Estimating a Parametric Component Lifetime Distribution from a Collection of Superimposed Renewal Processes

Wei Zhang  
*Iowa State University*

Ye Tian  
*Iowa State University*

Luis A. Escobar  
*Louisiana State University at Baton Rouge*

William Q. Meeker  
*Iowa State University, wqmeeker@iastate.edu*

Follow this and additional works at: [https://lib.dr.iastate.edu/stat_las_pubs](https://lib.dr.iastate.edu/stat_las_pubs)  
Part of the [Statistics and Probability Commons](https://lib.dr.iastate.edu/stat_las_pubs)

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/stat_las_pubs/286](https://lib.dr.iastate.edu/stat_las_pubs/286). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Estimating a Parametric Component Lifetime Distribution from a Collection of Superimposed Renewal Processes

Abstract
Maintenance data can be used to make inferences about the lifetime distribution of system components. Typically, a fleet contains multiple systems. Within each system, there is a set of nominally identical replaceable components of particular interest (e.g., 2 automobile headlights, 8 dual in-line memory module (DIMM) modules in a computing server, 16 cylinders in a locomotive engine). For each component replacement event, there is system-level information that a component was replaced, but no information on which particular component was replaced. Thus, the observed data are a collection of superpositions of renewal processes (SRP), one for each system in the fleet. This article proposes a procedure for estimating the component lifetime distribution using the aggregated event data from a fleet of systems. We show how to compute the likelihood function for the collection of SRPs and provide suggestions for efficient computations. We compare performance of this incomplete-data maximum likelihood (ML) estimator with the complete-data ML estimator and study the performance of confidence interval methods for estimating quantiles of the lifetime distribution of the component. Supplementary materials for this article are available online.

Keywords
Component reliability, Log-location-scale family, Maximum likelihood estimation, Recurrence data, Relative efficiency, Superposition of renewal processes

Disciplines
Statistics and Probability

Comments

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/stat_las_pubs/286
Estimating a Parametric Component Lifetime Distribution from a Collection of Superimposed Renewal Processes

Wei Zhang
Department of Statistics
Iowa State University
Ames, IA 50011
wzhang87@iastate.edu

Ye Tian
Department of Statistics
Iowa State University
Ames, IA 50011
tianye1984@gmail.com

Luis A. Escobar
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70808
luis@lsu.edu

William Q. Meeker
Department of Statistics
Iowa State University
Ames, IA 50011
wqmeeker@iastate.edu

February 14, 2016

Abstract

Maintenance data can be used to make inferences about the lifetime distribution of system components. Typically a fleet contains multiple systems. Within each system there is a set of nominally identical replaceable components of particular interest (e.g., two automobile headlights, eight DIMM modules in a computing server, sixteen cylinders in a locomotive engine). For each component replacement event, there is system-level information that a component was replaced, but not information on which particular component was replaced. Thus the observed data is a collection of superpositions of renewal processes (SRP), one for each system in the fleet. This paper proposes a procedure for estimating the component lifetime distribution using the aggregated event data from a fleet of systems. We show how to compute the likelihood function for the collection of SRPs and provide suggestions for efficient computations. We compare performance of this incomplete-data ML estimator with the complete-data ML estimator and study the performance of confidence interval methods for estimating quantiles of the lifetime distribution of the component.

Keywords

Component reliability; Log-location-scale family; Maximum likelihood estimation; Recurrence data; Relative efficiency; Superposition of renewal processes
1 Introduction

1.1 Background

Repairable systems arise and are of particular interest in many industrial reliability applications. Normally when there is a failure of a repairable system, a component replacement will restore the system operation. If we assume that a replaced component has the same lifetime distribution as the old one, then the observed recurrent event data can be represented by a renewal process. In practice, it is common that the system under observation contains a collection of similar replaceable components (e.g., valve seats or cylinders in a diesel locomotive engine). Typically the replacement data are available in an aggregate form (i.e., event time for each replacement is available, but we do not know which component underwent the replacement). In this case the aggregate data form a superposition of renewal processes (SRP) (see, for example, pages 47-51 of [Ascher and Feingold, 1984]). In the following, we use two examples to illustrate the data structure we described above.

1.2 Examples

Cylinder data

[Nelson (2003)] presents the recurrence data of a fleet of 120 diesel engines. Each engine has 16 cylinders and the cylinders can develop problems leading to leaks or low compression. Figure 1 shows the event plot for the cylinder replacement for a subset of 30 engines. The event plot tells us which engine (system-level information) each replacement comes from, but not the information about which cylinder position (or “socket”) inside the engine. The missing socket-level information makes it more difficult to estimate the component failure-time distribution.

![Engine cylinder replacement](image)

Figure 1: Event plot of diesel engine cylinder replacement
Automobile-component data

We also have recurrence data from an automotive system. For this application, there were 144,102 vehicles in the fleet, and each vehicle had two identical components. Similar to the cylinder data, the component (socket) level information of the replacements is unknown.

The estimation of the lifetime distribution of particular system components is important for many purposes, such as collecting information for future system design and maintenance planning for individual units. However, due to the missing information about position within a system for replacement events, it is challenging to estimate the lifetime distribution of the particular components. In this paper, we propose a method for estimating the component lifetime distribution from the aggregated event data consisting of an SRP for each system. With a model assumption for the component lifetime distribution, we derive the likelihood function for the observed recurrent data for the SRP, by considering all possible allocations of the recurrent events to sockets in the SRP. Then we obtain the ML estimates by maximizing the likelihood function.

1.3 Related work

Several other approaches have been explored to analyze the SRP recurrence data. Barlow and Proschan (1996) gave general discussions of SRPs and some limiting results. In general, the SRP will not be a renewal process unless the component renewal processes are homogeneous Poisson processes (HPP). Drenick (1960) showed that when the number of systems is large and the time is far away from the origin, an SRP behaves like a HPP (Drenick’s theorem). Khinchin (1956) showed that if the number of sockets in an SRP is large, the SRP behaves like a nonhomogeneous Poisson process (NHPP).

Krivtsov and Frankstein (2014) present statistical methods to distinguish between the situation in a multi-socket system (with the same type of component in each socket) where there is an SRP because the failure-time distribution is the same in all sockets versus the situation where all or most of the failures are coming from one socket because of a system problem (e.g., components are stressed more in one socket, relative to the others).

There have been few studies to estimate the component failure-time distribution from an SRP. For applications involving aircraft components, Peixoto (2009) proposed a method to estimate the failure-time distribution by assigning the event times to sockets randomly, and then using simulation to correct for bias. In Peixoto’s study, a second layer of simulation is needed to quantify statistical uncertainty (e.g. to compute confidence intervals).

