Nonparametric Analysis of the Order-Statistic Model in Software Reliability

Simon P. Wilson and Francisco J. Samaniego

Abstract—In the literature on statistical inference in software reliability, the assumptions of parametric models and random sampling of bugs have been pervasive. We argue that both assumptions are problematic, the first because of robustness concerns and the second due to logical and practical difficulties. These considerations motivate the approach taken in this paper. We propose a nonparametric software reliability model based on the order-statistic paradigm. The objective of the work is to estimate, from data on discovery times observed within a type I censoring framework, both the underlying distribution F from which discovery times are generated and N, the unknown number of bugs in the software. The estimates are used to predict the next time to failure. The approach makes use of Bayesian nonparametric inference methods, in particular, the beta-Stacy process. The proposed methodology is illustrated on both real and simulated data.

Index Terms—Beta-Stacy process, order statistics, reliability, testing strategies, nonparametric statistics, survival analysis.

✦

1 INTRODUCTION

SOFTWARE reliability models have received a lot of attention in recent decades, as software systems have become more pervasive and vital to the operation of many important aspects of modern life. These models attempt to describe the process of bug discovery in software, typically in the prerelease stage of software development, when the software is tested to ensure that it meets its specification and to remove errors. Among their important purposes is predicting the total number of errors that will be discovered in the code and optimizing the software development process, for example, deciding on optimal testing strategies and when to release software for use.

The data usually take the form of times between bug discoveries or counts of discovered bugs at known times. The structure of the data, and so a reasonable model for them, is very dependent on the type of software, the use to which the software is put, and the circumstances under which the data were collected. Because of this, a plethora of parametric models, based on many different modeling approaches and ideas about the structure of the data, have arisen that work well for certain data but lack robustness.

In this paper, we investigate one common class of software reliability model—the order-statistic model—and, in doing so, attempt to address the problem of achieving a better flexibility in the modeling approach. We also point out a deficiency in the sampling scheme that most software reliability models assume and offer a more reasonable

alternative. We consider a nonparametric approach on the basis that this provides the desired robustness in modeling. Bayesian inference is developed that infers the posterior distribution of the discovery time distributions and the number of bugs and is illustrated on real and simulated data. Once this is done, we show how the estimates may be used to predict next time to failure. As well as describing methodology for this class of models, we believe that it is an interesting extension of Bayesian nonparametric approaches.

Order-statistic models assume that there are an unknown number of bugs N in the software. Bugs are labeled $i = 1, \ldots, N$ and their discovery times are independent, specified by distribution functions $G_1, G_2, \ldots, G_n$. The observed sequence of discovery times are the order statistics of the G_i. The earliest order-statistic model for software reliability is the Jelinski-Moranda model [1], which assumes that the G_i are independent and identically distributed exponential random variables, although the authors did not categorize the model as of that type. The order-statistic model approach was first defined for use in software reliability in [2], although the idea was implicit in the earlier work of [3]. Here, the G_i were exponential (making the work of [1] a special case) but not necessarily identically distributed, equivalent to assuming that each bug failed as an independent Poisson process. There had been some work on order-statistic models in general reliability theory beforehand; see [4]. A more general order-statistic model, that moved beyond the exponential assumption was proposed in [5].

A feature of all these papers is that a parametric form for the G_i is assumed. Our work attempts to generalize to a nonparametric form for the G_i. However, in order to make inference practical, we work with a special case, where the G_i are identical; we explain why in Section 2. The work of [6] is the closest that we have found to our approach, where the general order-statistic model is treated nonparametrically.

- *S.P. Wilson is with the Department of Statistics, Trinity College Dublin, Dublin 2, Ireland. E-mail: simon.wilson@tcd.ie.*
- *F.J. Samaniego is with the Department of Statistics, University of California at Davis, One Shields Avenue, Davis, CA 95616. E-mail: fjsamaniego@ucdavis.edu.*

When N is assumed known, inference for an order-statistic model is typically straightforward. When N is unknown, as is the case with software testing data, the inference is a type of "how many kinds are there" problem, well-known in the literature on estimating numbers of distinct species. It is pointed out in [7] that, in some cases, the data contain little information about N; thus, inference can be more difficult. In this regard, a Bayesian approach has the advantage that it can make use of any prior knowledge about N to aid the inference. In [8], a Bayesian approach was described for an example of an exponential order-statistic model. For the general order-statistic model, inference procedures are described in [9] and [5], while [10] is an example of Bayesian inference.

There have been previous applications of Bayesian nonparametric methods in software reliability, but they have been restricted to a hierarchical semiparametric approach, where a parametric model for the discovery times is assumed, and a nonparametric model is assigned for the prior on a model parameter; see [11] and [12]. The approach is also related to work in warranty analysis, where the failure of systems under warranty is analyzed; see [13].

The paper is organized as follows: In Section 2, we define the order-statistic model and demonstrate the problems that can arise for inference when a parametric form of the model is assumed. In Section 3, we define the nonparametric form of the order-statistic model and describe a Bayesian inference procedure for the distribution function from which the discovery time distribution is defined and N, the number of bugs. Section 4 discusses how to present the inference and how to apply it to prediction. Section 5 gives examples using simulated and real data. In Section 6, we close with some concluding remarks.

2 A NONPARAMETRIC ORDER-STATISTIC MODEL

We assume that software has an unknown number N of bugs and is being tested. We denote the bug discovery times as $T_1, \ldots, T_n$ and they are the order statistics of a set of N independent random variables with distribution functions $G_1, \ldots, G_N$. Defining the distribution functions of the T_i to be $F_i(t)$, standard order-statistic theory defines the $F_i(t)$ in terms of the $G_i(t)$ and N; see [14]. For example, letting $G_i(t) = 1 - e^{-\theta t}$, $t \geq 0$, $\forall i$, it is shown in [2] that we have the well-known Jelinski-Moranda model [1], where

$$F_i(t \mid \theta, N) = 1 - \exp(-(N - i + 1)\theta t). \quad (1)$$

The model that we define and explore in this paper is a general order-statistic model, but the G_i are not defined parametrically. It is a very flexible extension to the Jelinski-Moranda model; the G_i are identical but are not given any parametric form. We let $G_i(t) = F(t)$, $\forall i$; hence, our model is defined by $F(t)$ and N.

