A Parametric Empirical Bayesian Software Reliability Model

DAMODARAN DURAI SWAMY1* and GOPAL GOVINDASAMY2

1 Centre for Reliability, STQC Directorate, Ministry of Communication & Information Technology, Government of India, Dr.VSI Estate, Thiruvanmiyur, Chennai-600041
2 Department of Statistics, University of Madras, Chepauk, Chennai-600005.

(Received on April 19, 2008; Final Revision on November 10, 2008)

Abstract: In this paper, a new parametric empirical Bayesian software reliability model is presented. Times between failures follow generalised exponential distribution with stochastically decreasing order on the failure rate functions of successive failure time intervals with the software tester’s intention to improve the software quality by the correction of each failure. With the Bayesian approach, the predictive distribution has been arrived at by combining generalised exponential time between failures and gamma prior distribution for the parameter namely failure rate. The expected time between failure measure, reliability function etc. have been obtained. The posterior distribution of the failure rate measure has been deduced and the mean failure rate is also obtained. For the parameter estimation, least square estimation method has been adopted. The proposed model has been applied to three sets of actual software failure data. It has been observed that the predicted failure times as per the proposed model are closer to the actual failure times. Sum of square errors criteria has been used for comparing the actual time between failure times and predicted time between failures.

Keywords: Software reliability, failure-rate, parameter and hyper-parameter, predictive & posterior distributions, least square estimate.

1. Introduction

Software reliability is defined as the probability of failure-free operation for a computer program in a specified environment for a specified period of time [15]. Over the past 30 years, active research have been carried out to study software reliability engineering, and many analytical models [2],[5],[7] have been proposed for software reliability assessment. Development of these models is based mainly on the failure history of the software and is classified according to the nature of the failure process. One of the main classes of models is times between failures. Over the last three decades the most common approach is that time between failures follows a known statistical distribution [3] and the parameters of which depend on the number of faults remaining in software during this interval. Within the non-Bayes framework, these parameters are thought to be unknown, but fixed quantities. A random sample of time between failures is collected. Based on these observed values, parameters of interest are estimated by Maximum Likelihood Estimation (MLE) or Least Square Estimation (LSE) methodologies. Parameter estimation in the existing Markov and Non homogeneous Poisson process (NHPP) models is not an easy task and sometimes these estimates do not provide adequate results [15]. The experiences from similar software assignments and previous information about the software development projects which can be combined with the available failure data in order to make more accurate estimation and
prediction. In Bayesian approach, this prior knowledge is being used effectively whereas in classical estimation technique this information cannot be utilised in a systematic manner.

In recent times, another main concern in the software industry is the limited budget and planned time for software development. Delay in introduction of a software product into the market or even overspending on the stringent budget results in financial crisis for the company. Increasing insistence on cost-effectiveness and limited time allocated for testing make it very difficult to obtain adequate sample of failures. Under the classical approach, such a small sample size would result in imprecise reliability estimates [15]. Furthermore, for safety-critical applications, it is seldom that we encounter software failure. However, its consequence is very disastrous such as loss of life. So, developers are faced with the pressure of obtaining accurate estimates based on scarce data. Attempts have been made in using Bayesian techniques to deal with these problems. Bayesian approach [8],[12] has proven to be an effective methodology since it allows the incorporation of prior information such as expert knowledge, historical data, etc. into the model, thus improving its prediction on software reliability while reducing testing time or sample size requirement. Many Bayesian models [1],[6],[10] have been proposed for the analysis of software failure data combined with previous knowledge in the form of so called prior distribution of the unknown parameter. For any model of the software failure process, there are number of parameters to be estimated using collected failure data. Usually, some information about these parameters is available and such information can be described as prior distribution of the parameters. The prior distributions contain some parameters and these are called as hyper-parameters. All the Bayesian models mentioned above [1],[6],[10] assume that the Time between Failures (TBF) follow exponential distribution. The time between failures of software system in several occasions may follow different patterns other than exponential distribution [9]. Taking this into consideration, a new parametric empirical Bayesian software reliability model has been presented and the predicted TBFs based on this model seem to be closer to the actual failure times.