There has been some other work to estimate component lifetime distribution without system-level information. Trindade and Haugh (1979, 1980) proposed a nonparametric estimator of the lifetime distribution, based on the deconvolution of the renewal equation. Baxter (1994) discussed a problem in telecommunications system component reliability where the recorded data are the numbers of failed components returned by the customers at a sequence of equally spaced time intervals (e.g., each month). The author derived a nonparametric estimator of the lifetime distribution function in a discretized manner and also fit a Weibull distribution to the nonparametric estimates. The methods proposed based on the renewal function deconvolution perform poorly if the nonparametric estimate of the distribution approaches 1 (e.g., if the expected number of events per socket approaches or exceeds 1). Tortorella (1996) proposed an
estimation procedure by building a pooled discrete renewal process model and estimating the component reliability based on a maximum likelihood-like method.

In this paper, we describe a likelihood-based method to estimate the component lifetime distribution. The only limitation is the computationally intensive nature of the estimation method for SRPs that have a large number of events (e.g., more than 15 events). Considering this, we know that the proposed method will be especially useful for dealing with a fleet of SRPs where each SRP only has a relatively small number of events (common in most applications). As long as the total number of systems is not too small, the total number of events in the fleet would be large enough to enable precise estimation of component reliability.

1.4 Overview

The remainder of this paper is organized as follows. Section 2 describes the data structure and the proposed model. Section 3 describes some basic results from combinatorics that are needed to compute the SRP likelihood. Using these basic results, Section 4 first derives the likelihood function for a single SRP (system) and then shows how to compute the likelihood for a whole fleet of SRPs (systems). Sections 5 and 6 illustrate the methods for two different applications. Section 7 provides information on the amount of computer time that is needed to compute the SRP likelihood as a function of $r$ and $m$. Section 8 describes a simulation study that compares the SRP ML estimator with that of the complete-data renewal process estimator. Section 9 provides some conclusions and the discussion of future work.

2 Data Structure and Model

2.1 Superposition of a renewal process

We consider a fleet of $n$ independent systems where each system contains $m$ components operating in $m$ sockets. When a component fails, the failed component is replaced by a new one in the same socket. We assume that the lifetime of a component, $T$, has a cdf $F(t; \theta)$ and pdf $f(t; \theta) = dF(t; \theta)/dt$, where $\theta$ is a vector of unknown parameters. For example, the Weibull distribution cdf is

$$F(t; \beta, \eta) = 1 - \exp\left(-\frac{t}{\eta}\right)^\beta,$$

where $\eta > 0$ is a scale parameter, $\beta > 0$ is a shape parameter, and $\theta = (\eta, \beta)$. With an iid assumption, the event history for a single socket is a renewal process (RP). To illustrate this, let $T_i$ denote the lifetime for the component before replacement $i = 1, 2, 3, \ldots$, then the event times in the socket, $T_j = \sum_{i=1}^j T_i$, $j = 1, 2, 3, \ldots$ form a RP (e.g., Cox 1962).

In our application, for each replacement, we only know the system index (identifying the system), but not the socket where the replacement was made within the system. Each system-level set of replacement times forms an SRP.
2.2 SRP data

Let \( \mathcal{H}_{\tau_c} = (\tau_1, \tau_2, \ldots, \tau_r, \tau_c) \) denote the observed event history of a single SRP with event times \( \tau_1 < \cdots < \tau_r \), and end-of-observation time \( \tau_c \) with \( \tau_c \geq \tau_r \). A data set will consist of \( n \) independent SRPs corresponding to the \( n \) systems in the fleet.

In summary, the assumptions in our model are: we assume that the component cdf is the same in all sockets in all system and over time, that the failures within a socket are independent, and that all sockets within one system have the same end-of-observation time \( \tau_c \), but that \( \tau_c \) can (and often will) differ from system to system. We also assume that the \( n \) systems in the fleet are independent.

2.3 Log-location-scale family of distributions

An alternative form for the Weibull cdf is

\[
F(t; \beta, \eta) = \Phi_{\text{sev}}\left(\frac{\log(t) - \mu}{\sigma}\right),
\]

where \( \mu = \log(\eta) \), \( \sigma = 1/\beta \), and \( \Phi_{\text{sev}}(z) = 1 - \exp[-\exp(z)] \) is the cdf for a standardized smallest extreme value distribution with location parameter \( \mu = 0 \) and scale parameter \( \sigma = 1 \). That is, if a positive random variable \( T \) follows a Weibull distribution, then \( \log(T) \) follows the smallest extreme value distribution.

The Weibull distribution is one important distribution in log-location-scale family distributions. By changing the definition of \( \Phi \), one can obtain other similar distributions in log-location-scale family, such as the lognormal, loglogistic, and the Fréchet distributions. For example, for the lognormal distribution, replace \( \Phi_{\text{sev}} \) with the standard normal cdf \( \Phi_{\text{nor}} \) and for the Fréchet distribution, replace \( \Phi_{\text{sev}} \) with the standard largest extreme value cdf \( \Phi_{\text{lev}}(z) = \exp[-z - \exp(-z)] \).

We illustrate the SRP estimation method by using the log-location-scale family because it contains the most commonly used statistical distributions in reliability applications (i.e., the Weibull and lognormal distributions). These distributions are used not only because they are flexible, but also because there are physical motivations justifying their use. For example, the Weibull distribution can be used to describe time to failure when the failure mechanism is cause by a minimum process (e.g., fracture of a brittle material). The lognormal distribution is commonly used to describe time to failure from a cumulative-damage process (e.g., chemical degradation or fatigue crack growth in metals, Section 4.6 of [Meeker and Escobar 1998]). Similarly, the Fréchet distribution will generally provide a useful description for failures driven by a maximum process (e.g., failure occurs when the last of several redundant components fails). Thus in this paper, we mainly focus on log-location-scale family distributions, although the general approach could be easily applied to other parametric probability distributions.
3 SRP Likelihood Preliminaries and Counting Methods

In this section, we define a likelihood for a single system (SRP) under the assumption that system-level information is available in a fleet of systems (set of independent and identical SRPs). Then the likelihood for the whole fleet is the product of all of the SRP likelihoods.

3.1 Data configurations

For a single system, let $r$ denote the number of observed events. Suppose there are $m$ statistically independent and identical renewal processes in $m$ sockets within a system, and the sockets are labeled sequentially 1, 2, 3, \ldots, $m$.

**Definition 1** A data configuration is the assignment of the $r$ observed event times to the $m$ sockets in one SRP.

In the complete data case, the true data configuration is known and we can write the likelihood directly according to actual assignment of events. Because the socket-level information is not available for incomplete data, numerous data configurations could lead to the observed $r$ events for the SRP. We obtain the likelihood for an SRP by enumerating all possible data configurations, computing the probability of the data for each of these data configurations, and summing all of these probabilities.

**Result 1** For a system (SRP) with $m$ sockets and $r$ events, the number of all possible data configurations leading to $r$ observed events is equal to $m^r$.

This result follows from noting that for each observed event, there are $m$ distinct ways to assign that event to one of the $m$ sockets. Therefore, according to the “multiplication counting principle,” there are $m^r$ possible data configurations that would generate these $r$ observed events in the SRP. For example, Figure 2 shows all $3^2 = 9$ data configurations for the situation where there are $m = 3$ sockets and $r = 2$ events ($\tau_1$, $\tau_2$).