Assuming a common F is a simplifying approximation, but we argue that the model is still useful—when the data model is in an appropriate neighborhood of the Jelinski-Moranda model, for example—and is a stepping stone to thinking about more complex nonparametric models. Indeed, without some assumption to link the G_i, a nonparametric analysis is a hopeless proposition. The

assumption of identical G_i allows us to implement an MCMC scheme for Bayesian inference. So, our work can be considered as a nonparametric generalization of the Jelinski-Moranda model or as a general order-statistic model with identical nonparametric discovery time distribution. We note that the nonparametric distribution can be multimodal, allowing one to think of the bugs as coming from separate subpopulations.

2.1 The Random Sampling Scheme

The usual assumption is that data consist of a set of k independent interdiscovery times $t_1, \ldots, t_k$, where k is fixed, having likelihood

$$L(\theta, N \mid t_1, \ldots, t_k) = \prod_{i=1}^{k} f_i(t_i \mid N, \theta),$$

where $f_i(t_i \mid N, \theta) = F_i'(t_i \mid N, \theta)$ is the density function of the ith interdiscovery time, parameterized by θ. We observe that a fixed and predetermined choice of k implies that one has assumed that $N \geq k$. This assumption is, in general, both risky and impossible to justify and runs counter to the experimenter's natural inclination to sample as many bugs as possible, both because so doing improves the software, through the removal/repair of the discovered bugs, and because the precision of one's estimates of model parameters tends to improve as the sample size increases.

Because of the tenuous nature of the "fixed k" assumption, we propose a type I censoring framework, where the software is observed for a fixed predetermined amount of time T^*. The protocol for bug discovery and removal is the same as that discussed above. A consequence of this alternative model is that the number k of bugs discovered in the time interval $(0, T^*)$ is a nonnegative random variable. While k can take on the value zero with positive probability, this probability is negligibly small in most applications of interest. This leads to $N - k$ discovery times to be right-censored at T^*.

This is a more appropriate and realistic framework for sampling software bugs than the fixed sample-size approach. Assume that the discovery times are independent with distribution function F and density f. We view our data to consist of both k and $t_1, \ldots, t_k$. Given N and F, the number of discovery times k before T^* is binomial with probability parameter $F(T^*)$. Given k and T^*, the order-statistic model gives the distribution of $t_1, \ldots, t_k$ to be the joint order statistics distribution of a sample of size k from F, given that $t_i < T^*$. Given this, we arrive at the likelihood for these data:

$$L(F, N \mid T^*, k, t_1, \ldots, t_k)$$
$$= P(t_1, \ldots, t_k, k \mid T^*, N, F)$$
$$= P(t_1, \ldots, t_k \mid T^*, k, N, F) \, P(k \mid T^*, N, F)$$
$$= \left[k! \prod_{i=1}^{k} \left(\frac{f(t_i)}{F(T^*)} \right) \right] \left[\binom{N}{k} F(T^*)^k (1 - F(T^*))^{N-k} \right]$$
$$= \frac{N!}{(N-k)!} (1 - F(T^*))^{N-k} \prod_{i=1}^{k} f(t_i),$$
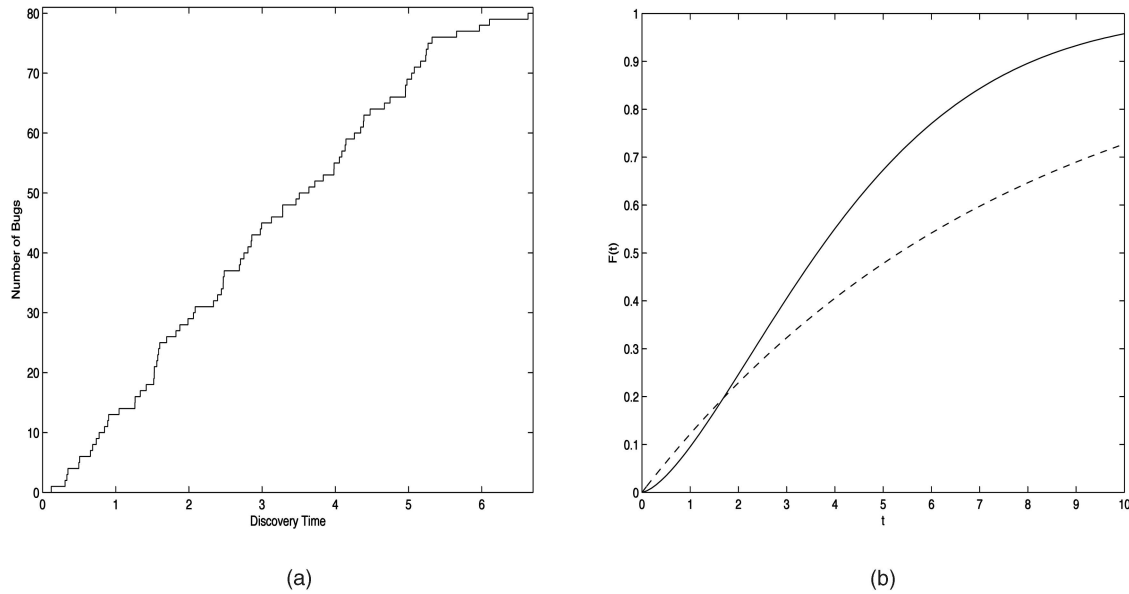$$t_1 \leq t_2 \leq \cdots \leq t_k \leq T^*. \quad (2)$$

Fig. 1. (a) The discovery times of the simulated Weibull data set. (b) The estimated F using MLE for the exponential case (dashed line) with the true Weibull F used to generate the data (solid line).

