\epsilon} + M \frac{\log(1+\epsilon\tau)}{\epsilon} \right] \\ &\quad - \sum_{i=1}^{n(\tau)} \log(1+\epsilon T_i) + \sum_{i=1}^{n(\tau)} \log(M+i). \end{aligned} \quad (5.3)$$

In order to maximize $R_2(M, \epsilon)$ we can first maximize with respect to M , keeping ϵ fixed, and then maximize with respect to ϵ . The first maximization is relatively easy because we can make use of the following theorem:

Theorem 5.1:

For n integer and ≥ 2 , and η real and > 0 , the function

$$h(M) := -n \log(M+\eta) + \sum_{i=1}^n \log(M+i)$$

has precisely one local extremum, which is realized at:

- $M=0$, for $\eta \leq n \left[\sum_{i=1}^n \frac{1}{i} \right]^{-1}$,
- $M=\infty$, for $\eta \geq \frac{n+1}{2}$,
- $M \in (0, \infty)$, for intermediate values of η .

The proof of theorem 5.1 can be found in Appendix B. Now $R_2(M, \epsilon)$ can be rewritten as:

$$\begin{aligned} R_2(M, \epsilon) &= n(\tau) \log n(\tau) - n(\tau) \log \left[\frac{\log(1+\epsilon\tau)}{\epsilon} \right] - n(\tau) \log(M+\eta(\epsilon)) \\ &\quad + \sum_{i=1}^{n(\tau)} \log(M+i) - \sum_{i=1}^{n(\tau)} \log(1+\epsilon T_i), \end{aligned} \quad (5.4)$$

where

$$\eta(\epsilon) := \sum_{i=1}^{n(\tau)} \frac{\log(1+\epsilon T_i)}{\log(1+\epsilon\tau)}.$$

Only the third and the fourth of the terms in the right-hand side of (5.4) depend on M . Their sum has a form

similar to the function $h(M)$ of theorem 3, and so $R_2(M, \varepsilon)$ has for ε fixed, exactly one local supremum, which occurs for $M=0$ if $\eta(\varepsilon) \leq n[\Sigma 1/i]^{-1}$, for $M=\infty$ if $\eta(\varepsilon) \geq (n+1)/2$ and for a finite, positive value of M otherwise.

Remark 5.1:

$\tilde{M}(\varepsilon)$, the value of M which maximizes $R_2(M, \varepsilon)$, can not be expressed explicitly, contrary to the α of (5.2). If ε increases, $\eta(\varepsilon)$ increases as well, and so will $\tilde{M}(\varepsilon)$. (*End of remark 5.1*)

Finally then, if we define:

$$R_1(\varepsilon) := R_2(\tilde{M}(\varepsilon), \varepsilon),$$

we have to maximize $R_1(\varepsilon)$. While doing this, one must bear in mind that $R_1(\varepsilon)$ might reach its maximum for $\varepsilon=0$. As a matter of fact, we have encountered this feature several times when analyzing real data-sets.

Remark 5.2:

We now return to the dataset mentioned in remark 3.1:

$$\begin{aligned} n(\tau) &= 3; \\ T_1 &= 1, T_2 = 399.9, T_3 = 400.1; \\ \tau &= 709.5. \end{aligned}$$

The function $R_1(\varepsilon)$ takes its global maximum for $\varepsilon=0$, a second local maximum for $\varepsilon=1.152$ and a local minimum for $\varepsilon=0.023$. This corresponds to a global maximum for $R_2(M, \varepsilon)$ for $\varepsilon=0$, $M=0$, a second local maximum for $R_2(M, \varepsilon)$ for $\varepsilon=1.152$, $M=9.8$, and a saddlepoint at $\varepsilon=0.023$, $M=0.0015$. (*End of remark 5.2*)

All results are easily obtained by a standard optimization procedure for one-dimensional functions, called *Golden section search* (see for instance Vetterling et al. (1985)).

5.2. Statistical properties of method II

Although we were not able to prove the following conjecture, practical results obtained made us strongly believe that it does hold:

Conjecture 5.1:

Method II yields consistent estimators.