This paper is organized as under. In section 1, we have given the notation and model assumptions. In section 2, we discuss about the proposed parametric empirical Bayesian software reliability model and discuss about the various reliability characteristics and other measures. The proposed model has been validated using three actual software reliability datasets and comparative graphs and tables are presented in section 3. The methodology for estimation of parameters is given in section 4. In section 5, conclusions of the paper are given.

**Notation**

- $T_i$ random variable (r.v.) which represents the actual time between failures $(i-1)^{th}$ and $i^{th}$ software failure
- $t_i$ realization of the r.v. $T_i$
- $\lambda_i$ failure rate (parameter) at $i^{th}$ failure time.
- $g(\lambda_i / \alpha, \psi(i))$ gamma prior distribution of the parameter $\lambda_i$
- $\alpha, \psi(i)$ hyper-parameters
- $\pi(\lambda_i / \alpha, \psi(i))$ posterior distribution of the parameter $\lambda_i$
Model Assumptions:

1) The time interval after repair of \((i-1)\)th and the occurrence of \(i\)th failure is a random variable following a Generalised Exponential (GE) distribution \([4]\):

\[
f(t_i / \lambda_i) = b \lambda_i e^{-\lambda_i t_i} (1 - e^{-\lambda_i t_i})^{b-1}; \quad \lambda_i > 0, t_i > 0
\]

where \(b\) is shape parameter and \(\lambda_i\) is scale parameter.

2) Upon observing a failure, faults are immediately removed within negligible time.

2. The Proposed Bayesian Software Reliability Model

The time between failures (TBF) of software System may not follow exponential distribution always. In several occasions TBFs may have different patterns and in order to handle these patterns, a new Bayesian software reliability model has been proposed in this paper by considering GE distribution \([14]\) which is more flexible. The GE pdf as given in equation (1) involves two parameters namely \(b\) and \(\lambda_i\). In this paper, for the simplicity and for the purpose of single parameter distribution, the GE distribution with shape parameter \(b=2\) has been considered. Thus, we have the GE pdf as

\[
f(t_i / \lambda_i) = 2\lambda_i e^{-\lambda_i t_i} (1 - e^{-\lambda_i t_i}); \quad \lambda_i > 0, t_i > 0
\]

where \(\lambda_i\) is the parameter namely failure rate at the \(i\)th failure and assume that \(\lambda_i \leq \lambda_{i-1}\) and it has stochastic nature as follows:

\[
P(\lambda_{i-1} \leq \lambda) \geq P(\lambda_i \leq \lambda) \quad \text{for all } i \geq 1
\]

Let us assume the parameter \(\lambda_i\) (failure rate) follows gamma prior distribution with the following pdf:

\[
g(\lambda_i / \alpha, \psi(i)) = \frac{[\psi(i)]^\alpha \lambda_i^{\alpha-1} \exp(-\psi(i)\lambda_i)}{\Gamma(\alpha)}; \quad \alpha > 0, \psi(i) > 0
\]

where \(\alpha\) is shape parameter and \(\psi(i)\) is the scale parameter depending on the number of detected faults. \(\psi(i)\) may take different forms namely, \(\beta_0 + \beta_1 i (\text{Linear})\), \(\beta_0 + \beta_1 i^2 (\text{Quadratic})\) and \(\exp(\beta_0 + \beta_1 i) (\text{Exponential})\). For the different forms of \(\psi(i)\), various software test environments may be described. Then the predictive distribution of \(t_i\) for the proposed Bayesian model can be derived as follows:

\[
f(t_i / \alpha, \psi(i)) = \int_0^\infty \left[2\lambda_i e^{-\lambda_i t_i} (1 - e^{-\lambda_i t_i})\right]^{\alpha-1} \lambda_i^{-\alpha} \exp(-\psi(i)\lambda_i) \frac{1}{\Gamma(\alpha)}
\]

\[
= \left[\frac{2\alpha\psi(i)^\alpha}{[t_i + \psi(i)]^{\alpha+1}} - \frac{2\alpha\psi(i)^\alpha}{[2t_i + \psi(i)]^{\alpha+1}}\right]
\]