2.2 Parametric Models

Parametric models for bug counting schemes have been shown to have some undesirable properties and can lack robustness. In particular, most of the models assume reliability growth, and if the data do not clearly demonstrate this, then maximum likelihood estimates of N can become infinite; see [15]. The problems that can occur with the Jelinski-Moranda model in this regard are described in [16].

To illustrate this lack of robustness of the parametric bug counting model, we fit the Jelinski-Moranda model to a simulated data set of $k = 80$ discovery times, with $T^* = 6.7$ ($t_{80} = 6.63$ being the final discovery time) from a Weibull order-statistic model with $N = 100$ and

$$F(t) = 1 - \exp(-0.1t^{1.5});$$

thus, $F_i(t) = 1 - \exp(-0.1(101 - i)t^{1.5})$. The data are displayed in Fig. 1a. The maximum likelihood estimate for the Jelinski-Moranda model, for which $F(t) = 1 - \exp(-\theta t)$, is found numerically to be $(\hat{N}, \hat{\theta}) = (137, 0.13)$. The estimated $F(t) = 1 - \exp(-0.13t)$ is plotted in Fig. 1b, along with the true F, and we see that the true F has been very badly estimated, even though this particular Weibull is only modestly different from the Jelinski-Moranda model, where F is an exponential distribution.

3 BAYESIAN NONPARAMETRIC ANALYSIS

In this section, we describe a Bayesian inference procedure that computes the posterior distribution of F and N from discovery time data under the type I censoring scheme of Section 2.

A Bayesian analysis requires a likelihood and prior distribution for F and N. Given our model in terms of F and N, application of the expression for the likelihood in (2) gives us a likelihood

$$L(F, N \mid t_1, \ldots, t_k, T^*)$$
$$= \frac{N!}{(N - k)!}(1 - F(T^*))^{N-k} \prod_{i=1}^{k}(F(t_i) - F(t_i-)), \quad (3)$$

where we have replaced the probability density $f(t_i)$ with the term $F(t_i) - F(t_i-)$, referring to the limit from the left, appropriate where there is a point mass at t_i.

For a prior distribution, we assume that N and F are independent a priori. This is principally for mathematical convenience. However, a priori dependence between N and F may be desirable since, if N increases as the expected value of F increases, we may model the same beliefs about the discovery times. Thus, positive dependence between N and the expected value of F may be appropriate. In practice, we will assume rather flat, noninformative priors that we show have little effect on the posterior distribution.

3.1 Prior Specification for N

Several techniques have been proposed for specifying a prior for N in the context of bug counting, for example, the use of software metrics [17] or the use of elicitation techniques [18]. In these cases, analysis of the sensitivity of the posterior to the prior on N will be essential. One might also use noninformative proper priors, such as discrete uniform on a finite range $\{0, \ldots, N_{max}\}$, where N_{max} is a suitably chosen upper bound, such as the number of characters in the code.

3.2 Prior Specification for F

The usual prior distribution on the set of possible distribution functions is the Dirichlet process prior [19]. It is a distribution over the set of all discrete distribution functions and is defined by a finite measure β on the sample space, which is the expected value of the prior. The expected distribution has measure $\beta/\beta(\mathbb{R})$. If the data $x_1, \ldots, x_n$ are uncensored, then the posterior distribution is also a Dirichlet process prior, but the measure β changes to have point masses at the observed points $x_1, \ldots, x_n$. The size of

$\beta(\mathbb{R})$ controls the relative importance that is placed on the prior; by making this small, the posterior will be dominated by the data.

Inspired by the Jelinski-Moranda model, we assume that the mean of the prior model is the exponential distribution with failure rate θ; thus, $\beta(t) = Be^{-\theta t}$, for $B > 0$, where $B = \beta(\mathbb{R})$ is the prior weight. We may then specify a hyperprior distribution for θ or assign it a value. Here, we adopt an empirical Bayes approach and specify θ to be the MLE of the Jelinski-Moranda model. In our experience, the inference is not sensitive to the choice of θ if one keeps the prior weight B small relative to k.

3.3 The Posterior Distribution of (N, F)

The objective of the Bayesian analysis is to compute the posterior distribution of F and N given $t_1, \ldots, t_k, T^*$. This is done by simulating values of F and N using Gibbs sampling. That is to say, we simulate a distribution function F from $P(F \mid N, t_1, \ldots, t_k, T^*)$ and then a value of N from $P(N \mid F, t_1, \ldots, t_k, T^*)$. After repeated alternate sampling of F and N from these distributions, the sampled F and N are from $P(F, N \mid t_1, \ldots, t_k, T^*)$. Below, we describe how to sample from the two full conditional distributions.

3.4 Sampling from the Posterior Distribution of F Given N

The key to being able to simulate a realization of F given N is to consider the data as purely observations from F. As we observed when defining (3), if N is known, then the order-statistic model that we propose allows us to interpret our data as k observations $t_1, \ldots, t_k$ from F and $N - k$ observations right-censored at T^*. The presence of right-censored data means that the posterior distribution is no longer represented by a Dirichlet process but is of a more general form defined through a stochastic process called the beta-Stacy process; Walker et al. [20] discuss beta-Stacy processes in detail, and we refer the reader to that paper for a comprehensive description.