When trying to prove Conjecture 5.1 we encountered a third alternative way of constructing an useful estimator for θ . The idea is to regard the failure times T_i as the outcomes of $n(\tau)$ independent variables from a *Pareto*(a, b) distribution that is truncated at τ . We condition on N and then maximize the partial likelihood with respect to a and b . It can be shown that the the total likelihood now is maximized by substituting a simple function of the data and the maximizers \hat{a} and \hat{b} for N . This approach and a proof of its consistency, hopefully together with a proof of conjecture 5.1, will be discussed in a next paper appearing in *Statistica Neerlandica* (1992).

6. Comparison of the results of method I and method II

In this section we will discuss the results of some simulation experiments. We generated failure times according to the Littlewood model (2.4) with $\alpha_0=1$, $\varepsilon_0=1$ and different values for N_0 (100, 1000 and 10000). The exposure period was kept fixed: $\tau=4$. We compare the initial estimator $\tilde{\theta}$ (suggested in step 3 on page 11), a 5-step Newton-Raphson modification θ^{nr} applied on $\tilde{\theta}$ and the maximizer of the likelihood $\hat{\theta}$. Note that the initial estimator $\tilde{\theta}$ does not exist for all datasets (see remarks 4.1 and 4.3). We did repeat the simulation experiments and parameter estimations for each value of N_0 . For each value of N_0 we repeated the simulation experiments and parameter estimations $K_0=1000$ times. In table 2, K denotes the actual number of cases in which the curves $V(\lambda)$ and $W(\mu)$ intersected each other, leading to an initial estimator $\tilde{\theta}$.

N_0	K
100	521
1000	940
10000	1000

Table 2: Number of successful trials to construct $\tilde{\theta}$

For each value N_0 we picked randomly 5 datasets from those leading to an initial estimator $\tilde{\theta}$, and computed the associated θ^{nr} and $\hat{\theta}$. These three estimators, together with the corresponding values of the loglikelihood function are given in table 3. From the figures in table 3 we may draw the following conclusions:

- (1) All estimators become better as N_0 gets larger. This is not strange as we have more statistical information.
- (2) α can be estimated relatively well already for small N_0 ; this in contrary to ε which has for $N_0=10000$ still a large range.
- (3) The value of the likelihood function in θ^{nr} is larger than in $\tilde{\theta}$, and in $\hat{\theta}$ again larger than in θ^{nr} , as might be expected. The differences are in some cases very small, indicating the extreme flatness of the likelihood in the neighbourhood of its maximum.
- (4) For almost all datasets the value of the log likelihood function at θ^{nr} is already higher than $\log L_0$, the value in the true parameter θ_0 .
- (5) Different triples can lead to almost same curve (up to τ) and likelihood. The Littlewood class is very broad in the sense that different parameter triples can give a very good fit to the same data. Predictions of events after τ will differ, of course.
- (6) In two cases (dataset 2 and 3) the likelihood function takes its absolute maximum on the boundary $\varepsilon=0$. We checked that in both cases there are no other (local) maxima. The data here falsely suggests the Jelinski-Moranda model.

In figure 2 we give plots of the observed and expected number of faults detected versus time, according to the Littlewood model with parameter θ_0 , $\tilde{\theta}$, θ^{nr} and $\hat{\theta}$. In figure 2a we used dataset 5, in figure 2b dataset 9 and in figure 2c dataset 13. From figure 2b we see once more very clearly, that the estimators are strongly influenced by stochastic deviations of the observed counting process from the expected $m(t, \theta_0) = \mathbf{E}n(t)$.

(a)	1	2	3	4	5
$\log L_0$	183.3	212.8	191.0	196.8	198.7
\tilde{N}	83	87	84	114	112
$\tilde{\alpha}$	1.156	0.991	1.083	0.950	0.782
$\tilde{\varepsilon}$	0.414	0.056	0.243	1.440	0.860
$\log \tilde{L}$	184.5	215.5	192.6	196.6	198.0
N^{nr}	85	83	84	109	101
α^{nr}	1.047	1.078	1.037	1.047	1.011
ε^{nr}	0.404	0.056	0.239	1.454	0.932
$\log L^{nr}$	184.6	216.0	192.7	196.7	198.8
\hat{N}	79	85	80	88	98
$\hat{\alpha}$	0.994	1.094	1.022	1.067	1.081
$\hat{\varepsilon}$	0.042	0.000	0.000	0.408	0.922
$\log \hat{L}$	185.2	216.2	193.2	197.1	198.9