Then the Mean Time To Failure (MTTF) can be obtained as follows:

\[
E[T_i] = \int_0^\infty t_i \left[\frac{2\alpha\psi(i)^\alpha}{[t_i + \psi(i)]^{\alpha+1}} - \frac{2\alpha\psi(i)^\alpha}{[2t_i + \psi(i)]^{\alpha+1}}\right] dt_i = \frac{3\psi(i)}{2(\alpha-1)}
\]
The cumulative distribution function of the predictive distribution of \( t_i \) is given by,

\[
F(t_i \mid \alpha, \psi(i)) = 1 + \frac{\psi(i)^\alpha}{[2t_i + \psi(i)]^\alpha} - \frac{2\psi(i)^\alpha}{[2t_i + \psi(i)]^{1+\alpha}}
\]

The reliability function of the predictive distribution is,

\[
R(t_i \mid \alpha, \psi(i)) = 1 - F(t_i \mid \alpha, \psi(i)) = \frac{2\psi(i)^\alpha}{[t_i + \psi(i)]^\alpha} - \frac{\psi(i)^\alpha}{[2t_i + \psi(i)]^\alpha}
\]

The failure rate function of the predictive distribution can be derived as follows:

\[
f(t_i \mid \lambda_i) = \frac{2\alpha \psi(i)^\alpha}{[t_i + \psi(i)]^{\alpha+1}} - \frac{2\alpha \psi(i)^\alpha}{[2t_i + \psi(i)]^{\alpha+1}}
\]

The posterior distribution of the parameter \( \lambda_i \) is as follows:

\[
\pi(\lambda_i \mid \alpha, \psi(i)) \propto \left[ \prod_{i=1}^{n} f(t_i \mid \lambda_i) \right] g(\lambda_i \mid \alpha, \psi(i))
\]

3. Model validation of the proposed model with various actual software reliability data sets:

The proposed model has been validated using three actual software reliability datasets viz. NTDS & AT&T [12] and RAC [13]. Actual and predicted failure times of these three data sets are compared.
Table 1: Comparison of predictive power of the proposed model with actual failure times for AT&T data set [12]