For our purposes, it is sufficient to know that the posterior of F can be written as $F(t) = 1 - \exp(-Z(t))$, where $Z(t)$ is a beta-Stacy process. Such processes have countably many points of discontinuity and can be written as the sum of independent increments $W_1, W_2, \ldots$ at those discontinuity points and a continuous component. It is shown in [21] that, with a Dirichlet prior having measure $\beta(t)$ and data observed at $x_1, x_2, \ldots$, some of which may be right censored, the posterior for F can be written $F(t) = 1 - \exp(-Z(t))$, where

$$Z(t) = Z_c(t) + \sum_{i : x_i \text{ uncensored}} I(x_i < t) W_i, \qquad (4)$$

where I is the indicator function, $Z_c(t)$ is a continuous Levy process which they define (also see [21]), the jumps W_i have density function

$$f_i(w) \propto (1 - e^{-w})^{N\{x_i\}-1}$$
$$\times \exp\left(-w\left[\beta(x_i) + \sum_j I(x_j \geq x_i) - N\{x_i\}\right]\right), \qquad (5)$$

and where $N\{x_i\}$ is the number of exact (noncensored) observations at x_i.

Our goal is to sample from this process. We will see that, to sample N, it is sufficient to compute F at the t_i, $i = 1, \ldots, k$, and T^*. Therefore, we sample the jumps $W_1, \ldots, W_k$ that occur at $t_1, \ldots, t_k$—there is no jump at T^* because there is no uncensored observation—and sample Z_c between successive observations, e.g.,

$$Z_c(t_1), Z_c(t_2) - Z_c(t_1), \ldots, Z_c(t_k) - Z_c(t_{k-1})$$

and $Z_c(T^*) - Z_c(t_k)$, from which

$$Z_c(t_i) = Z_c(t_1) + \sum_{j=2}^{i} [Z_c(t_j) - Z_c(t_{j-1})].$$

Then, $F(t_i-)$ and $F(t_i)$ are given by

$$F(t_i-) = 1 - \exp(-Z(t_i-))$$
$$= 1 - \exp\left(-Z_c(t_i) - \sum_{j=1}^{i-1} W_j\right) \qquad (6)$$

and

$$F(t_i) = 1 - \exp(-Z(t_i))$$
$$= 1 - \exp\left(-Z_c(t_i) - \sum_{j=1}^{i} W_j\right). \qquad (7)$$

To sample the jumps, it is easy to sample from the distribution in (5). In most cases, all the $t_1, \ldots, t_k$ are distinct, so $N\{t_i\} = 1$ and the W_i are exponentially distributed with a mean $[\beta(t_i) + N - i]^{-1}$. The sampling of the continuous component is more complex, but follows exactly the developments in [21]. A Markov chain Monte Carlo method is needed; in fact, two chains are required for each sample of $Z_c(t_i) - Z_c(t_{i-1})$. For the reader's convenience, the sampling method is described in Appendix A.

3.5 Sampling from the Posterior Distribution of N Given F

Given F, the full conditional distribution of N is by Bayes' law:

$$P(N \mid F, t_1, \ldots, t_k, T^*) \propto L(F, N \mid t_1, \ldots, t_k, T^*) P(N)$$
$$= \frac{N!}{(N-k)!} (1 - F(T^*))^{N-k} \left(\prod_{i=1}^{k} (F(t_i) - F(t_i-))\right) P(N)$$
$$\propto \frac{N!}{(N-k)!} (1 - F(T^*))^N P(N), \qquad (8)$$

for $N \geq k$, with

$$1 - F(T^*) = \exp\left(-Z_c(T^*) - \sum_{i=1}^{k} W_i\right)$$
$$= \exp\left((-Z_c(t_1) - (Z_c(T^*) - Z_c(t_k))\right)$$
$$- \sum_{i=2}^{k} (Z_c(t_i) - Z_c(t_{i-1})) - \sum_{i=1}^{k} W_i\right);$$

hence, to sample N it is sufficient to have sampled the W_i and Z_c between successive discovery times and from t_k to T^*, as described in the last Section 3.4.

We directly compute the probabilities of (8) and sample N by the inverse distribution function method.

3.6 MCMC Convergence

As with all MCMC methods, we must assess if the sampling method has converged. Lack of convergence is usually tested by plotting traces of the sampled quantities in F—the W_i and the continuous components $Z_c(t_i) - Z_c(t_{i-1})$—and N, by applying statistical tests for autocorrelations and other signs of nonconvergence. However, here the issue is complicated because each sampled component $Z_c(t_i) - Z_c(t_{i-1})$ is itself computed from the output of two MCMC chains (see Appendix A). A possibility is that we judge F and N to be sampled from a chain in equilibrium, but that this is from unconverged chains for the $Z_c(t_i) - Z_c(t_{i-1})$. Further, given that $2(k+1)$ chains must be run to get each sample of F, it is impractical to go through each of these chains individually, looking for signs of nonconvergence.

We attempt to address this problem by running each of the chains required for $Z_c(t_i) - Z_c(t_{i-1})$ for a different randomly chosen number of iterations. An example is given in Appendix B, where the length of the first chain was chosen uniformly between 7,000 and 25,000 iterations; this pertains to a parameter that is denoted λ. The second chain samples a sequence of η values denoted by $\tau_1, \ldots, \tau_\eta$. This chain was constructed by a uniformly random burn-in period between 5,000 and 50,000 iterations, and the number of iterations between each sample τ_j was uniformly chosen between 2,000 and 15,000. Then the computed values of λ and $Z_c(t_i) - Z_c(t_{i-1})$ are compared against the number of iterations in the MCMC chain used to compute them. If the chain lengths are sufficient, we should see no relationship between number of iterations and the computed values. If we see no evidence from this test of nonconvergence or bad mixing, we proceed to look at the samples of F and N in the usual way to test for nonconvergence.

4 DISPLAYING THE POSTERIOR OF F, MODEL FIT, AND PREDICTING TIME TO NEXT DISCOVERY

In this section, we describe how to display the posterior distribution in an informative manner, how to assess the performance of the fitted model, and how to use the posterior distribution of N and F that was simulated in Section 3 to compute the distribution of the next time to failure.

