Combining imprecise Bayesian and maximum likelihood estimation for reliability growth models

Lev V. Utkin, Svetlana I. Zatenko, Frank P.A. Coolen

St.Petersburg Forest Technical Academy, Durham University

`lev.utkin@mail.ru, s_lana2004@mail.ru, frank.coolen@durham.ac.uk`

Abstract

A new framework is explored for combining imprecise Bayesian methods with likelihood inference, and it is presented in the context of reliability growth models. The main idea of the framework is to divide a set of the model parameters of interest into two subsets related to fundamentally different aspects of the overall model, and to combine Walley's idea of imprecise Bayesian models related to one of the subsets of the model parameters with maximum likelihood estimation for the other subset. In accordance with the first subset and statistical data, the imprecise Bayesian model is constructed, which provides lower and upper predictive probability distributions depending on the second subset of parameters. These further parameters are then estimated by a maximum likelihood method, based on a novel proposition for maximum likelihood estimation over sets of distributions following from imprecise Bayesian models for the other subset of parameters. Use of this hybrid method is illustrated for reliability growth models and regression models, and some essential topics that need to be addressed in order to fully justify and further develop this framework are discussed.

1. Introduction

GROWTH MODEL: $X_1, ..., X_n$ is a series of r.v. X_i is governed by a probability distribution function $p_i(x \mid \mathbf{b}, \mathbf{d})$ depending on two vectors of parameters $\mathbf{b}$ and $\mathbf{d}$. $\mathbf{b}$ contains parameters of the probability distribution under consideration. $\mathbf{d}$ characterizes the growth, i.e., the growth is modelled by a function $f(i, \mathbf{d})$. In software reliability analysis, the function f shows how parameters $\mathbf{b}$ of p_i change with the number of corrected errors or faults i.

A typical regression model is

$$Y = f(\mathbf{X}, \mathbf{d}) + \epsilon.$$

Here $\mathbf{X} = (1, X_1, ..., X_n)$; $\mathbf{d}$ is the vector of parameters; ϵ are uncorrelated random errors, usually assumed to have expected value 0 and unknown variance σ^2. $\mathbf{d}$ can be a set of growth parameters, for instance, coefficients in a linear regression model, $\mathbf{b} = (\sigma^2)$.

2. The likelihood principle for constructing standard models

Let $\mathbf{K} = (k_1, ..., k_n)$ be a realization of $X_1, ..., X_n$, with k_i non-negative integers. If probability distributions $p_i(k_i \mid \mathbf{b}, \mathbf{d})$ of the r.v. X_i, $i = 1, ..., n$, are known, then the standard way for obtaining the parameters $\mathbf{b}$ and $\mathbf{d}$ is to maximize the likelihood function

$$L(\mathbf{K} \mid \mathbf{b}, \mathbf{d}) = \prod_{i=1}^{n} p_i(k_i \mid \mathbf{b}, \mathbf{d})$$

over a set of $\mathbf{b}$ and $\mathbf{d}$.

3. Maximization of the likelihood function over a set of distributions

Suppose that the r.v. X_i is governed by an unknown CDF $F_i(k)$ which is only known to belong to the set $\mathcal{M}_i(\mathbf{d})$ defined by the lower and upper CDFs

$$\underline{F}_i(k \mid \mathbf{d}) = \inf_{\mathcal{M}_i(\mathbf{d})} F(k), \qquad (1)$$

$$\overline{F}_i(k \mid \mathbf{d}) = \sup_{\mathcal{M}_i(\mathbf{d})} F(k). \qquad (2)$$

The likelihood function can be written in the following form:

$$L(\mathbf{K} \mid \mathbf{d}) = \Pr\{X_1 = k_1, ..., X_n = k_n\}.$$

Proposition 1 *Suppose that discrete r.v. $X_1, ..., X_n$ are governed by a probability distribution $F(k)$ from sets $\mathcal{M}_i$ defined by bounds (1)-(2), respectively. If $X_1, ..., X_n$ are independent, then there holds*

$$\max_{\mathcal{M}_1, ..., \mathcal{M}_n} \Pr\{X_1 = k_1, ..., X_n = k_n\} = \prod_{i=1}^{n} \left\{\overline{F}_i(k_i) - \underline{F}_i(k_i - 1)\right\}. \qquad (3)$$

4. A general scheme of the model construction

1. We divide the set of parameters into two subsets. The first subset contains the parameters $\mathbf{b}$ of the assumed probability distribution p of the r.v. $X_1, ..., X_n$. The second subset consists of the growth parameters $\mathbf{d}$.
2. For the assumed distribution p, we choose an appropriate type of the conjugate prior $\pi(\mathbf{b} \mid \mathbf{c})$ with parameters $\mathbf{c}$.
3. We construct the corresponding Bayesian imprecise model on the basis of results of Walley or Quaeghebeur and de Cooman. We replace the parameters $\mathbf{c}$ by new parameters including the hyperparameter s. The produced set $\mathcal{P}$ depends on the hyperparameter s.
4. By using n observations $k_1, ..., k_n$, we write the lower $\underline{F}_i(k \mid \mathbf{d}, s)$ and upper $\overline{F}_i(k \mid \mathbf{d}, s)$ predictive CDFs as functions of the parameters $\mathbf{d}$ and the hyperparameter s for every debugging period. These functions form the sets $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$.

