Semi-Parametric Approach for Software Reliability Evaluation Using Mixed Gamma Distributions

Hiroyuki OKAMURA, Takumi HIRATA and Tadashi DOHI
Department of Information Engineering, Graduate School of Engineering
Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, 739-8527 Japan
dohi@rel.hiroshima-u.ac.jp

Abstract

This paper proposes a semi-parametric software reliability model (SRM) based on a mixed gamma distribution, so-called the mixed gamma SRM. In addition, we develop the parameter estimation method for the mixed gamma SRM. Concretely, the estimation method is based on the Bayesian estimation and the parameter estimation algorithm is described by MCMC (Markov chain Monte Carlo) method with grouped data. In numerical experiment, we compare the mixed gamma SRM with several parametric SRMs, and examine the performance on the fitting ability of proposed model.

Keywords: Software reliability, Semi-parametric approach, Mixed gamma model, Bayesian estimation, MCMC

1 Introduction

Software reliability is a significant attribute of software quality. The quantitative software reliability is defined as the probability that no failure occurs during a certain time period. Thus probabilistic models are needed to evaluate quantitative software reliability from field data (failure data). In fact, a vast amount of software reliability models (SRMs) have been proposed and developed from various points of view during the last four decades. Specifically, non-homogeneous Poisson process (NHPP) based SRMs have played a central role to estimate the number of remaining faults as well as the quantitative software reliability [1, 2].

In general, there are two categories of SRMs: parametric and non-parametric SRMs. When evaluating the software reliability using parametric NHPP-based SRMs, we first decide a set of candidates used in the software reliability estimation. The statistical parameter estimation is executed for all the candidates to determine their model parameters fitting to observed software failure data. After estimating model parameters, we choose the best model in the sense of information criteria such as AIC (Akaike information criterion). However, it is not easy to decide a good set of candidates that include the best model describing the software failure occurrence process.

In the non-parametric approach, we do not need to decide the candidates before estimating model parameters. For example, kernel density estimation, which is a typical non-parametric approach, defines a kernel function for each data point. Without any density function, it gives an estimated density function based on the set of kernel functions.
The simplest example of non-parametric estimation is the empirical distribution. Also, neural network models are also categorized to one of the non-parametric methods. Several papers discussed the applicability of non-parametric approaches to the software reliability evaluation. Generally, the non-parametric estimation gives highly accurate estimators in the case where many samples are observed. However, we rarely observe many failure time or count data in the software reliability evaluation. In addition, the drawback of non-parametric approach is not to estimate the density for truncated area. In general, software failure data behave similar to truncated data, and thus non-parametric approach does not work well.

In this paper, we consider semi-parametric approach for the quantitative software reliability. Semi-parametric models, which are the intermediate modeling framework between parametric and non-parametric models, can represent every stochastic behavior of data by systematically augmentation of the number of parameters. Unlike non-parametric models, semi-parametric models can provide the prediction of future behavior of software failure occurrence.

This paper proposes a semi-parametric SRM based on mixed gamma distribution, so-called mixed gamma SRM. For this semi-parametric model, we consider the Bayesian estimation. In general, the Bayesian estimation is commonly used in the parameter estimation of semi-parametric models. Roughly speaking, the maximum likelihood (ML) estimation, which is frequently utilized in the software reliability evaluation, causes overfitting problem due to the fact that semi-parametric models consist of many model parameters to be estimated. Compared with ML estimation, it is well known that the Bayesian method avoids the overfitting problem. This paper first give the definition of mixed gamma SRM and develop Bayesian estimation for mixed gamma SRM using Markov chain Monte Carlo (MCMC) method.

2 Mixed Gamma SRM

The mixed gamma SRM is defined as the NHPP model where failure occurrence time follows a mixed gamma distribution. The c.d.f. and the p.d.f. of mixed gamma distribution are given by

\[ F(t; \pi, \alpha, \beta) = \int_0^t f(u; \pi, \alpha, \beta)du, \]  \hspace{1cm} (1) \\
\[ f(t; \pi, \alpha, \beta) = \sum_{k=1}^{m} \pi_k \mathcal{G}(t; \alpha_k, \beta_k), \] \hspace{1cm} (2)

where

\[ \pi = \{\pi_1, ..., \pi_m\}, \quad \sum_{k=1}^{m} \pi_k = 1, \quad \pi_k \geq 0, \]  \hspace{1cm} (3) \\
\[ \alpha = \{\alpha_1, ..., \alpha_m\}, \quad \beta = \{\beta_1, ..., \beta_m\}. \]  \hspace{1cm} (4)