We assume that L samples of $F(t)$ and N have been generated. Let $(F(t)^{(l)}, N^{(l)})$ be the lth sample from the posterior distribution of F and N. For each observed time, we estimate the posterior mean of $F(t_i)$ and $F(t_i-)$ as the mean of the sampled values of $F(t_i)$ and $F(t_i-)$, for $i = 1, \ldots, k$ using (6) and (7). Estimates of $F(t)$ between observation times are obtained by linear interpolation of $Z(t)$. Similarly, taking high and low percentiles of the $F(t_i)^{(l)}$ gives upper and lower bounds to the posterior at each observed time.

Model evaluation is done by comparing the observed failure times with estimates of its distribution. The ith observed discovery time t_i is the ith order statistic from an independent sample of size N with distribution F. Hence,

$$F_i(t) = \sum_{j=i}^{N} \binom{N}{j} F(t)^j (1 - F(t))^{N-j},$$

from which an obvious predicted value is the posterior median. We compare this posterior median with t_i. For each sample of $F(t)$ and N, we compute its ith order statistic distribution:

$$F_i(t)^{(l)} = \sum_{j=i}^{N^{(l)}} \binom{N^{(l)}}{j} [F(t)^{(l)}]^j (1 - F(t)^{(l)})^{N^{(l)}-j}.$$

By interpolating between the t_i, any percentile of $F_i(t)^{(l)}$ can be computed; we denote the 100α percentile as $F_{i,\alpha}^{(l)}$. The posterior mean of the 100α percentile of the distribution of the ith observed discovery time can be computed by

$$F_{i,\alpha} \approx \frac{1}{L} \sum_{l=1}^{L} F_{i,\alpha}^{(l)}. \tag{9}$$

The median is calculated from (9) with $\alpha = 0.5$. A prediction interval for the ith discovery time is then the 2.5 percent and 97.5 percent points of the posterior distribution of $F_i(t)$, again calculated by (9). Because of the right censoring in the data, we may not estimate $F_i(t)$ well in the right tail. This feature is quite familiar in nonparametric estimation; for example, the Kaplan-Meier estimator of an underlying survival function never attains the value zero when the largest observation is a censored failure time. This is particularly the case for i near k, and more so if the posterior distribution of N places large probability close to k.

The distribution of the discovery time of the next bug is the $(k+1)$th order statistic, left-truncated at T^* since we did not observe it by this time. If $N = k$, then there are no more bugs to be discovered and $T_{k+1} = \infty$ with probability 1. We define this in terms of the reliability function $\bar{F}_{k+1}(t \,|\, T_{k+1} \geq T^*) = 1 - F_{k+1}(t \,|\, T_{k+1} \geq T^*)$:

$$\bar{F}_{k+1}(t \,|\, T_{k+1} \geq T^*) = \frac{1 - F_{k+1}(t)}{1 - F_{k+1}(T^*)}$$

$$= \begin{cases} \dfrac{\sum_{j=0}^{k} \binom{N}{j} F(t)^j (1-F(t))^{N-j}}{\sum_{j=0}^{k} \binom{N}{j} F(T^*)^j (1-F(T^*))^{N-j}} & \text{if } N \geq k+1, \\ 1, & \text{if } N = k, \end{cases} \tag{10}$$

for $t \geq T^*$, which we compute in a similar way to the computation of $F(t)$. For the lth sample, we compute

$$\bar{F}_{k+1}(t \,|\, T_{k+1} \geq T^*)^{(l)} = 1$$

if $N^{(l)} = k$ and

$$\bar{F}_{k+1}(t \,|\, T_{k+1} \geq T^*)^{(l)}$$

$$= \frac{\sum_{j=0}^{k} \binom{N^{(l)}}{j} [F(t)^{(l)}]^j (1 - F(t)^{(l)})^{N^{(l)}-j}}{\sum_{j=0}^{k} \binom{N^{(l)}}{j} [F(T^*)^{(l)}]^j (1 - F(T^*)^{(l)})^{N^{(l)}-j}}$$

if $N^{(l)} \geq k+1$. We then take the mean of the $\bar{F}_{k+1}(t)^{(l)}$ as our estimate of $\bar{F}_{k+1}(t\,|\,T_{k+1} \geq T^*)$. Uncertainty in $\bar{F}_{k+1}(t)$ is described by looking at percentiles of the $\bar{F}_{k+1}(t)^{(l)}$, e.g., the 2.5th and 97.5th percentiles provides a 95 percent probability interval. Since $\bar{F}_{k+1}(t) = 1$, $\forall t$ with probability $P(N = k\,|\,t_1, \ldots, t_k, T^*)$, the 100α percentile of the posterior distribution of $F_{k+1}(t)$ will be 1,

$$\forall \alpha \leq P(N = k\,|\,t_1, \ldots, t_k, T^*).$$

In practice, we find the inference easier to interpret if we report $P(N = k\,|\,t_1, \ldots, t_k, T^*)$ and $\bar{F}_{k+1}(t)$ conditional on $N \geq k+1$, where we only consider samples for which $N^{(l)} > k$, e.g., the posterior mean is

$$\bar{F}_{k+1}(t\,|\,T_{k+1} \geq T^*, N \geq k+1)$$
$$\approx \frac{1}{|\{l\,|\,N^{(l)} > k\}|} \sum_{l:\,N^{(l)} \geq k+1} \bar{F}_{k+1}(t\,|\,T_{k+1} \geq T^*)^{(l)},$$
$$t \geq T^*.$$

Note that, to estimate F_{k+1}, we require samples of $F(t)^{(l)}$ for $t > T^*$. We only sampled values of F up to $t = T^*$ in Section 3, so to do this, we need to generate a sample of the beta-Stacy process $Z(t)$ for $t > T^*$. Since the Gibbs sampler of Section 3 gives us samples of $Z(T^*)$ and there are no jumps in the beta-Stacy process for $t > T^*$—they only occur at $t_1, \ldots, t_k$—what is required is to simulate the continuous part of the process from T^* to t. This we do in the manner of Section 3 for as many different values of t as we require.