(b)	6	7	8	9	10
$\log L_0$	3746.8	3862.8	3604.1	3801.1	3563.4
\tilde{N}	1505	1246	828	942	973
$\tilde{\alpha}$	0.804	0.958	0.900	1.008	0.885
$\tilde{\varepsilon}$	2.541	1.972	0.183	0.697	0.818
$\log \tilde{L}$	3744.0	3862.8	3603.0	3801.4	3566.6
N^{nr}	1443	1179	819	937	971
α^{nr}	0.750	0.998	0.955	1.031	0.903
ε^{nr}	2.046	1.779	0.189	0.708	0.845
$\log L^{nr}$	3745.8	3863.2	3603.9	3801.5	3566.7
\hat{N}	1044	1041	908	948	1029
$\hat{\alpha}$	0.902	1.050	1.020	1.039	0.895
$\hat{\varepsilon}$	0.986	1.223	0.745	0.782	1.089
$\log \hat{L}$	3747.6	3863.9	3606.9	3801.5	3566.9

(c)	11	12	13	14	15
$\log L_0$	56728.7	55770.9	56476.9	55908.5	55976.8
\tilde{N}	9693	9580	10114	9709	9607
$\tilde{\alpha}$	1.004	0.975	0.984	0.947	0.947
$\tilde{\varepsilon}$	0.812	0.774	0.996	0.762	0.726
$\log \tilde{L}$	56729.0	55769.9	56477.2	55913.7	55966.5
N^{nr}	9692	9502	10062	9743	9452
α^{nr}	1.014	1.003	1.002	0.940	1.010
ε^{nr}	0.828	0.787	1.006	0.766	0.757
$\log L^{nr}$	56729.3	55770.9	56477.5	55913.8	55971.7
\hat{N}	10028	9789	10225	9819	10321
$\hat{\alpha}$	1.012	1.002	0.999	0.940	1.002
$\hat{\varepsilon}$	0.993	0.926	1.082	0.800	1.176
$\log \hat{L}$	56730.4	55771.7	56477.7	55913.8	55977.9

Table 3: Numerical results for $\tilde{\theta}$, θ^{nr} and $\hat{\theta}$ with (a) $\theta_0=(100,1,1)$, (b) $\theta_0=(1000,1,1)$, (c) $\theta_0=(10000,1,1)$.

Figure 2a: N=100

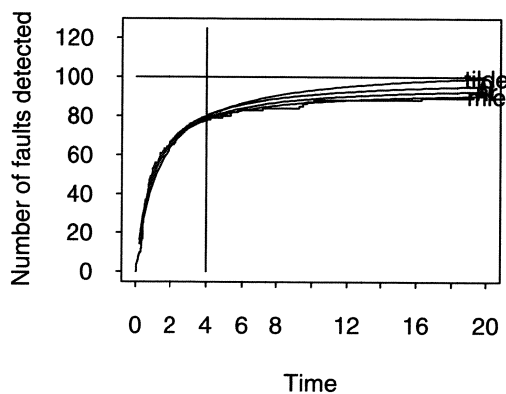


Figure 2b: N=1000

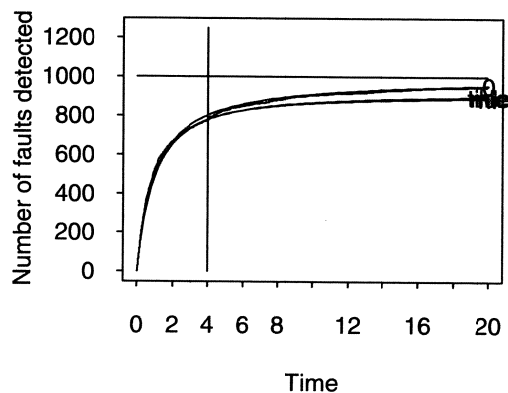


Figure 2c: N=10000

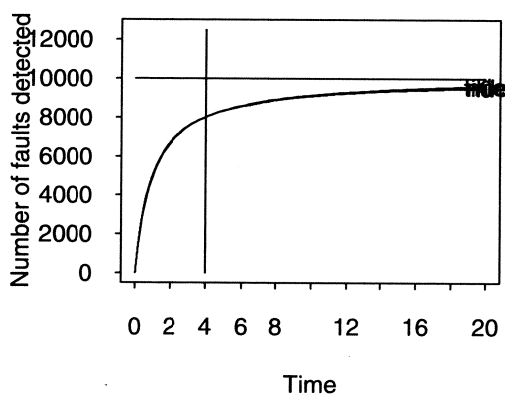


Figure 2: Observed and expected counting process according to Littlewood's model with parameters θ_0 , θ , θ^{rr} and $\hat{\theta}$.