According to the order statistics model [3], we get the following probability mass function of the cumulative number of software failures \( N(t) \) before time \( t \):

\[ P(N(t) = k) = \frac{(\omega F(t; \pi, \alpha, \beta))^k}{k!} \exp \left( -\omega F(t; \pi, \alpha, \beta) \right). \]  \hspace{1cm} (5)
Obviously, the above probability mass function corresponds to the NHPP probability mass function with the mean value function \( \omega F(t; \pi, \alpha, \beta) \), where \( \omega \) indicates the average number of total failures during the software life cycle.

3 Bayesian Computation

As mentioned before, ML estimation for semi-parametric models often causes the overfitting problem. Therefore, this paper considers Bayesian estimation for mixed gamma SRM.

Suppose that software failure occurrences are observed as grouped data. That is, the grouped data collect the number of failure occurrences as a bin. Let \( t_i \) and \( x_i \) be the cumulative testing time indicating breaks of bins and the number of failures observed in the \( i \)-th bin, where we assume \( t_0 = 0 \).

Consider the Bayesian estimation with the grouped data. Bayesian estimation is the well-established framework for parameter estimation based on prior information. The key idea behind the Bayesian estimation is to regard model parameters as random variables. Let \( p(\theta) \) and \( D \) denote the prior information about a parameter set \( \theta \) and observed data set, respectively. From Bayes theorem, the posterior information, i.e., the updated information after obtaining \( D \) is given by

\[
p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta},
\]

where \( p(D \mid \theta) \) is the likelihood of \( D \) on the fixed parameter set \( \theta \). Taking account of the normalizing condition of posterior distributions, Eq. (6) can be alternatively expressed without the normalizing constant \( \int p(D \mid \theta)p(\theta)d\theta \):

\[
p(\theta \mid D) \propto p(D \mid \theta)p(\theta).
\]

The computation of normalizing constant causes analytical or numerical integration over the domain of parameter set \( \theta \). Only for some specified cases, we can obtain the closed forms of normalizing constants; for example, when the conjugate prior distributions are applied, we have normalizing constants implicitly. In most situations, however, any numerical technique has to be utilized for evaluating posterior distributions.

The MCMC (Markov chain Monte Carlo) is a versatile method to evaluate posterior distributions in Bayesian estimation. The MCMC can be regarded as sampling-based approximation of posterior distributions. Except for the use of conjugate prior distribution, it is difficult to derive the closed forms of posterior distributions. Therefore, we have to apply any specific sampling method to obtain samples from target posterior distributions. The Gibbs sampling and the Metropolis-Hasting method are representative sampling methods in the MCMC.

In the Gibbs sampling, one generates the target joint posterior distribution based on conditional posterior distributions. Let \( p(\theta_1, \ldots, \theta_m \mid D) \) be the target joint posterior distribution of parameters \( \theta_1, \ldots, \theta_m \). When one can generate samples from the conditional posterior distribution \( p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m, D) \), the Gibbs sampler is given by the
following scheme:

\[
\hat{\theta}_1 \sim p(\theta_1 | \hat{\theta}_2, \ldots, \hat{\theta}_m, D), \\
\hat{\theta}_2 \sim p(\theta_2 | \hat{\theta}_1, \hat{\theta}_3, \ldots, \hat{\theta}_m, D), \\
\vdots \\
\hat{\theta}_i \sim p(\theta_i | \hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_m, D), \\
\vdots \\
\hat{\theta}_m \sim p(\theta_m | \hat{\theta}_1, \ldots, \hat{\theta}_{m-1}, D),
\]

where \( \hat{\theta} \sim p(\cdot) \) indicates that the sample \( \hat{\theta} \) is generated from the probability distribution \( p(\cdot) \). The above sampling can be regarded as Markov chain with state space \((\theta_1, \ldots, \theta_m)\).

In the Gibbs sampling, the stationary distribution of this Markov chain is exactly equal to the joint posterior distribution. Therefore, we obtain the samples from the joint posterior distribution by repeating the above sampling scheme.