5 EXAMPLES

5.1 Simulated Weibull Data

To see how well the inference procedure works, it is applied to the simulated Weibull data described in Section 2.2 and Fig. 1. The maximum likelihood estimate from the Jelinski-Moranda model for these data is $(\hat{\theta}, \hat{N}) = (0.13, 137)$. The prior on F is therefore taken to be Dirichlet with mean as an exponential distribution with $\theta = 0.13$ and a small weight $B = 1$. A uniform prior on $\{0, 1, \ldots, 1,000\}$ is placed on N.

Ten thousand samples of N and F were generated. The convergence of the chains was assessed using the ideas of Section 3.6. The assessment showed no signs of nonconvergence, and we determined adequate chain lengths for the sampling of λ to be 100,000 iterations, while for the τ_j, there was a burn-in of 20,000 iterations, with successive τ_j taken every 10,000 iterations. Details of how the chain lengths are determined are in Appendix B.

Fig. 2 shows the sampled values and autocorrelation function of $F(t_1)$ and N, respectively. The chain for $F(t_1)$ shows good mixing and no sign of nonconvergence. It is typical of those for other $F(t_i)$. However, the sampled values for N are not as well behaved and show persistent autocorrelation. This is due to the very long tail in the distribution of N that the Gibbs sampler takes time to explore. It is clear that the prior upper bound to N of 1,000 has truncated the posterior distribution.

Fig. 3 shows the posterior distribution of F, as described in Section 4, and a histogram of the sampled values of N. The plot of F also shows the Weibull distribution that generated the data and the Kaplan-Meier estimate for F
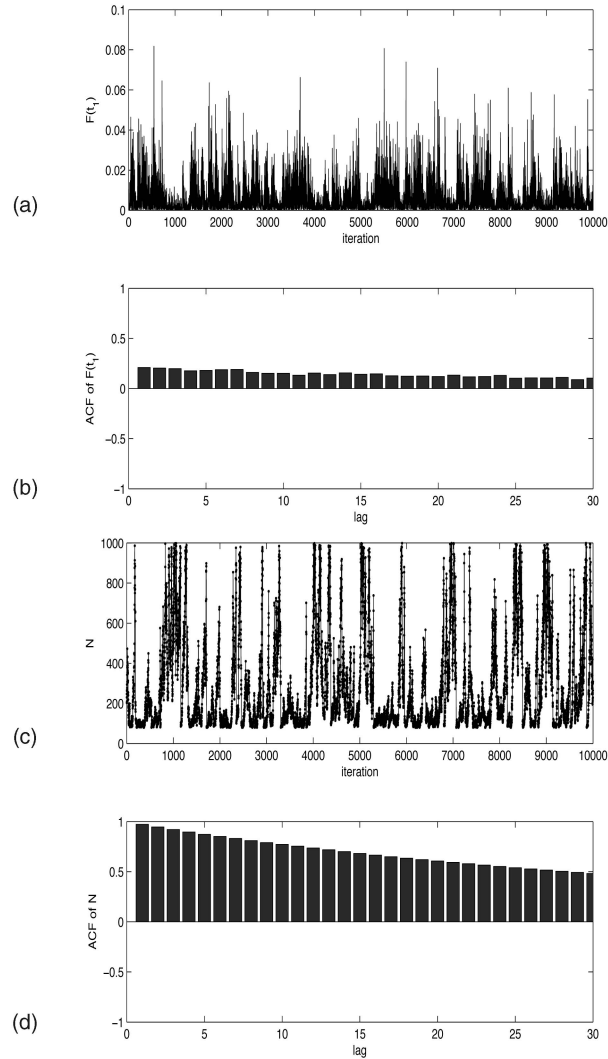

Fig. 2. (a) The sampled values of $F(t_1)$. (b) Their autocorrelation function for the simulated Weibull order-statistic model data. (c) The sampled values of N. (d) Their autocorrelation function.

given that we assume $N = 100$ (the true value). We note that the true F lies within the pointwise posterior probability interval for F, although the posterior mean for F is not close to the true F. However, we note that our mean estimate for F is very close to the Kaplan-Meier estimate, but that the latter is constructed given the true value of N. The posterior median of N is 180 and the posterior (2.5 percent, 97.5 percent) interval is (82, 913). The posterior mean is estimated to be 281. The true value of N is the 18th percentile of the sampled values.