<table>
<thead>
<tr>
<th>Failure number</th>
<th>Actual cumulative inter failure times in CPU units</th>
<th>Actual time between failure times in CPU units</th>
<th>Proposed model predictive cumulative time between Failures</th>
<th>Proposed model predictive time between failures</th>
<th>Error sum of squares (ESS) for the proposed model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.5</td>
<td>5.5</td>
<td>13.07365518</td>
<td>13.07365518</td>
<td>57.36025271</td>
</tr>
<tr>
<td>2</td>
<td>7.33</td>
<td>1.83</td>
<td>16.93317553</td>
<td>3.85920535</td>
<td>4.118952851</td>
</tr>
<tr>
<td>3</td>
<td>10.08</td>
<td>2.75</td>
<td>23.36570944</td>
<td>6.432533917</td>
<td>13.56105605</td>
</tr>
<tr>
<td>4</td>
<td>80.97</td>
<td>70.89</td>
<td>32.37125693</td>
<td>9.00547483</td>
<td>3829.685463</td>
</tr>
<tr>
<td>5</td>
<td>84.91</td>
<td>3.94</td>
<td>43.94981798</td>
<td>11.57856105</td>
<td>58.34761491</td>
</tr>
<tr>
<td>6</td>
<td>99.89</td>
<td>14.98</td>
<td>58.10139259</td>
<td>14.15157462</td>
<td>0.686288616</td>
</tr>
<tr>
<td>7</td>
<td>103.36</td>
<td>3.47</td>
<td>74.82598078</td>
<td>16.72458818</td>
<td>175.6841079</td>
</tr>
<tr>
<td>8</td>
<td>113.32</td>
<td>9.96</td>
<td>94.12358253</td>
<td>19.29760175</td>
<td>87.19080644</td>
</tr>
<tr>
<td>10</td>
<td>144.59</td>
<td>19.88</td>
<td>140.4378267</td>
<td>24.44362888</td>
<td>20.82670858</td>
</tr>
<tr>
<td>11</td>
<td>152.4</td>
<td>7.81</td>
<td>167.4544692</td>
<td>27.01664245</td>
<td>368.8951142</td>
</tr>
<tr>
<td>12</td>
<td>166.99</td>
<td>14.59</td>
<td>197.0441252</td>
<td>29.58965602</td>
<td>224.9896806</td>
</tr>
<tr>
<td>13</td>
<td>178.41</td>
<td>11.42</td>
<td>229.2067948</td>
<td>32.16266958</td>
<td>430.2583414</td>
</tr>
<tr>
<td>14</td>
<td>197.35</td>
<td>18.94</td>
<td>263.9424779</td>
<td>34.73568315</td>
<td>249.5036062</td>
</tr>
<tr>
<td>15</td>
<td>262.65</td>
<td>65.3</td>
<td>301.2511746</td>
<td>37.30869672</td>
<td>783.5130595</td>
</tr>
<tr>
<td>16</td>
<td>262.69</td>
<td>0.04</td>
<td>341.1328849</td>
<td>39.88171028</td>
<td>1587.361878</td>
</tr>
<tr>
<td>17</td>
<td>388.36</td>
<td>125.67</td>
<td>383.5876088</td>
<td>42.45472385</td>
<td>6924.782185</td>
</tr>
<tr>
<td>18</td>
<td>471.05</td>
<td>82.69</td>
<td>428.6153462</td>
<td>45.02773742</td>
<td>1418.466023</td>
</tr>
<tr>
<td>19</td>
<td>471.5</td>
<td>0.45</td>
<td>476.2160972</td>
<td>47.60070908</td>
<td>2223.193318</td>
</tr>
<tr>
<td>20</td>
<td>503.11</td>
<td>31.61</td>
<td>526.3898617</td>
<td>50.17376455</td>
<td>344.6133543</td>
</tr>
<tr>
<td>21</td>
<td>632.42</td>
<td>129.31</td>
<td>579.1366398</td>
<td>52.74677812</td>
<td>5861.926945</td>
</tr>
<tr>
<td>22</td>
<td>680.02</td>
<td>47.6</td>
<td>634.4564315</td>
<td>55.31979168</td>
<td>59.59518363</td>
</tr>
</tbody>
</table>

Total 24834.38324

Fig. 1: Comparison of proposed model with actual failure times – AT&T Data Set
4. Parameter Estimation for the proposed model

The parameters of the proposed model have been obtained by the Method of Least Squares Estimation by minimizing the following equation:

\[ S(\alpha, \beta_0, \beta_1) = \sum_{i=1}^{n} (t_i - E[T_i])^2 = \sum_{i=1}^{n} \left( t_i - \left[ \frac{3\psi}{2(\alpha - 1)} \right] \right)^2 \]

From equation (6), the above expression reduces to the following:

\[ = \sum_{i=1}^{n} \left( t_i - \left[ \frac{3(\beta_0 + \beta_1 t_i^2)}{2(\alpha - 1)} \right] \right)^2 \] (14)

The estimation of the hyper-parameters \( \alpha, \beta_0, \beta_1 \) of the above equation, have been obtained by using the NCSS [11] software and the respective values of the \( \alpha, \beta_0, \beta_1 \) for the different actual software failure data sets have been obtained. The various measures viz. the Mean Time To Failure (MTTF) [7], Mean failure rate of the posterior distribution of the parameter etc. have been obtained for the various actual software failure data sets.