To build an MCMC algorithm for the mixed gamma SRM, we define the following probability densities:

\[
D(\pi; \rho) = \frac{\Gamma(\sum_{k=1}^m \rho_k) \prod_{k=1}^m \pi_k^{\rho_k-1}}{\prod_{k=1}^m \Gamma(\rho_k)}, \\
\rho = (\rho_1, \ldots, \rho_m),
\]

where \( \pi_k \geq 0, \sum_{k=1}^m \pi_k = 1, \rho_k \geq 0 \).

\[
B(\mathbf{x}; \rho) = \begin{cases} 
\rho_1 & \text{if } \mathbf{x} = \{1, 0, \ldots, 0\}, \\
\rho_2 & \text{if } \mathbf{x} = \{0, 1, \ldots, 0\}, \\
\vdots & \\
\rho_m & \text{if } \mathbf{x} = \{0, 0, \ldots, 1\}, \\
0 & \text{otherwise}, 
\end{cases}
\]

\( \rho = (\rho_1, \ldots, \rho_m) \),

where \( \rho_k \geq 0, \sum_{k=1}^m \rho_k = 1 \). The prior distributions are given by

\[
p(\omega) = \mathcal{G}(\omega; a_\omega, b_\omega), \\
p(\pi) = \mathcal{D}(\pi; \rho), \\
p(\alpha_k) = \mathcal{G}(\alpha_k; a_{\alpha_k}, b_{\alpha_k}), \\
p(\beta_k) = \mathcal{G}(\beta_k; a_{\beta_k}, b_{\beta_k}).
\]

Let \( \mathbf{Z}_u \in \mathbf{Z} \) be a latent parameter vector, where \( Z_{u,k} \in \mathbf{Z}_u, k = 1, \ldots, m \), is an indicator random variable meaning the failure experienced at time \( u \) belongs to the \( k \)-th mixture component. If the failure experienced at time \( u \) belongs to \( k \)-th mixture component, \( Z_{u,k} = 1 \), otherwise, \( Z_{u,k} = 0 \). Consequently, the likelihood function of the mixed gamma SRM
can be written as follows.

\[
p(D, \tilde{N}, S, \bar{S}, Z | \omega, \pi, \alpha, \beta) = e^{-\omega} \prod_{i=1}^{n} \prod_{j=1}^{m} \omega \prod_{k=1}^{m} (\pi_k G(S_{i,j}; \alpha_k, \beta_k))^{Z_{S_{i,j},k}} \times \prod_{i=1}^{n} \prod_{j=1}^{m} (\pi_k G(t_{i,j}; \alpha_k, \beta_k))^{Z_{t_{i,j},k}} \times \prod_{j=1}^{\tilde{N}} \prod_{k=1}^{m} (\pi_k G(S_{n+1,j}; \alpha_k, \beta_k))^{Z_{S_{n+1,j},k}}, \quad (18)
\]

where

\[
\nu = \sum_{i=1}^{n} (x_i + y_i) + \tilde{N}, \quad (20)
\]

\[
\zeta_k = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{S_{i,j},k} + y_i Z_{t_{i,j},k} + \sum_{j=1}^{\tilde{N}} Z_{S_{n+1,j},k}, \quad (21)
\]

\[
\xi_k = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{S_{i,j},k} S_{i,j} + \sum_{j=1}^{\tilde{N}} y_i Z_{t_{i,j},k} + \sum_{j=1}^{\tilde{N}} Z_{S_{n+1,j},k} S_{n+1,j}, \quad (22)
\]

Moreover, we consider two following cases on prior information:

- Non-informative prior case: The prior distributions are not given.
- Informative prior case: The prior distributions are given.

Then we have the Gibbs sampling as the following scheme:

- Generate samples of latent parameters:
  \[
  \tilde{N}|\omega, \pi, \alpha, \beta; D \sim P(\tilde{N}; \omega (1 - F(t_n; \pi, \alpha, \beta))) \quad (23)
  \]
  \[
  S_{i,j}|\pi, \alpha, \beta; D \sim f(S_{i,j}|t_{i-1} < S_{i,j} < t_i; \pi, \alpha, \beta), \quad (24)
  \]
  for all \( i = 1, \ldots, n \) and \( j = 1, \ldots, x_i \),
  \[
  S_{n+1,j}|\pi, \alpha, \beta; D \sim f(S_{n+1,j}|S_{n+1,j} > t_n; \pi, \alpha, \beta), \quad (25)
  \]
  for all \( j = 1, \ldots, \tilde{N} \),
  \[
  Z_{S_{i,j}}|\pi, \alpha, \beta; D \sim \mathcal{B}(Z_{S_{i,j}}; \psi_{S_{i,j}}), \quad (26)
  \]
  for all \( i = 1, \ldots, n \) and \( j = 1, \ldots, x_i \),
  \[
  Z_{t_{i,j}}|\pi, \alpha, \beta; D \sim \mathcal{B}(Z_{t_{i,j}}; \psi_{t_{i,j}}), \quad (27)
  \]
  for all \( i = 1, \ldots, n \),
  \[
  Z_{S_{n+1,j}}|\pi, \alpha, \beta; D \sim \mathcal{B}(Z_{S_{n+1,j}}; \psi_{S_{n+1,j}}), \quad (28)
  \]
  for all \( j = 1, \ldots, \tilde{N} \),
where
\[ \psi_u = \{ \pi_1 \mathcal{G}(u; \alpha_1, \beta_1), \ldots, \pi_m \mathcal{G}(u; \alpha_m, \beta_m) \} \left[ \sum_{k=1}^{m} \pi_k \mathcal{G}(u; \alpha_k, \beta_k) \right]^{-1}. \] (29)