Further, in Fig. 4a, we compare the observed values with the predicted, and we see that the model has made good predictions of the observed values. In Fig. 4b is the prediction for the reliability function of the time to next failure, $R_{81}(t) = (1 - F(t))^{N-80}$, conditional on $N \geq 81$. In this case, we estimate $P(N = 80\,|\,\text{data}) = 0.008$. This plot also shows the lower and upper 10th percentile of $R_{81}(t)$, the true Weibull reliability function from which the data were generated and the Jelinski-Moranda F (the prior). We see that the estimated function overestimates the true reliability
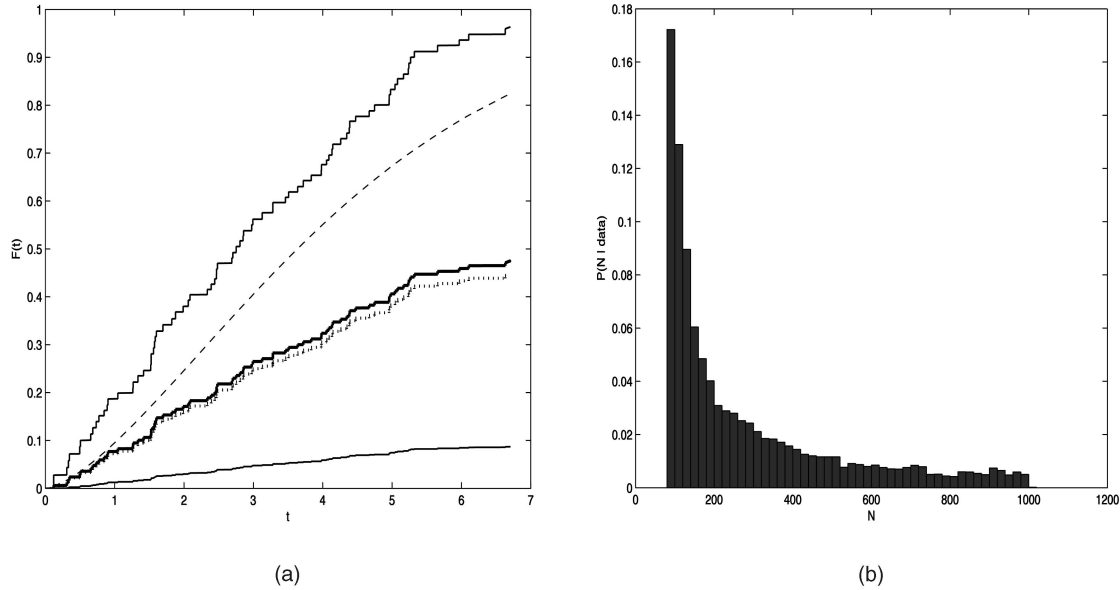
(a)

(b)

Fig. 3. Analysis of the Weibull order-statistic simulated data. (a) The true and posterior $F(t)$: pointwise posterior mean (thick solid line), pointwise posterior 2.5 and 97.5 percentiles (thin solid lines) and true F (dashed line) and the Kaplan-Meier estimate of $F(t)$ given the true value $N = 100$ (dotted line). (b) A relative histogram of posterior samples of N (grouped in intervals of 20).

from the simulated model $R_{81}(t) = \exp(-1.5t^{1.5})$, which is mainly due to the underestimation of N, but that the true function lies well within the prediction bounds.

We also assessed the robustness of the posterior of F to the value of θ. Setting $\theta = 1.0$, we saw that, in spite of the large misspecification, the posterior of F was not different. The small value of B has ensured that the results are insensitive to the choice of prior distribution parameter.

This example shows that the model has fitted the data well (Fig. 4a); in particular, we have estimated F to be close to the Kaplan-Meier estimate given the true N. There is high

posterior variance in both F and N (Fig. 3), the price that one pays for adopting such a general model. The posterior distribution of N is sensitive to the choice of prior. However, experimentation showed that the model fit and reliability prediction (as in Fig. 4) were not sensitive to the upper bounds on the uniform prior that were higher than 1,000.

5.2 The NTDS Data

The Naval Tactical Data System was a large software project for the US Navy. Data on the times of bugs detected in the testing phase of one of the modules form the NTDS data set, which originally appeared in [1].
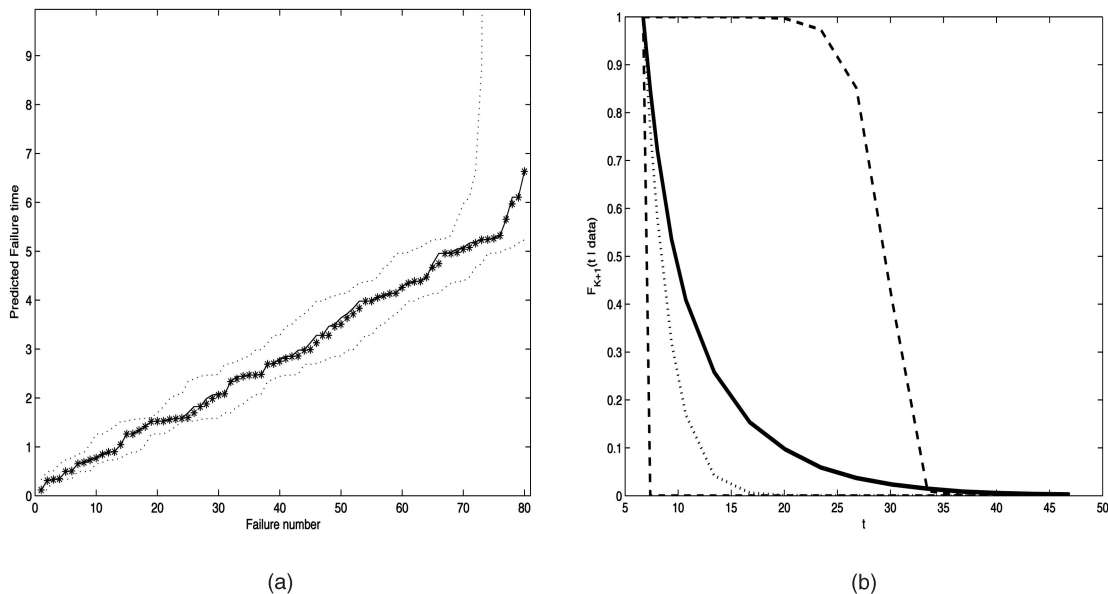

(a)

(b)

Fig. 4. (a) Observed bug discovery times of the Weibull order-statistic model data plotted with the posterior median of the median discovery time (solid line), and the posterior mean of the 10th and 90th percentiles (dotted lines). (b) The posterior reliability function of the 196th time to failure given $N \geq 196$: pointwise posterior mean (solid line), pointwise posterior 2.5th and 97.5th percentiles (dashed lines), and the true reliability function (dotted line).
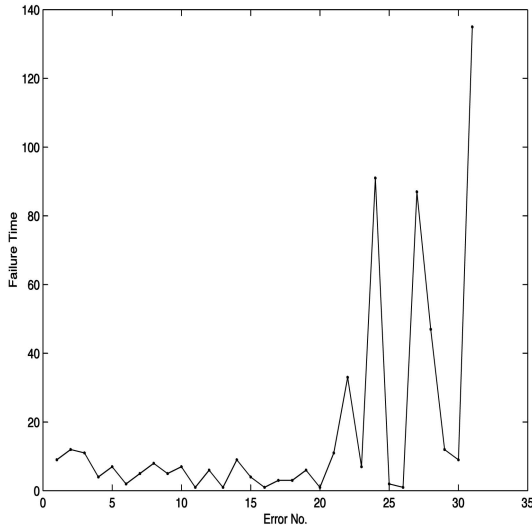
Fig. 5. The NTDS data set.