To obtain the likelihood of the $3^2 = 9$ independent data configurations, we sum the joint probabilities for each of the $3^2 = 9$ data configurations to get the marginal probability for the observed SRP. For continuous data, it is convenient to use the density approximation instead of the actual probability in likelihood calculation. Then the likelihood (proportional to the probability of the data) for this simple SRP in Figure 2 is

$$L = \sum_{i=1}^{9} L_i = 3 \times f(\tau_1) f(\tau_2 - \tau_1) S(\tau_c - \tau_2) [S(\tau_c)]^2 + 6 \times f(\tau_1) S(\tau_c - \tau_1) \times f(\tau_2) S(\tau_c - \tau_2) S(\tau_c)$$

where $\tau_c$ is the end-of-observation time, $f(t)$ and $S(t)$ are respectively the probability density function and survival functions for the failure time distribution.
Figure 2: All $3^2$ data configurations for the SRP with $m = 3$ and $r = 2$. Note that the event times are the same in each data configuration.

Therefore the likelihood contribution of the $3^2 = 9$ data configurations are reduced to two unique terms according to their contribution to the SRP likelihood. The first term $f(\tau_1)f(\tau_2 - \tau_1)S(\tau_c - \tau_2)[S(\tau_c)]^2$ corresponds to the situation where the $r = 2$ events occur in a single socket. The socket label could be 1, 2, or 3 (as shown in the first row of Figure 2). All three cases are equivalent in the sense that each one provides an equal likelihood contribution to the SRP likelihood $L$. The second term $f(\tau_1)S(\tau_c - \tau_1) \times f(\tau_2)S(\tau_c - \tau_2)S(\tau_c)$ corresponds to the situation where the $r = 2$ events occur in two different sockets. According to the last two rows of Figure 2, the labels of the two sockets could be (1, 2), (1, 3), (2, 1), (2, 3), (3, 1) or (3, 2). Similarly, each of these six label permutations has the same likelihood contribution. The remainder of this section describes how to handle the general case of $r$ events in $m$ sockets, providing a basis to compute the likelihood for an SRP.

3.2 Partitions of an integer

Definition 2 A partition of a positive integer $r$ is a list of nonincreasing positive integers that sum to $r$.

In our application, a partition indicates how the $r$ events could have occurred in the $m$ sockets, without regard to time order or socket label. For a positive integer $r$, let $h$ denote the total number of possible distinct partitions. Note there is no closed form equation for $h$, but we can
obtain $h$ by using recursive computations. The $i$th particular partition of the $r$ events is denoted by $E_i^r = (r_1, r_2, \ldots, r_h)$, where $i = 1, 2, \ldots, h$, and $E_i^r$ is sometimes called the “shape” of the partition, see [Hankin and West (2007)]. Here $l_i \leq m$ is the length of the partition (i.e., the number of sockets that contain events), and $\sum_{i=1}^{h} r_i = r$.

For example, for a SRP with $r = 3$ events and $m \geq r$ sockets, there are $h = 3$ partitions:

1. $E_1^3 = (3)$: All three events take place in one socket.
2. $E_2^3 = (2, 1)$: Two events take place in one socket, the other event occurs in another socket.
3. $E_3^3 = (1, 1, 1)$: The three events take place in three different sockets.

Because of the restriction $l_i \leq m$, we are, in general, dealing with restricted partitions. For example if $r = 3$ events and there are $m = 2$ sockets, the partition $E_3^3 = (1, 1, 1)$ is not possible.

### 3.3 Set partitions of a partition of an integer

**Definition 3** A set partition enumerates the distinct equivalence relations corresponding to a particular partition $E_i^r$.

See [Hankin and West (2007)] for a more detailed description of set partitions, an algorithm to compute them, other applications, and references for the underlying theory. In our application, set partitions correspond to unique-likelihood configurations, given by the arrangement of the event times within the $l_i$ event-containing sockets corresponding to a particular SRP. In general, for a given partition $E_i^r$, [Hankin and West (2007)] give the result that there are

$$s_i = \frac{r!}{\prod_{j=1}^{l_i} r_j! \times \prod_{j=1}^{q_i} f_j!}$$

equivalence relations (statistically unique-likelihood configurations) where $q_i$ is the number of unique digits in the partition $E_i^r$ and $f_1 \ldots f_{q_i}$ are the frequencies with which the unique digits appear in the partition.

To illustrate this with a simple example, consider data with $r = 3$ events and $m = 4$ system sockets. For one particular partition $E_2^3 = (2, 1)$, the length of the partition is $l_2 = 2$, which indicates that 2 of the $m = 4$ sockets contain events. We will refer to the $l_2 = 2$ sockets that have events as socket A and socket B. The allocation of the $r = 3$ events to the $l_2 = 2$ event-containing sockets results in $s_2 = 3!/(2!1! \times 1!1!) = 3$ different unique-likelihood configurations given in Table 1.
Consider the unique-likelihood configuration \((\tau_1, \tau_3 | \tau_2)\) corresponding to partition \((2, 1)\) (shown in the first row of Table 1) as an example. This unique-likelihood configuration describes the situation where \(\tau_1, \tau_3\) occur in event-containing socket A, and \(\tau_2\) takes place in event-containing socket B. There are no events in the other \((m - 2) = 2\) sockets.

Then the corresponding likelihood of given unique-likelihood configuration \((\tau_1, \tau_3 | \tau_2)\) is

\[
f(\tau_1)f(\tau_3 - \tau_1)S(\tau_c - \tau_3) \times f(\tau_2)S(\tau_c - \tau_2) \times [S(\tau_c)]^2
\]

where \(f(\tau_1)f(\tau_3 - \tau_1)S(\tau_c - \tau_3)\) and \(f(\tau_2)S(\tau_c - \tau_2)\) are the likelihoods for event-containing socket A and event-containing socket B, respectively. The remaining two sockets containing no events has a likelihood \([S(\tau_c)]^2\). The likelihoods corresponding to all \(s_i = 3\) unique likelihood configurations are listed in the fourth column of Table 1.

### 3.4 Equivalent socket permutations

In the example in Section 3.3, there are \(k_i = 4!/(4 - 2)! = 4!/2! = 12\) ways that the \(l_i = 2\) event-containing sockets can be allocated to the \(m = 4\) system sockets, as shown in Table 2. All of these \(k_i = 12\) arrangements are statistically equivalent in the sense that they have exactly the same likelihood and only differ in socket labels.
Table 2: All arrangements of the $l_i = 2$ event-containing sockets for partition $(2, 1)$ among the $m = 4$ system sockets

<table>
<thead>
<tr>
<th>System sockets containing events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event-containing</td>
</tr>
<tr>
<td>socket A 1 1 1 2 2 2 3 3 3 4 4 4</td>
</tr>
<tr>
<td>Event-containing</td>
</tr>
<tr>
<td>socket B 2 3 4 1 3 4 1 2 4 1 2 3</td>
</tr>
<tr>
<td>Arrangement number</td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
</tbody>
</table>

More generally, the $l_i$ event-containing sockets can be arranged within the $m$ system sockets

$$k_i = \frac{m!}{(m-l_i)!}$$
different ways, where $(m-l_i)!$ is equal to 1 when $m = l_i$. That is, the $i^{th}$ partition $E_i^r = (r_1, r_2, \ldots, r_{l_i})$ has $k_i$ statistically equivalent socket permutations. Note that when all system sockets contain at least one event $l_i = m$ and then $k_i = m!$.