Acknowledgement

This research was carried out under a grant of the Netherlands Foundation for Applied Technology (STW). The authors are indebted to Prof. Duistermaat for some helpful comments concerning the analytical results of Lemma 4.1 and Theorem 4.1.

Appendix A: Proof of lemma 4.2.

For the proof of lemma 4.2 we will need two other lemma's.

Lemma A.1:

The function

$$f(u) := \frac{(1+u)\log(1+u)}{u}, \quad u > -1$$

is strict monotone increasing.

Proof of lemma A.1:

Calculating the first derivative of f we find:

$$f'(u) := \begin{cases} \frac{u - \log(1+u)}{u^2} & \text{if } u > -1, u \neq 0; \\ \frac{1}{2} & \text{if } u = 0. \end{cases}$$

As $u > \log(1+u)$ both for $-1 < u < 0$ and for $u > 0$, we find that $f'(u) > 0$ for all $u > -1$. This proves lemma 2. \square

Lemma A.2:

If $\lambda_1 < \lambda_2 < \lambda_3$ and $v_1 < v_2 < v_3$ are six real numbers, then

$$\begin{vmatrix} e^{\lambda_1 v_1} & e^{\lambda_2 v_1} & e^{\lambda_3 v_1} \\ e^{\lambda_1 v_2} & e^{\lambda_2 v_2} & e^{\lambda_3 v_2} \\ e^{\lambda_1 v_3} & e^{\lambda_2 v_3} & e^{\lambda_3 v_3} \end{vmatrix} > 0.$$

Proof of lemma A.2:

Can be found in any textbook on matrix theory, see e.g. Gantmacher (1954). \square

We now return to the proof of lemma 4.2. Differentiating (4.10)-(4.11) yields:

$$\frac{dV_2}{dV_1}(\lambda) := \frac{\frac{s_3}{1+s_3\lambda}\log(1+s_1\lambda) - \frac{s_1}{1+s_1\lambda}\log(1+s_3\lambda)}{\frac{s_2}{1+s_2\lambda}\log(1+s_1\lambda) - \frac{s_1}{1+s_1\lambda}\log(1+s_2\lambda)}; \quad (\text{A.1})$$

$$\frac{dW_2}{dW_1}(\mu) := \frac{\frac{-x_3}{1-x_3\mu}\log(1-x_1\mu) - \frac{-x_1}{1-x_1\mu}\log(1-x_3\mu)}{\frac{-x_2}{1-x_2\mu}\log(1-x_1\mu) - \frac{-x_1}{1-x_1\mu}\log(1-x_2\mu)}. \quad (\text{A.2})$$

Now with

$$u_i := \frac{\log(1+s_i\lambda_0)}{\lambda_0}, \quad i=1,2,3.$$

we have:

$$\frac{\log(1-x_i\mu_0)}{\log(1-x_1\mu_0)} = \frac{\log(1+s_i\lambda_0)}{\log(1+s_1\lambda_0)} = \frac{u_i}{u_1}, \quad i=2,3.$$

The derivatives (A.1)-(A.2) hence can be written as:

$$\frac{dV_2}{dV_1}(\lambda_0) := \frac{\frac{s_3}{1+s_3\lambda_0}u_1 - \frac{s_1}{1+s_1\lambda_0}u_3}{\frac{s_2}{1+s_2\lambda_0}u_1 - \frac{s_1}{1+s_1\lambda_0}u_2}; \quad (\text{A.3})$$

$$\frac{dW_2}{dW_1}(\mu_0) := \frac{\frac{x_3}{1-x_3\mu_0}u_1 - \frac{x_1}{1-x_1\mu_0}u_3}{\frac{x_2}{1-x_2\mu_0}u_1 - \frac{x_1}{1-x_1\mu_0}u_2}. \quad (\text{A.4})$$