Mean failure rate of the posterior distribution (\( E[\hat{\lambda}_i] \)) of the parameter \( \hat{\lambda}_i \) is:

\[ = 2^n \alpha^n (\alpha + 1)^n \left[ \prod_{i=1}^{n} \left[ \frac{[\psi(i)]^\alpha}{[t_i + \psi(i)]^{\alpha+2}} \right] - \prod_{i=1}^{n} \left[ \frac{[\psi(i)]^\alpha}{[2t_i + \psi(i)]^{\alpha+2}} \right] \] \] (15)

The value of one of the hyper parameters \( \alpha \) is assumed to be fixed and the other hyper parameter \( \psi(i) \) takes different forms of linear, quadratic and exponential. All the three
forms have been verified with three data sets for fitting. The quadratic form is found to be fitting better compared to other forms. Hence the MTTF and mean failure rate measure of the posterior distribution have been calculated for the quadratic form of $\psi(i)$.

The reliability characteristics for three software failure datasets are furnished in the following table.

**Table 2: Reliability characteristics for three software failure data sets**

<table>
<thead>
<tr>
<th>Software failure Data Set &amp; Estimated values of the hyper parameters $\alpha, \beta_0, \beta_1$</th>
<th>Mean Time To Failure (MTTF)</th>
<th>Mean failure rate of the posterior distribution of $\lambda_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T Data: $\alpha=1.813, \beta_0=5.04, \beta_1 = 0.523$</td>
<td>At $n^{th}$ failure time 634.35, At $(n+1)^{th}$ failure time 692.35</td>
<td>At $n^{th}$ failure time 1.181E-100, At $(n+1)^{th}$ failure time 1.12E-106</td>
</tr>
<tr>
<td>NTDS Data: $\alpha=1.49, \beta_0=6.53, \beta_1 = 0.072$</td>
<td>At $n^{th}$ failure time 224.57, At $(n+1)^{th}$ failure time 240.16</td>
<td>At $n^{th}$ failure time 1.3819E-58, At $(n+1)^{th}$ failure time 2.9442E-61</td>
</tr>
<tr>
<td>RAC Data: $\alpha=2.64, \beta_0=18.095, \beta_1 =1.01$</td>
<td>At $n^{th}$ failure time 298.13, At $(n+1)^{th}$ failure time 336.21</td>
<td>At $n^{th}$ failure time 5.146E-51, At $(n+1)^{th}$ failure time 2.692E-55</td>
</tr>
</tbody>
</table>

5. Conclusions

The generalized exponential distribution failure time models are well suited for the software reliability assessment. With Bayesian approach, a predictive distribution has been arrived at by combining generalised exponential failure time along with Gamma prior distribution for the parameter namely failure-rate. It has been observed that the predicted failure times for the proposed model are seen to be closer to the actual failure times for the data sets. This model is simple and easy to adopt. And also even for the smaller sample observed failure times also this parametric empirical Bayesian model will give better results.

References


**Damodaran Duraiswamy** is a Senior Scientist in Centre for Reliability (CFR), Government of India, Chennai. He is a Certified Reliability Professional (CRP) and a Certified Software Test Manager (CSTM) by Standardization Testing and Quality Certification Directorate (STQC), Department of Information Technology (DIT), Government of India. He has been trained in Six Sigma Characterizations and Optimization by BEQI, Bangalore, India. He has an extensive experience in the area of Reliability Engineering. He specializes in Reliability Estimation, Life Data Analysis, Software Development and Software Reliability and Statistical Process Control techniques. He obtained his Master’s degree in Statistics from the University of Madras, Chennai, India and is currently pursuing research in the area of Software Reliability in the Department of Statistics, University of Madras, Chennai, India.

**Gopal Govindasamy** is Professor and Head of the Department of Statistics, University of Madras, Chennai. His current areas of research are Majorization, Reliability and Survival analysis. He has to his credit 22 research papers in National/International journals. He has guided so far 6 Ph.D. scholars. He has also introduced Actuarial Science in the postgraduate program in the University of Madras. Institute of Actuaries of India has conferred him with the title Hon. Fellow of Institute of Actuaries of India (Hon.FIAI) for promoting actuarial education in India. He obtained his Master’s degree and Ph.D. in Statistics from the University of Madras, Chennai, India.