3.5 Summary of SRP partitioning and examples

Sections 3.2, 3.3, and 3.4 describe a natural method for enumerating all data configurations of a given number of events $r$ in one SRP having $m$ sockets:

a) Enumerate all possible integer partitions $E_i^r$ of $r$ and for each partition,

b) Enumerate the unique-likelihood configurations, and

c) Compute the number of socket permutations.

Table 3 lists all partitions of the integer $r$ for $1 \leq r \leq 6$, assuming one SRP with $m = 16$ sockets. For each partition, the number of unique-likelihood configurations $s_i$ and socket permutations $k_i$ are also given. Note that when $m \geq r$, $B_r = \sum_{i=1}^{h} s_i$ is the number of partitions of an $r$-element set, known as the Bell number (Rota, 1964). For a fixed value of $r$, the number of data configurations is $m^r$ according to Result 1. Using the procedure described in the previous sections and illustrated by the examples in Table 3, the number of data configurations can also be computed as $\sum_{i=1}^{h} k_i s_i$. That is, $\sum_{i=1}^{h} k_i s_i = m^r$. For example, $r = 5$, $h = 7$, the number of data configurations is $\sum_{i=1}^{7} k_i s_i = 1048576 = 16^5$.

The examples in Table 3 show that the number of unique-likelihood configurations $s_i$ increases rapidly with the number of events $r$ in one SRP. The amount of time required to compute an SRP likelihood will be approximately proportional to $s_i$. The number of unique-likelihood configurations in this table correspond to Bell numbers in these cases because $m > r$. 
Table 3: Examples for an SRP with \( m = 16 \) sockets

<table>
<thead>
<tr>
<th>( r )</th>
<th>( i )</th>
<th>Partition of ( r )</th>
<th>Length of partition ( l_i )</th>
<th>Socket permutations ( k_i = \frac{m!}{(m - l_i)!} )</th>
<th>Unique-likelihood configurations ( s_i )</th>
<th>( B_r = \sum_{i=1}^{h} s_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>(2)</td>
<td>1</td>
<td>( 16!/(16 - 1)! = 16 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(1, 1)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>(3)</td>
<td>1</td>
<td>( 16!/(16 - 1)! = 16 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>(2, 1)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>(1, 1, 1)</td>
<td>3</td>
<td>( 16!/(16 - 3)! = 3360 )</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>(4)</td>
<td>1</td>
<td>( 16!/(16 - 1)! = 16 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>(3, 1)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>(2, 2)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>(2, 1, 1)</td>
<td>3</td>
<td>( 16!/(16 - 3)! = 3360 )</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>(1, 1, 1, 1)</td>
<td>4</td>
<td>( 16!/(16 - 4)! = 43680 )</td>
<td>15</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>(5)</td>
<td>1</td>
<td>( 16!/(16 - 1)! = 16 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>(6)</td>
<td>1</td>
<td>( 16!/(16 - 1)! = 16 )</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>(5, 1)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>(4, 2)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>(4, 1, 1)</td>
<td>3</td>
<td>( 16!/(16 - 3)! = 3360 )</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>(3, 3)</td>
<td>2</td>
<td>( 16!/(16 - 2)! = 240 )</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>(3, 2, 1)</td>
<td>3</td>
<td>( 16!/(16 - 3)! = 3360 )</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>(3, 1, 1, 1)</td>
<td>4</td>
<td>( 16!/(16 - 4)! = 43680 )</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>(2, 2, 2)</td>
<td>3</td>
<td>( 16!/(16 - 3)! = 3360 )</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>(2, 2, 1, 1)</td>
<td>4</td>
<td>( 16!/(16 - 4)! = 43680 )</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>(2, 1, 1, 1)</td>
<td>5</td>
<td>( 16!/(16 - 5)! = 524160 )</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>(1, 1, 1, 1, 1)</td>
<td>6</td>
<td>( 16!/(16 - 6)! = 5765760 )</td>
<td>1</td>
<td>203</td>
</tr>
</tbody>
</table>

4 SRP likelihood

The likelihood for a single SRP is the sum of the likelihoods for all possible data configurations that could have led to the observed SRP. In Section 3, we outlined a general procedure to enumerate all data configurations. In this section, we will give a more formal and complete description of the likelihood for a single SRP (corresponding to one system in the fleet) and show how to compute the log likelihood for a fleet of similar systems.
4.1 The likelihood for a single system

Consider an SRP with \( m \) sockets. Given the observed event history \( H_{\tau_c} = (\tau_1, \tau_2, \ldots, \tau_r, \tau_c) \) of the SRP has events at time \( \tau_1 < \cdots < \tau_r, \) \( r \) is a positive integer, and \( \tau_c \geq \tau_r. \) Thus there are \( R = r \) observed events.

We use \( \cup_{d=1}^{m^r} D^r_d \) to denote all the possible data configurations that could lead to the observed event history \( H_{\tau_c} = (\tau_1, \tau_2, \ldots, \tau_r, \tau_c). \) To enumerate all data configurations, we start by enumerating all \( h \) partitions of \( r \) observed events. There are \( s_i \) distinct unique-likelihood configurations \( B_{i,j}^r \), \( j = 1, 2, \ldots, s_i \) for each socket permutation \( l, l = 1, 2, \ldots, k_i. \) All \( k_i \) socket permutations are statistically equivalent because they correspond to the same set partition (unique-likelihood configuration) and differ only in the socket labels. Thus, as described in Section 3, all the data configurations can be represented by

\[
\cup_{d=1}^{m^r} D^r_d = \cup_{i=1}^{h} \left[ \cup_{l=1}^{k_i} \left( \cup_{j=1}^{s_i} B_{i,j}^r \right) \right].
\]

**Result 2** The likelihood \( L(\theta; H_{\tau_c}) \) for the observed SRP, where \( \theta \) is a vector of unknown parameters, is defined as

\[
L(\theta; H_{\tau_c}) = \Pr (H_{\tau_c}; \theta) = \Pr (H_{\tau_c} \cap R; \theta)
\]

\[
= \Pr (H_{\tau_c} \cap \left( \bigcup_{d=1}^{m^r} D^r_d \right); \theta)
\]

\[
= \Pr \left( H_{\tau_c} \cap \left\{ \bigcup_{i=1}^{h} \left[ \bigcup_{l=1}^{k_i} \left( \bigcup_{j=1}^{s_i} B_{i,j}^r \right) \right] \right\}; \theta \right)
\]

\[
= \sum_{i=1}^{h} \Pr \left( H_{\tau_c} \cap \left( \bigcup_{l=1}^{k_i} \left( \bigcup_{j=1}^{s_i} B_{i,j}^r \right) \right); \theta \right)
\]

\[
= \sum_{i=1}^{h} k_i \sum_{j=1}^{s_i} \Pr (H_{\tau_c} \cap B_{i,j}^r; \theta)
\]

\[
= \sum_{i=1}^{h} k_i \sum_{j=1}^{s_i} L_{i,j} (\theta; H_{\tau_c})
\]

where \( L_{i,j} (\theta; H_{\tau_c}) \) is the likelihood corresponding to the \( j^{th} \) unique-likelihood configuration \( B_{i,j}^r \) for the \( i^{th} \) partition \( E_i^r = (r_1, r_2, \ldots, r_{l_i}). \)