From lemma A.1 it follows directly that the denominators of (A.3) and (A.4) are respectively negative and positive. So inequality (4.12) is equivalent to

$$\begin{aligned} & \left[\frac{s_3}{1+s_3\lambda_0}u_1 - \frac{s_1}{1+s_1\lambda_0}u_3 \right] \left[\frac{x_2}{1-x_2\mu_0}u_1 - \frac{x_1}{1-x_1\mu_0}u_2 \right] \\ & > \left[\frac{s_2}{1+s_2\lambda_0}u_1 - \frac{s_1}{1+s_1\lambda_0}u_2 \right] \left[\frac{x_3}{1-x_3\mu_0}u_1 - \frac{x_1}{1-x_1\mu_0}u_3 \right] \end{aligned} \quad (\text{A.5})$$

and (A.5) is again equivalent to:

$$\begin{vmatrix} \frac{s_1}{1+s_1\lambda_0} & u_1 & \frac{x_1}{1-x_1\mu_0} \\ \frac{s_2}{1+s_2\lambda_0} & u_2 & \frac{x_2}{1-x_2\mu_0} \\ \frac{s_3}{1+s_3\lambda_0} & u_3 & \frac{x_3}{1-x_3\mu_0} \end{vmatrix} > 0. \quad (\text{A.6})$$

Notice that for $i=1,2,3$:

$$\frac{s_i}{1+s_i\lambda_0} = \frac{1-e^{-\lambda_0 u_i}}{\lambda_0} = \int_0^{u_i} e^{-\lambda_0 v} dv; \quad (\text{A.7})$$

$$\frac{x_i}{1-x_i\mu_0} = \frac{e^{c\lambda_0 u_i} - 1}{\mu_0} = \frac{c\lambda_0}{\mu_0} \int_0^{u_i} e^{c\lambda_0 v} dv; \quad (\text{A.8})$$

$$u_i = \int_0^{u_i} 1 dv. \quad (\text{A.9})$$

Substituting (A.7)-(A.9) in (A.6) yields another equivalent inequality:

$$\frac{c\lambda_0}{\mu_0} \int_{v_1=0}^{u_1} \int_{v_2=u_1}^{u_2} \int_{v_3=u_2}^{u_3} \begin{vmatrix} e^{-\lambda_0 v_1} & 1 & e^{c\lambda_0 v_1} \\ e^{-\lambda_0 v_2} & 1 & e^{c\lambda_0 v_2} \\ e^{-\lambda_0 v_3} & 1 & e^{c\lambda_0 v_3} \end{vmatrix} dv_3 dv_2 dv_1 > 0. \quad (\text{A.10})$$

As a direct consequence of lemma 3 the determinant in the integrand of (A.10) is strictly positive. Therefore inequality (A.10) and equivalently (4.12) are satisfied. Lemma 4.2 is now completely proved. \square

Appendix B: Proof of theorem 5.1

It is easy to check that

$$\begin{aligned} \frac{d}{dM} h(M) &= \frac{-n}{M+\eta} + \sum_{i=1}^n \frac{1}{M+i}; \\ \frac{d^2}{dM^2} h(M) &= \frac{n}{(M+\eta)^2} - \sum_{i=1}^n \left[\frac{1}{M+i} \right]^2. \end{aligned}$$

If for a finite value M' of M

$$\frac{d}{dM} h(M') = 0, \quad (\text{B.1})$$

then

$$\begin{aligned} \frac{d^2}{dM^2} h(M') &= \frac{n}{(M'+\eta)^2} - \sum_{i=1}^n \left[\frac{1}{M'+i} \right]^2 \\ &= \frac{1}{n} \left[\sum_{i=1}^n \frac{1}{M'+i} \right]^2 - \sum_{i=1}^n \left[\frac{1}{M'+i} \right]^2 \\ &< 0, \end{aligned} \quad (\text{B.2})$$

as $n \geq 2$. So there is at most one finite value M' for M for which (B.1) holds, because if there were two of them, say $M'_1 < M'_2$, then we would have $\frac{d^2}{dM^2} h(M) \geq 0$ for either one of them, or there would be an intermediate value M'_3 , $M'_1 < M'_3 < M'_2$ for which $\frac{d}{dM} h(M'_3) = 0$, $\frac{d^2}{dM^2} h(M'_3) \geq 0$. Both options are, however, excluded by (B.2).

- **Case 1:** $\eta \leq n \left[\sum_{i=1}^n \frac{1}{i} \right]^{-1}$.