- Generate the samples of parameters from (conditional) posterior distributions.
  - Non-informative prior case:
    \[ \omega|\pi, \alpha, \beta, \bar{N}; D \sim \mathcal{G}(\omega; 1 + \nu, 1), \] (30)
    \[ \pi|\bar{N}, Z; D \sim \mathcal{D}(\pi; \zeta), \] (31)
    \[ \alpha_k|\beta_k, \bar{N}, S, S', Z; D \sim p(D|\omega, \pi, \alpha, \beta, \bar{N}, S, Z), \] for all \( k = 1, \ldots, m, \) (32)
    \[ \beta_k|\alpha_k, \bar{N}, S, S', Z; D \sim \mathcal{G}(\beta; 1 + \alpha_k \zeta_k, \xi_k), \] for all \( k = 1, \ldots, m. \) (33)
  - Informative prior case:
    \[ \omega|\pi, \alpha, \beta, \bar{N}; D \sim \mathcal{G}(\omega; a_\omega + \nu, b_\omega + 1), \] (34)
    \[ \pi|\bar{N}, Z; D \sim \mathcal{D}(\pi; \rho + \zeta), \] (35)
    \[ \alpha_k|\beta_k, \bar{N}, S, S', Z; D \sim p(D|\omega, \pi, \alpha, \beta, \bar{N}, S, Z)p(\alpha_k), \] for all \( k = 1, \ldots, m, \) (36)
    \[ \beta_k|\alpha_k, \bar{N}, S, S', Z; D \sim \mathcal{G}(\beta; a_{\beta_k} + \alpha_k \zeta_k, b_{\beta_k} + \xi_k), \] for all \( k = 1, \ldots, m. \) (37)

Note that we apply the data augmentation technique [4] for the latent variables. For the sampling of \( \alpha_k \) in both cases, we should apply the Metropolis-Hastings method [5].

### 4 Numerical Examples

In this section, we investigate the efficacy of our semi-parametric SRM. In the experiment, we use failure data collections in The Data & Analysis Center for Software 1. In particular, the grouped data of System ss1a and System 1 are used for our experiment. System ss1a was collected from the system testing of an operating system. System 1 was collected from the system testing of a real time command and its control in the military application.

To compare the fitting ability of semi-parametric model, we consider some of parametric NHPP-based SRMs (exponential SRM [6], gamma SRM [7], truncated normal SRM [8] and log-normal SRM [8]). Also we suppose that the prior distributions of mixed gamma SRM are not given.

Tables 1 and 2 present the goodness-of-fit and the predictive performance of all the models in System aa1a and System 1, respectively. The goodness-of-fit is evaluated by the free energy [9]. The model with smaller free energy fits better to the observed data. Also, the error of prediction is computed by using the observed number of faults as follows:

\[ \frac{1}{T} \sum_{i=1}^{T} \left\{ \hat{\Lambda}(t_i; \omega, \theta) - \hat{\Lambda}(t_{i-1}; \omega, \theta) - (x_i + y_i) \right\}^2. \] (38)

\(^{1}\text{DACS The software reliability dataset http://www.dacs.dtic.mil/databases/sled/surel.shtml}\)
Let $B$ be the number of MCMC samples. The estimated value function in the above equation is given by

$$\hat{\Lambda}(t; \omega, \theta) = \frac{1}{B} \sum_{i=1}^{B} \omega^{(i)} F(t; \theta^{(i)}).$$  \hfill (39)