To express the likelihood \( L_{i,j} (\theta; H_{\tau_c}) \), we first obtain the likelihoods corresponding to all sockets within the system and calculate the sum. Among the \( m \) sockets of the SRP, there are \( l_i \) event-containing sockets and \( (m - l_i) \) sockets in which no events occurred. For the \( (m - l_i) \) sockets with zero events, the likelihood is

\[
[S(\tau_c)]^{(m-l_i)}.
\]

For event-containing socket \( a, a = 1, 2, \ldots, l_i, \) we need to relabel the \( r_a \) event times that occur in socket \( a \) as \( \tau_{a,1}, \tau_{a,2}, \ldots, \tau_{a,r_a}. \) Then using the density approximations to replace the actual
probability statements for the reported event times and letting $\tau_0 = 0$, the likelihood for event-containing sockets is proportional to
\[
f(\tau_{a,1}) \times f(\tau_{a,2} - \tau_{a,1}) \times \cdots \times f(\tau_{a,r_a} - \tau_{a,r_a-1}) \times S(\tau_c - \tau_{a,r_a}).
\] (3)

Consequently

\[
\Pr \left( H_{\tau_c} \cap B_{i,j} ; \theta \right) \propto L_{i,j}(\theta; H_{\tau_c})
\]
and can be computed as the product of (2) and (3).

4.2 The likelihood for a fleet of systems

Multiple systems are common in most applications that we have encountered and usually having system label information makes it computationally feasible to compute the SRP likelihood directly. Section 4.1 showed how to compute the likelihood for a single SRP, and notation indicating system index was suppressed. To compute the likelihood of a fleet of $n$ systems having identical SRPs, it is necessary to label systems sequentially as $k = 1, 2, \ldots, n$.

Suppose that there are $m_k$ sockets and $r_k$ events in system $k$, $k = 1, 2, \ldots, n$. Let $H^k_{\tau_c} = (\tau^k_1, \tau^k_2, \ldots, \tau^k_{r_k}, \tau^k_c)$ denote the observed event history of system $k$, where observed failure times $\tau^k_1 < \cdots < \tau^k_{r_k}$, and end-of-observation time $\tau^k_c \geq \tau^k_{r_k}$. Then the likelihood for system $k$, $L_k(\theta; H^k_{\tau_c})$, is computed using the procedure described in Section 4.1.

Result 3 Under the assumption that all systems in the fleet are independent and have identically-distributed component failure times, the total log likelihood for the fleet is

\[
\mathcal{L}\left[\theta; (H^1_{\tau_c}, H^2_{\tau_c}, \ldots, H^n_{\tau_c})\right] = \sum_{k=1}^{n} L_k(\theta; H^k_{\tau_c})
\]
\[
= \sum_{k=1}^{n} \log \left[ L_k(\theta; H^k_{\tau_c}) \right].
\] (4)

5 Application to the Engine Cylinder Replacement Data

5.1 Data description

In this section, we apply the proposed maximum likelihood estimation and confidence interval procedures to analyze the diesel engine cylinder replacement data. Here is a basic description of the cylinder data:

a) The fleet has $n = 120$ engines (systems).

b) Each engine has $m = 16$ cylinders (sockets).

c) There is a total of 156 events.
5.2 Reparameterization to improve log likelihood maximization performance

The log likelihood \( L(\theta; (H_1^{\tau_1}, H_2^{\tau_2}, \ldots, H_n^{\tau_n})) \) for a fleet of multiple SRPs was defined in Section 4.2. Then the maximization of the log likelihood gives the idML estimates \( \hat{\theta} \) where “id” refers to “incomplete data” because there is a lack of socket-level information. Then the relative likelihood takes the form

\[
R(\theta) = \frac{L(\theta; (H_1^{\tau_1}, H_2^{\tau_2}, \ldots, H_n^{\tau_n}))}{L(\hat{\theta}; (H_1^{\tau_1}, H_2^{\tau_2}, \ldots, H_n^{\tau_n}))}
\]

To fit a Weibull distribution to the cylinder data, the parameters of interest are \( \theta = (\mu, \sigma) = (\log(\eta), 1/\beta) \), with \( \sigma > 0 \) (\( \beta > 0 \)), as described in Section 2.3. The choice of parameterization can affect the accuracy of some statistical inferences (e.g., Wald-based confidence intervals depend on parameterization). In addition, there is an expectation that the normal distribution approximation underlying the Wald method will be better on the log scale, which is unrestricted in sign. Because the scale parameter \( \sigma > 0 \), we first used the unrestricted parameters \( \theta = (\mu, \log(\sigma)) \) to simplify the optimization. The left plot in Figure 3 presents the contours of the relative likelihood function \( R(\theta) \) for a Weibull distribution with parameterization \( \theta = (\mu, \log(\sigma)) \). The plot shows there is a positive correlation between \( \mu \) and \( \log(\sigma) \). This is because for the cylinder data, the fraction of sockets with events is \( \approx 156/[120 \times 16] \approx 0.081 \), which means that we expect to have a good estimation of quantiles up to \( t_{0.08} \). Therefore the estimator of the location parameter \( \mu = \log(\eta) = \log(t_{0.632}) \) will be highly correlated with the estimator of the scale parameter \( \sigma \). To eliminate the strong correlation between the parameter estimators, we replaced \( \eta = t_{0.632} \) with the quantile \( t_{0.08} \) and used an alternative unrestricted parameterization \( \theta = (\log(t_{0.08}), \log(\sigma)) \). As seen in the right plot of Figure 3, the likelihood contours suggest that the parameter estimators are approximately uncorrelated, which is a useful property for both maximizing the likelihood and for statistical inference which will be discussed in the following sections.

5.3 Model fitting

5.3.1 Fitting cylinder data using log-location-scale distributions

Besides the Weibull distribution, the Lognormal, Loglogistic and Fréchet distributions were also fit to the cylinder data. Table 4 presents the log likelihood for each of these distributions. One can see that the idML estimation using the Fréchet distribution has the largest log likelihood, suggesting that the Fréchet distribution gives the best fit among the four log-location-scale distributions. Figure 4 shows the cdf estimate for each of these distributions. The dashed line represents the average censoring time for all 120 systems. The fitted distributions are close to each other before the average censoring time. After this average censoring time, the estimates diverge as extrapolation is involved.
Figure 3: Likelihood contours for the cylinder data using two different parameterizations. The $\theta = (\mu, \log(\sigma))$ parameterization is on the left and the $\theta = (\log(t_{0.08}), \log(\sigma))$ parameterization is on the right panel. The dot in the center of the contours corresponds to the idML estimates of the parameters $\theta$.