It is easy to check that $\frac{d}{dM} h(0) \leq 0$. From this and (B.2) it follows immediately that $\frac{d}{dM} h(M) < 0$ for all $M > 0$. Hence $M=0$ provides the unique local maximum of $h(M)$.

- **Case 2:** $\eta \geq \frac{n+1}{2}$.

We find that for all $M \geq 0$:

$$\begin{aligned} \frac{d}{dM}h(M) &= \frac{-n}{M+\eta} + \sum_{i=1}^n \frac{1}{M+i} \\ &\geq \frac{-n}{M+\frac{n+1}{2}} + \sum_{i=1}^n \frac{1}{M+i} \\ &\geq 0. \end{aligned}$$

So the supremum is reached at $M = \infty$.

- **Case 3:** $n \left[\sum_{i=1}^n \frac{1}{i} \right]^{-1} < \eta < \frac{n+1}{2}$.

We have $\frac{d}{dM}h(0) > 0$, but

$$\begin{aligned} \lim_{M \rightarrow \infty} M^2 \frac{d}{dM}h(M) &= n\eta - \sum_{i=1}^n i \\ &< 0. \end{aligned}$$

So there exists a M_0 , $0 < M_0 < \infty$, such that for all $M > M_0$ we have $M^2 \frac{d}{dM}h(M) < 0$ and hence also $\frac{d}{dM}h(M) < 0$. Thus again there exists a finite positive value M' for which $\frac{d}{dM}h(M') = 0$ and the unique local maximum of the function $h(M)$ is realised.

This completes the proof of theorem 5.1. □

References

- Aalen, O.O. (1978), Non-parametric inference for a family of counting processes. *Annals of Statistics* **6** 701-726.
- Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1992), *Statistical Methods Based on Counting Processes*, Springer-Verlag, New-York. To appear.
- Borgan, Ø. (1984), Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics* **1** 1-16.
- Gantmacher, F.R. (1954), *The Theory of Matrices*. Chelsea Publishing Company, New York.
- Geurts, W.A.J., Hasselaar, M.M.A. & Verhagen, J.H. (1988), *Large Sample Theory for Statistical Inference in Several Software Reliability Models* Report MS-R8807, Centre for Mathematics and Computer Science, Amsterdam.
- Greenberg, M. (1966), *Lectures on Algebraic Topology*, W.A. Benjamin, Inc., Menlo Park, California.
- Jacod, J. (1975), Multivariate point processes: Predictable projection, Radon-Nikodym derivatives, representation of martingales. *Z. Wahrsch. verw. Geb.*, **31**, 235-253.
- Jelinski, Z. & Moranda, P. (1972), Software Reliability Research. *Statistical Computer Performance Evaluation* 465-484.
- Kurtz, T.G. (1981), *Approximation of Population Processes* Regional Conference Series in Applied Mathematics **36**, SIAM Philadelphia, Pennsylvania.
- Kurtz, T.G. (1983), Gaussian approximations for Markov chains and counting processes. *Bulletin of the International Statistical Institute* **50** 361-375.

- Le Cam, L. (1990), Maximum likelihood: an introduction. *International Statistical Review*, Volume 58, No. 2, 153-171.
- Littlewood, B. (1980), Theories of software reliability: How good are they and how can they be improved? *IEEE Transactions on Software Engineering* 6 489-500.
- Moek, G. (1983), *Software Reliability Models on Trial: Selection, Improved Estimation, and Practical Results*. Report MP 83059 U, National Aerospace Laboratory NLR, Amsterdam.
- Moek, G. (1984), Comparison of some software reliability models for simulated and real failure data. *International Journal of Modelling & Simulation*, Volume 4, No. 1, 29-41.
- Musa, J.D. (1975), A theory of software reliability and its application. *IEEE Transactions on Software Engineering*, 3, 312-327.
- Musa, J.D., Iannino, A. & Okumoto, K. (1987), *Software Reliability: Measurement, Prediction, Application*. McGraw-Hill Book Company, New York.
- van Pul, M.C. (1990), *Asymptotic Properties of Statistical Models in Software Reliability*. Report BS-R9011, Centre for Mathematics and Computer Science, Amsterdam. (To appear in SJS, 1992)
- Vetterling, W.T., Teukolsky, S.A., Press, W.H. & Flannery, B.D. (1985), *Numerical Recipes: The art of Scientific Computing* Cambridge University Press, Cambridge.