After completing the four steps of the first task, the sets $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$ have been derived and these sets do not depend on $\mathbf{b}$ or $\mathbf{c}$ and they depend only on $\mathbf{d}$, the hyperparameter s, and the number of debugging periods i. The second task is to estimate the parameters $\mathbf{d}$, it consists of two steps.

1. The likelihood function $L(\mathbf{K} \mid \mathbf{d}, s)$ is derived by applying Proposition 1.
2. Values of the parameters $\mathbf{d}$ for a fixed s should be chosen in such a way that makes $L(\mathbf{K} \mid \mathbf{d}, s)$ achieve its maximum.

5. A software run reliability growth model and the imprecise beta-geometric growth model

Let X be a run lifetime of software, that is, X is a discrete r.v. taking the value k if the software fails during the k-th run after $k-1$ successful runs, $p(k) = \Pr\{X = k\}$.

we assume that the r.v. X is governed by the geometric distribution with parameter r and the probability mass function

$$p(k \mid r) = (1-r)^{k-1} r, \ k = 1, 2, ...,$$

Suppose that the probability $r = r_i$ is a r.v. having a beta distribution with prior parameters α and $\beta + f(i, \varphi)$. Here $f(i, \varphi)$ is a function characterizing the software reliability growth. Assume for simplicity that $f(i, \varphi) = (i-1) \cdot \varphi$.

Denote the parameters of the i-th posterior beta distribution after n observations

$$\alpha^* = \alpha + n - 1, \ \ \beta_i^* = \beta + D_i(\varphi),$$

where

$$D_i(\varphi) = K_n + f(i, \varphi), \ K_n = \sum_{j=1}^{n-1} (k_j - 1).$$

The predictive CDF for the i-th step of the software debugging after n observations is

$$F_i(k \mid \varphi, \alpha, \beta) = \int_0^1 (1 - (1-p)^k) \cdot \mathrm{Beta}(\alpha^*, \beta_i^*) \mathrm{d}p = 1 - \frac{\mathrm{B}(\alpha^* + \beta_i^*, k)}{\mathrm{B}(\beta_i^*, k)}.$$

By replacing $\alpha = s\gamma$, $\beta = s - s\gamma$, we write

$$F_i(k \mid \varphi, \gamma, s) = 1 - \frac{\mathrm{B}(s + n - 1 + D_i(\varphi), k)}{\mathrm{B}(s - s\gamma + D_i(\varphi), k)}.$$

The lower bound for $\mathcal{M}_i(\varphi)$ is

$$\underline{F}_i(k, t \mid s, b) = 1 - I\left(\frac{T_i(t, b)}{\tau(t_n, b) + T_i(t, b)}, k + 1, s + K_n\right),$$

The upper bound is

$$\overline{F}_i(k, t \mid s, b) = 1 - I\left(\frac{T_i(t, b)}{s + \tau(t_n, b) + T_i(t, b)}, k + 1, K_n\right).$$

The likelihood function maximized over $\mathcal{M}_i(\varphi)$ by given s and φ is

$$\max_{\mathcal{M}(\varphi)} L(\mathbf{K} \mid \varphi, s) = \prod_{i=1}^{n} \left(\frac{\mathrm{B}(C_i, \ k_i - 1)}{\mathrm{B}(s + D_i(\varphi), \ k_i - 1)} - \frac{\mathrm{B}(C_i, \ k_i)}{\mathrm{B}(D_i(\varphi), \ k_i)}\right).$$

Here $C_i = s + n - 1 + D_i(\varphi)$.

The lower and upper software run failure functions after the n-th software failure are

$$\underline{F}_{n+1}(k, s) = 1 - \frac{\mathrm{B}(s + n + D_{n+1}(\varphi_0), \ k)}{\mathrm{B}(s + D_{n+1}(\varphi_0), \ k)},$$

$$\overline{F}_{n+1}(k, s) = 1 - \frac{\mathrm{B}(s + n + D_{n+1}(\varphi_0), \ k)}{\mathrm{B}(D_{n+1}(\varphi_0), \ k)}.$$

6. NHPP software reliability models and the imprecise negative binomial growth model

The non-homogeneous Poisson process (NHPP): Let $X_i = N(t_i) - N(t_{i-1})$ be the random number of failures between t_{i-1} and t_i. For any time points $0 < t_1 < t_2 < ...$ (for ease of notation, let $t_0 = 0$), the probability that the number of failures between t_{i-1} and t_i is k, $k = 0, 1, 2, ...$, can be written as

$$\Pr\{N(t_i) - N(t_{i-1}) = k\} = \frac{\{m(t_i) - m(t_{i-1})\}^k}{k!} e^{\{-(m(t_i) - m(t_{i-1}))\}}. \qquad (4)$$

Here $m(t)$ is the mean number of failures occurring up to time t.