Table 4: Log likelihood for the cylinder data for all the four log-location-scale distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fréchet</td>
<td>$-1164.8$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$-1168.5$</td>
</tr>
<tr>
<td>Loglogistic</td>
<td>$-1174.0$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$-1174.7$</td>
</tr>
</tbody>
</table>

Figure 4: Fitted log-location-scale distributions for the cylinder data.
5.3.2 MCF plots

In this section, we estimate the mean cumulative function (MCF) of the replacements for the cylinder data, as shown in Figure 5. The solid curves correspond to the nonparametric MCF described, for example, in Nelson (2003), and the dashed curves correspond to the parametric MCF estimates based on the idML estimates for the Weibull (left plot) and Fréchet (right plot) distributions. Because there is no closed form for the parametric MCF, the Weibull and Fréchet fitted MCFs were obtained by simulation. This MCF plot is a useful diagnostic plot to check whether the idML estimate is a good candidate for describing the cylinder data. By comparing the two MCF plots in Figure 5, we see that the Fréchet distribution fitted MCF has a better agreement with the empirical MCF in comparison to the Weibull distribution fitted MCF, which is consistent with the fact that the Fréchet distribution gives largest log likelihood therefore a better fit to the data. One simulation study (details not shown here) indicated that the deviation seen in the right plot of Figure 5 could have come from random noise and there is no statistical evidence for a departure from the Fréchet distribution.

![MCF plots](image)

Figure 5: Weibull (left) and Fréchet (right) fitted MCF plots for the cylinder data. The solid curve and dashed curves correspond to empirical and fitted MCF, respectively.

5.4 Interval estimation

5.4.1 LR confidence intervals

Let \( \theta = (\theta_1, \theta_2) \) be the unknown parameters where \( \theta_1 \) is the parameter of interest and \( \theta_2 \) is a nuisance parameter. In our examples \( \theta \) would be \([\log(t_p), \log(\sigma)]\) (when estimating \( t_p \)) or \([\log(\sigma), \log(t_p)]\) (when estimating \( \beta = 1/\sigma \)). Then the relative profile likelihood for \( \theta_1 \) is

\[
R(\theta_1) = \max_{\theta_2} \left[ \frac{L(\theta_1, \theta_2; (H^1_{c_1}, H^2_{c_2}, \ldots, H^n_{c_n}))}{L(\hat{\theta}_1, \hat{\theta}_2; (H^1_{c_1}, H^2_{c_2}, \ldots, H^n_{c_n}))} \right]
\]

where \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are the idML estimates of \( \theta_1 \) and \( \theta_2 \) respectively. The likelihood ratio statistic is \( LR(\theta_1) = -2 \log R(\theta_1) \) and the asymptotic distribution of \( LR(\theta_1) \) is \( \chi^2_1 \) when evaluated at the
true value of $\theta_1$. Let $\chi^2_{(1,1-\alpha)}$ denote the $(1 - \alpha)$ quantile of a $\chi^2_1$ and suppose the roots of the equation $LR(\theta_1) - \chi^2_{(1,1-\alpha)} = 0$ are $\theta_{1L}$ and $\theta_{1U}$ with $\theta_{1L} \leq \theta_{1U}$. Then the likelihood ratio (LR) confidence interval for $\theta_1$ is $[\theta_{1L}, \theta_{1U}]$ with a nominal confidence level of $(1 - \alpha)$.

### 5.4.2 Wald confidence intervals

For a quantity of interest $\theta_1$, let $\hat{s}_{\theta_1}$ denote the estimate of the standard error of $\tilde{\theta}_1$, which is computed as a function of the elements of the Hessian matrix evaluated at $\tilde{\theta}$. Then the 100$(1 - \alpha)$% Wald confidence interval for $\theta_1$ is $\left[ \hat{\theta}_1 \pm z(1-\alpha) \times \hat{s}_{\tilde{\theta}_1} \right]$, where $1 - \alpha$ is the nominal confidence level and $z(1-\alpha)$ is the 1 - $\alpha$ quantile of the standard normal distribution.

To illustrate the interval estimation for the cylinder data we use the best-fitting Fréchet distribution. The point estimates, 90% LR intervals, and Wald intervals for the Fréchet shape parameter $\beta$ and some quantiles of interest are given in Table 5. Furthermore, the profile likelihoods, LR intervals and Wald intervals for $\beta$ and $t_{0.1}$ for the cylinder data are given in Figure 6. The LR interval and Wald interval give similar results, which suggests that the quadratic approximation for the relative log likelihoods corresponding to these parameters justifies the use of the simpler Wald confidence intervals, and is the reason that there is little difference between the LR and Wald intervals in Table 5.

Table 5: Point estimates and confidence intervals for the quantities of interest for the cylinder data

<table>
<thead>
<tr>
<th>Quantity of interest</th>
<th>idML estimates</th>
<th>90% confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.24</td>
<td>[1.11, 1.38]</td>
</tr>
<tr>
<td>$t_{0.001}$</td>
<td>712</td>
<td>[650, 769]</td>
</tr>
<tr>
<td>$t_{0.01}$</td>
<td>987</td>
<td>[931, 1038]</td>
</tr>
<tr>
<td>$t_{0.05}$</td>
<td>1396</td>
<td>[1344, 1450]</td>
</tr>
<tr>
<td>$t_{0.1}$</td>
<td>1725</td>
<td>[1655, 1807]</td>
</tr>
</tbody>
</table>

Figure 7 is a Weibull probability plot of the idML estimates for a series of quantiles, together with a set of 90% pointwise confidence intervals, obtained from the likelihood-based procedure. Note that the LR confidence intervals can be used to obtain a confidence interval on either the proportion of failing at a particular time (draw a vertical line on Figure 7) or on a quantile at a particular probability (draw a horizontal line on Figure 7). This equivalence is demonstrated by Hong et al. (2008).
Figure 6: Fréchet distribution profile likelihoods for $\beta$ and $t_{0.1}$ for cylinder data. The solid vertical lines and dashed vertical lines are corresponding to LR intervals and Wald intervals, respectively.

Figure 7: Fréchet probability plot of the idML estimates and the 90% LR confidence bands for cylinder data.

6 Application to the Automobile-Component Data

6.1 Data description

This section describes the analysis of the automobile-component data. All times were linearly scaled to preserve confidentiality. For this application there are 144,102 vehicles (systems) in the fleet, all of which were censored at 50.24 time units. Each vehicle contains two identical
components (i.e., there are $m = 2$ sockets). The data are summarized in Table 6. There were 2,174 (1.51%) vehicles that had at least one replacement and 141,928 vehicles had no replacements. There was a total of 2397 replacements in the fleet. Therefore the fraction of sockets with events is $2,174/(144,102 \times 2) = 0.0075$.