In Table 1, the gamma SRM is the best performance in terms of both data fitting and prediction. The mixed gamma SRM is the second best among them. In fact, from the definition of the mixed gamma SRM, the gamma SRM is a special case of the mixed gamma SRM. In this experiment, the number of gamma distributions in the mixed gamma distribution is fixed as $m = 2$. In addition, we did not provide the prior information for the mixed gamma SRM. Thus the mixed gamma SRM is inferior to the gamma SRM. However, it can be improved by applying the selection of the number of gamma components. Figures 1 through 5 depict the mean value functions fitted to the cumulative failures experienced in System ss1a. From these figures, it is found that the observed failures gradually decreases as testing time is elapsed. The large number of gamma components is not needed in the data having such tendency.

On the other hand, Table 2 showed that the mixed gamma SRM fitted well to System 1. However, the prediction does not work well. Similar to System ss1a, Figures 6 through 10 illustrate the mean value functions in System ss1a. From these figures, the data tendency has a suddenly changes just before the end of testing. In general, this kind of data requires a complex SRM like the mixed gamma SRM with a number of gamma components in terms of the data fitting ability. However, the prediction is quite difficult, because the data tendency has been changed before the end of testing. That is, this result indicates that the prior information is needed to perform the accurate prediction in such case.

5 Conclusions

This paper has proposed the semi-parametric SRM based on the mixed gamma distributions and has discussed the MCMC algorithm for the mixed gamma SRM. By applying the semi-parametric model, we do not need to decide the set of candidates from a vast amount of parametric SRMs before estimating their model parameters. This is a great merit from the practical point of view. In the experiment, we have examined the fitting ability of semi-parametric SRM. In future, we compare the semi-parametric model with non-parametric models to reveal the performance of prediction ability.

Acknowledgments

This research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), Grant No. 21510167 (2009-2011), Grant No. 23500047 (2011–2013) and Grant No. 23510171 (2011–2013). This paper is an extension of work originally reported in The 7th International Conference on Mathematical Methods in Reliability, –Theory, Methods and Applications – (MMR-2010) [10].
References


Table 1. The goodness-of-fit and the predictive performance (System ss1a).

<table>
<thead>
<tr>
<th></th>
<th>free energy</th>
<th>predictive error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed gamma SRM</td>
<td>141.93</td>
<td>0.388046</td>
</tr>
<tr>
<td>Exponential SRM</td>
<td>146.39</td>
<td>0.435578</td>
</tr>
<tr>
<td>Gamma SRM</td>
<td>141.06</td>
<td>0.366585</td>
</tr>
<tr>
<td>Truncated normal SRM</td>
<td>145.68</td>
<td>0.444543</td>
</tr>
<tr>
<td>Logarithmic normal SRM</td>
<td>143.19</td>
<td>0.480146</td>
</tr>
</tbody>
</table>

Table 2. The goodness-of-fit and the predictive performance (System 1).

<table>
<thead>
<tr>
<th></th>
<th>free energy</th>
<th>predictive error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed gamma SRM</td>
<td>67.946</td>
<td>3.65034</td>
</tr>
<tr>
<td>Exponential SRM</td>
<td>96.937</td>
<td>2.87066</td>
</tr>
<tr>
<td>Gamma SRM</td>
<td>80.961</td>
<td>3.89452</td>
</tr>
<tr>
<td>Truncated normal SRM</td>
<td>74.621</td>
<td>4.38434</td>
</tr>
<tr>
<td>Logarithmic normal SRM</td>
<td>82.453</td>
<td>3.44016</td>
</tr>
</tbody>
</table>

Figure 1. Mean value function of mixed gamma SRM (System ss1a).
Figure 2. Mean value function of exponential SRM (System ss1a).

Figure 3. Mean value function of gamma SRM (System ss1a).
Figure 4. Mean value function of truncated normal SRM (System ss1a).

Figure 5. Mean value function of log-normal SRM (System ss1a).
Figures 6 and 7 illustrate the mean value function of two different SRM models: mixed gamma SRM and exponential SRM, respectively, for System 1. The graphs show the cumulative number of faults over time, comparing observed faults with expected faults.

**Figure 6.** Mean value function of mixed gamma SRM (System 1).

**Figure 7.** Mean value function of exponential SRM (System 1).
Figure 8. Mean value function of gamma SRM (System 1).

Figure 9. Mean value function of truncated normal SRM (System 1).
Figure 10. Mean value function of log-normal SRM (System 1).