The predictive probability of k failures during time t under condition that K failures were observed during time T is ($\alpha^* = \alpha + K$ and $\beta^* = \beta + T$)

$$P(k) = \int_0^\infty \frac{(\lambda t)^k e^{-\lambda t}}{k!} \mathrm{Gamma}(\alpha^*, \beta^*) \mathrm{d}\lambda = \frac{\Gamma(\alpha^* + k)}{\Gamma(\alpha^*) k!} \left(\frac{\beta^*}{\beta^* + t}\right)^{\alpha^*} \left(\frac{t}{\beta^* + t}\right)^k. \qquad (5)$$

Let $m(t; a, b) = a \cdot \tau(t, b)$. The parameter λ of the Poisson distribution in (5) and the argument t can be replaced by the parameter a and the discrete time $\tau(t_i, b) - \tau(t_{i-1}, b)$, respectively. In fact, by replacing λ by a, we get the Poisson process with a scaled time of the software testing, i.e., every time interval $[t_{i-1}, t_i]$ is replaced by the interval $[\tau(t_{i-1}, b), \tau(t_i, b)]$. Then we can write the predictive CDF of the number of failures in the time interval between t_i and t ($t \in [t_i, t_{i+1}]$) after n observation periods as follows:

$$F_i(k, t | \mathbf{c}, b) = 1 - \frac{\mathrm{B}_{q(i,t)}(k + 1, \alpha + K_n)}{\mathrm{B}(k + 1, \alpha + K_n)} = 1 - I\left(q(i, t), k + 1, \alpha + K_n\right).$$

Here $t_0 = 0$, $k_0 = 0$,

$$q(i, t) = \frac{T_i(t, b)}{\beta + \tau(t_n, b) + T_i(t, b)},$$

$$T_i(t, b) = \tau(t, b) - \tau(t_i, b), \ K_n = \sum_{j=1}^{n} k_j,$$

$\mathrm{B}_q(k+1, r)$ is the incomplete Beta-function with $I(q, k, r)$ the regularized incomplete Beta-function.

We choose all vectors (α, β) within the triangle $(0, 0)$, $(s, 0)$, $(0, s)$. This implies that all possible prior 'rates of occurrence of failures' are represented, as the prior allows interpretation of $\alpha / \beta = \gamma$ as this rate. This prior set leads to the lower and upper bounds for $\mathcal{M}_i(b)$ by $t \in [t_i, t_{i+1}]$

$$\underline{F}_i(k, t \mid s, b) = 1 - I\left(\frac{T_i(t, b)}{\tau(t_n, b) + T_i(t, b)}, k + 1, s + K_n\right),$$

$$\overline{F}_i(k, t \mid s, b) = 1 - I\left(\frac{T_i(t, b)}{s + \tau(t_n, b) + T_i(t, b)}, k + 1, K_n\right).$$

The next step is to maximize the likelihood function over the set of b

$$L(\mathbf{K} | b, s) = \prod_{i=1}^{n} \left(\overline{F}_i(k_i, t_i \mid s, b) - \underline{F}_i(k_i - 1, t_i \mid s, b)\right).$$

7. Regression model (general scheme)

$$Y = \mathbf{X}\mathbf{d} + \epsilon.$$

Here $\mathbf{X} = (1, X_1, ..., X_n)$; $\mathbf{d} = (d_0, ..., d_n)^{\mathrm{T}}$ is the vector of parameters; ϵ are random errors or noise having zero mean and the unknown variance σ^2.

Let us construct the imprecise Bayesian model for ϵ. If ϵ is governed by some probability distribution $p(z \mid \sigma)$ and there is the corresponding conjugate distribution $\pi(\sigma \mid \mathbf{c})$, then we can find the predictive CDF $F_n(z \mid s, \gamma)$ after having n observations $(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)$ depending on new parameters s, γ and its bounds $\underline{F}(z \mid s)$, $\overline{F}(z \mid s)$.

Denote $z_i = y_i - \mathbf{x}_i\mathbf{d}$ and $\mathbf{Z} = (z_1, ..., z_n)$. Then

$$\max_{\mathcal{M}} L(\mathbf{Z} \mid s) = \prod_{i=1}^{n} \left(\overline{F}(z_i \mid s) - \underline{F}(z_i - 1 \mid s)\right).$$

Denote $z_i = y_i - \mathbf{x}_i\mathbf{d}$. Hence

$$\max_{\mathcal{M}} L(\mathbf{Z} \mid s) = \prod_{i=1}^{n} \left(\overline{F}(y_i - \mathbf{x}_i\mathbf{d} \mid s) - \underline{F}(y_i - \mathbf{x}_i\mathbf{d} - 1 \mid s)\right).$$

Now we can find parameters $\mathbf{d}$ by maximizing the obtained likelihood function.

ISIPTA 2009, 6th International Symposium on Imprecise Probability: Theory and Applications, July 2009, Durham, United Kingdom