Table 6: Distribution of number of replacements at each time

<table>
<thead>
<tr>
<th>Time</th>
<th>Number of replacements</th>
<th>Number of right-censored observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.14</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>6.28</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>9.42</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>12.56</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>15.70</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td>18.84</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>21.98</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>25.12</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>28.26</td>
<td>129</td>
<td>0</td>
</tr>
<tr>
<td>31.40</td>
<td>128</td>
<td>0</td>
</tr>
<tr>
<td>34.54</td>
<td>167</td>
<td>0</td>
</tr>
<tr>
<td>37.68</td>
<td>229</td>
<td>0</td>
</tr>
<tr>
<td>40.82</td>
<td>281</td>
<td>1</td>
</tr>
<tr>
<td>43.96</td>
<td>267</td>
<td>1</td>
</tr>
<tr>
<td>47.10</td>
<td>293</td>
<td>1</td>
</tr>
<tr>
<td>50.24</td>
<td>354</td>
<td>144099</td>
</tr>
</tbody>
</table>

6.2 Model fitting

The Weibull, Fréchet, Lognormal and Loglogistic distributions were fitted to the automobile-component data. The results in Table 7 show that the Weibull gives the largest likelihood among the four distributions. In Figure 8, the empirical MCF (solid curve) agree well with the Weibull fitted MCF (dashed curve). This suggests that the Weibull distribution provides a good choice to describe the automobile-component data.

Table 7: Log likelihoods for the automobile-component data for all the four log-location-scale distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>−21206.6</td>
</tr>
<tr>
<td>Loglogistic</td>
<td>−21207.8</td>
</tr>
<tr>
<td>Lognormal</td>
<td>−21299.3</td>
</tr>
<tr>
<td>Fréchet</td>
<td>−21393.1</td>
</tr>
</tbody>
</table>
6.3 Interval estimation

The point estimates and the 90% LR and Wald confidence intervals for the Weibull shape parameter $\beta$ and several quantiles of interest, $t_{0.001}$, $t_{0.005}$, $t_{0.01}$ and $t_{0.05}$, are shown in Table 8. The Weibull probability plot of the idML estimates for a set of quantiles, with a set of 90% pointwise likelihood-based confidence intervals is given in Figure 9. Because of the large number of vehicles in the fleet, there is almost no statistical error in the estimates.

Table 8: Point estimates and LR/Wald intervals for parameters of interest of automobile component data

<table>
<thead>
<tr>
<th>Quantity of interest</th>
<th>idML</th>
<th>90% confidence intervals LR</th>
<th>90% confidence intervals Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>2.08</td>
<td>[2.01, 2.15]</td>
<td>[2.01, 2.15]</td>
</tr>
<tr>
<td>$t_{0.001}$</td>
<td>18.11</td>
<td>[17.42, 18.80]</td>
<td>[17.42, 18.80]</td>
</tr>
<tr>
<td>$t_{0.005}$</td>
<td>39.33</td>
<td>[38.62, 40.05]</td>
<td>[38.61, 40.04]</td>
</tr>
<tr>
<td>$t_{0.01}$</td>
<td>54.96</td>
<td>[54.08, 55.89]</td>
<td>[54.06, 55.87]</td>
</tr>
<tr>
<td>$t_{0.05}$</td>
<td>120.41</td>
<td>[116.50, 124.67]</td>
<td>[116.33, 124.48]</td>
</tr>
</tbody>
</table>
Figure 9: Weibull probability plot of the idML estimates and the 90% LR confidence bands for automobile component data.

7 Computation Time for the Likelihood of a Single System

Evaluation of the SRP likelihood will be computationally intensive for large values of $r$. Figure 10 describes the amount of computer time required to compute the likelihood for a single system if we vary the number of events $r$ within one system from 1 to 20. The computations were done using a C++ program running on a Linux computer using a 3.07 GHz processor. Different curves correspond to different values of $m$ (numbers of sockets) in the system. Note that there is very little difference between the curves for $m = 8$ to 64. The plot shows that the computation time goes up exponentially with the number of events $r$. The idML procedure provides instantaneous estimation when the number of events $r$ within an SRP is less than 8. The practical limit on the number of events within a SRP that the idML estimation can handle depends on the number of sockets $m$ in the SRP. If there are only $m = 2$ sockets in the SRP, the idML procedure is able to provide an estimation for up to $r = 20$ events in the SRP (takes about one minute for computing the likelihood once). This is because only a few of the partitions for $r = 20$ are possible to occur when $m = 2$. For a large number of sockets $m$, especially when $m \geq r$ and all partitions of $r$ are feasible, the computation of the idML procedure is much slower. For example, it takes about a half hour to compute the likelihood once when the number of events $r = 15$ and the number of sockets is $m = 16$ or more.

Because the log likelihood for the whole fleet is the sum of all system log likelihoods, the computation time goes up linearly with the number of systems $n$. It is for this reason that in our simulation, we kept the number of events in each system fixed, and increased the number of systems to obtain larger number of events in the fleet.
Figure 10: Computation time required to compute the likelihood for a single system, when the number of events $r$ goes from 1 to 20.

8 Simulation Study

This section describes a simulation study to study the performance of the idML estimator, and to compare it with the complete-data ML (cdML) estimator (i.e., data when socket information is available). The simulation results are summarized briefly in this section (see the supplementary materials for more extensive results). In this paper, we mainly use the Weibull distribution to illustrate the simulation results. The simulations based on the other log-location-scale family distributions give similar results.

8.1 Design of the simulation study

8.1.1 Censoring schemes and experimental factors

The simulation is designed to mimic the replacement history for a fleet of $n$ systems (SRPs). For the simulation we used a Weibull component lifetime distribution with scale parameter $\eta = 1$. For simplicity, we assume that all systems have the same number of $m$ sockets.

In this study, we consider two censoring methods:

1. Failure censoring: the number of events before the end-of-observation time is specified/fixed and the end-of-observation time is random.
2. Time censoring: the end-of-observation time is specified/fixed and the number of events before the end-of-observation time is random.

For failure censoring, the end-of-observation time $\tau_c = \tau_r$, while for time censoring $\tau_c > \tau_r$. For time censoring, the number of events within each SRP is random and could be relatively large for some SRPs, which would significantly slow down computations. To save computing time, we used failure censoring for the main simulation and time censoring for a smaller simulation to compare the results.

The factors used in the failure-censoring (time-censoring) simulation experiments were:

- The number of sockets in each system $m$.
- The number $r$ (expected number $E(R)$) of events per system.
- The number $n \times r$ (expected number $n \times E(R)$) of events for the fleet.
- The Weibull shape parameter $\beta$.

### 8.1.2 Factor levels

We conducted simulations at all combinations of the following levels of the factors.

- $m$: 2, 4, 8, 16, 32
- $r$: 2, 4, 8
- $n \times r$: 8, 16, 32, 64, 128, 256
- $\beta$: 1, 3

Note that the number of systems $n$ is controlled according to the number of events $r$ and the number of events for the fleet $n \times r$ (or expected number of events $n \times E(R)$) when studying asymptotic behavior of estimation properties.

For each combination of the factor levels we simulated and computed the idML and cdML estimates for $B = 5,000$ data sets. The quantities of interest include the Weibull shape parameter $\beta$ and various quantiles $t_p$ of the Weibull distribution. The following section will summarize the most relevant and interesting results from the simulation experiments.

### 8.2 Simulation Results

#### 8.2.1 Variance and bias of the idML estimator

We use the cdML estimator as the reference because with all information available, this estimator is expected to perform well. The difference between the performances of the idML and cdML
estimators indicates how much information is lost if we do not know the socket identity for the component replacement events.

Figure 11 provides summaries of the idML and cdML estimates for 5,000 simulated data sets, corresponding to two factor level combinations: Case 1 (left) and Case 2 (right). The solid line in the center corresponds to the true Weibull distribution ($\eta = 1$ and $\beta = 3$), while the dashed line and dotted line in the center show the median of the 5,000 idML and cdML estimates respectively, for a set of quantiles ranging from 0.0001 to 0.99. These two lines describe the central tendency of the distributions for the idML and cdML estimators. The upper and lower pairs of curves are the 0.05 and 0.95 quantiles respectively, of the idML (dashed curves) and cdML (dotted curves) estimates of the Weibull cdf.

As seen in Figure 11 for Case 1 where $n = 8$, $m = 4$, and $r = 8$, the idML estimator has a slightly more median bias and somewhat more variability than the cdML estimator for estimating smaller probabilities. For Case 2 where the number of sockets in each system increases to $m = 32$, the performance of the idML estimator is similar to that of the cdML estimator in the sense that both of them have small bias and similar amount of variance. Note that the total number of events for the whole fleet is $n \times r = 8 \times 8 = 64$ for both Case 1 and Case 2, which explains the similar performance of the cdML estimates in the two cases. The reason that the different behaviors of the idML estimator in the two cases is the number of sockets $m$, with $m = 4$ for Case 1 and $m = 32$ for Case 2. Then the average number of events per socket are $r/m = 8/4 = 2$ and $r/m = 8/32 = 0.25$ for Case 1 and Case 2, respectively. Therefore it is most likely that each socket has zero or at most one event in the simulated SRPs in Case 2, and not much information is lost due to the lack of socket identity information.

![Graph](image)

Figure 11: The median, 0.05 quantile and 0.95 quantile of both the idML and cdML estimates of quantiles ranging from 0.0001 and 0.99 with $m = 4$ on the left and $m = 32$ on the right.

### 8.2.2 Relative efficiency of the idML estimator

In this section, we study the asymptotic behavior of the relative efficiency (RE) of the idML estimator and what factors affect the RE. The RE of the idML estimator relative to the cdML estimator is computed as the ratio of the respective MSE estimates, and quantifies how much information is lost if we do not know the socket identity of the event times. The simulation results
show that the RE of idML estimator gradually converges to a limiting value as the number of events in the fleet increases. Regarding to the effect of $\beta$, the idML estimator tend to have higher RE for large values of $\beta$. In addition, when the number of socket $m$ is fixed, the idML estimator tends to have a higher RE when there are fewer events within one system. While keeping the number of events $r$ within each system fixed, the idML estimator tends to have higher RE when there is a larger number of sockets $m$ in one SRP.

To better understand when the idML estimators will be relatively efficient, we conducted a small simulation study to obtain insight (see supplementary materials). In some situations, the data provide good information about the particular partition (or a small number of partitions) that were likely to have led to the observed data for a system. Situations leading to such data will have idML estimators with relatively high efficiency. For example if a system has 32 sockets with a Weibull component failure-time distribution and typical systems have only a small number of events (say fewer than 4), it is unlikely that there would be more than 2 events in one slot, especially if the Weibull shape parameter is large (say 3 or more). If, on the other hand, the expected number of events in a system is greater than or equal to the number of sockets in the system, there is little information in the data about the which particular partitions might have led to the particular SRP outcomes.

8.2.3 Confidence interval performance

This section compares confidence interval procedures based on inverting the likelihood ratio test and the Wald approximation, as described in Section 5.4. The simulation results show that the coverage probability for the upper bound tends to be greater than the nominal coverage, while the coverage probability for the lower bound tends to be lower than the nominal. The coverage probabilities for both upper and lower bound gradually converge to certain limiting values when there are relatively large number of events in the fleet. The Wald interval procedure tends to converge at a slower rate than the LR procedure. Similar to the results for the RE of idML estimator, both the Wald and the LR intervals have better coverage probability for smaller values of $r/m$ (i.e., smaller fraction of sockets with events). Interestingly, the Wald intervals always provide good approximation to the nominal coverage in estimating the Weibull shape parameter $\beta$ comparing to the LR interval. For estimating distribution quantiles, however, the LR confidence intervals have a better performance than the Wald-based intervals, and generally provide a good coverage probability if the total number of events in the fleet is not too small.

9 Concluding Remarks and Areas for Future Research

In this paper we proposed a likelihood-based procedure for estimating the lifetime distribution of a component from the aggregated event data for a fleet with multiple systems. This idML estimation method performs well especially when the number of events for each SRP is relatively small and the number of systems is sufficiently large.

Some possible areas for future research are:

1. In this paper, we applied the likelihood-based idML estimation to log-location-scale family
distributions. It is, however, easy to extend this procedure to other parametric distributions.

2. Our method requires the assumption that the component failure-time distribution is the same in all sockets across all systems. In some applications where systems are operated in different environments, covariate adjustment would be required to obtain meaningful results. It would be straightforward to extend the ML method given here to a model that allows for system-level covariate information for either static covariates (e.g., Hong and Meeker, 2010) or dynamic (time-varying) covariates (e.g., Hong and Meeker, 2013).

3. In order to assess the distributional goodness of fit of the proposed idML estimation, it would be useful to develop some nonparametric methods to estimate the cdf of the component lifetime distribution and quantify its statistical uncertainty to the use of probability plots, as is commonly done with usual censored data.

4. An approximate likelihood computation could be developed that uses only high-probability partitions, allowing for more events within one system. A reasonable range of $\beta$ values can be used to compute the partition probabilities. Then the partitions with high probabilities (e.g., larger than some specified threshold) can be identified and used to compute the likelihood. This could make the likelihood evaluation more efficient.

5. Based on the computationally intensive nature of the idML estimation method, it would be useful to apply certain highly parallel computing (e.g., GPU) for likelihood computations.

6. For situation where there is a large number of sockets and/or systems, it would be appropriate to apply an estimation procedure based on a non-homogeneous Poisson process approximation to the SRP.

Acknowledgement

We would like to thank the referees, associate editor, Editor Peihua Qiu, Dan Nordman and Vivekananda Roy for providing comments that helped us make important improvements to the paper. C codes from Robin K. S. Hankin’s R package “Partitions” were helpful for the implementation idML estimation procedure.

References