Fig. 5 displays the first $k = 31$ discovery times in the data, with $T^* = 600$ ($t_{31} = 540.0$ being the final discovery time). The MLE in this case is $(\hat{N}, \hat{\theta}) = (31, 0.0064)$. The prior on F is assumed to be a Dirichlet process with exponential distribution mean having $\theta = 0.0064$, and a small weight $B = 1$. A uniform prior on $\{0, 1, \ldots, 200\}$ was placed on N.

The MCMC was run for 10,000 iterations. Fig 6a shows the estimated posterior of F along with the prior, which is also the MLE estimate from the JM model. The prior and the posterior agree quite closely. Fig. 6b shows the sampled values of N. The posterior median of N is 31, and the (2.5 percent, 97.5 percent) probability interval is (31, 32). The posterior mean is 31.1. Moving to model assessment, Fig. 7a shows the predicted discovery times with the observed, as described in Section 4. It shows that, once again, the posterior distribution function has predicted well the observed discovery times.

Fig. 7b shows the posterior reliability function

$$R_{32}(t) = 1 - F_{32}(t) = (1 - F(t))^{N-31}.$$

The posterior probability that $N = 31$ (that is, that there are no further failures) is 0.94. The distribution compares well with the most recently observed discovery times and, in this case, agrees well with the reliability fitted from the JM model. Again, we see that the mean estimate for $R_{K+1}(t)$ performs well, although there is considerable uncertainty in the estimate. The MLE from the JM model is $\hat{N} = k = 31$, so it predicts that there are no new failures.

6 DISCUSSION

In this paper, we have presented a Bayesian nonparametric approach to the treatment of bug discovery time data. We have argued that the assumption of a parametric model leaves the investigator vulnerable to the consequences of model misspecification. For the same reasons that the empirical distribution function or the Kaplan-Meier estimator might be preferred to a parametric alternative, one might well prefer a nonparametric analysis in assessing the reliability of a piece of software. The model is a nonparametric form of the order-statistic model. The inferential approach presented here has the natural robustness of a nonparametric analysis and has, as well, the ability to incorporate, through the modeling of prior information, pertinent intuition or expert knowledge that might be available in the application of interest.

The approach we have taken has consciously eschewed the "random sampling" assumption that is typical of traditional treatments of the estimation problem under study. Our developments assume, instead, that the software being tested is subjected to such testing over a fixed and predetermined interval of time. Under this type I censoring scheme, the number of bugs found is a random variable, and the inference developed is based on the random discovery times between the (random) number of bugs found.
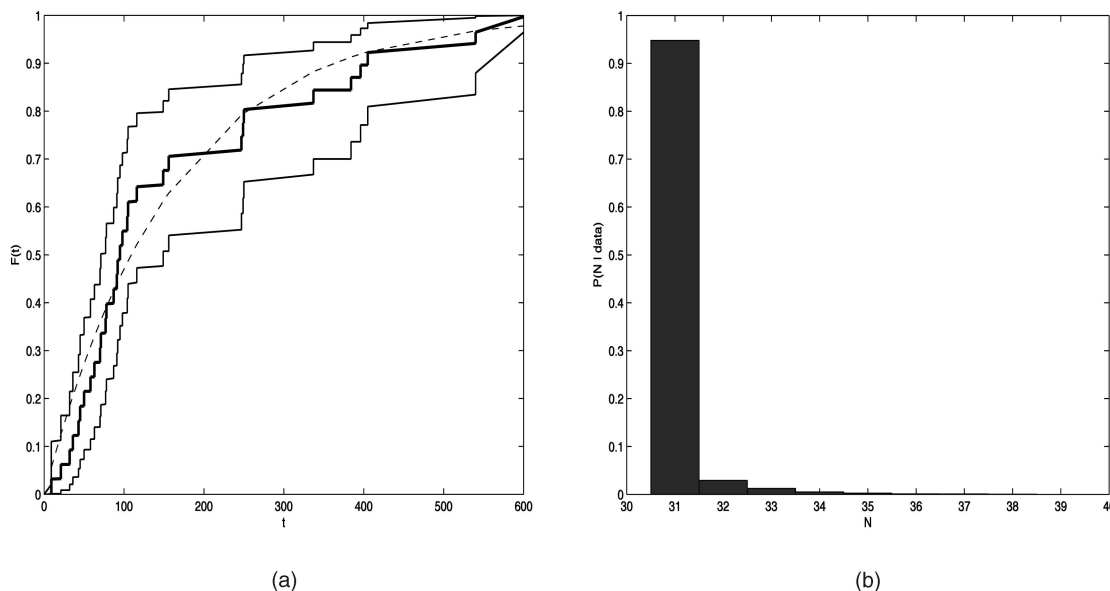

(a)

(b)

Fig. 6. (a) The prior and posterior of $F(t)$ for the NTDS data. Lines are pointwise posterior mean (thick solid line), pointwise posterior 2.5th and 97.5th percentiles (thin solid lines), and prior mean (dashed line). (b) A relative histogram of posterior samples of N.

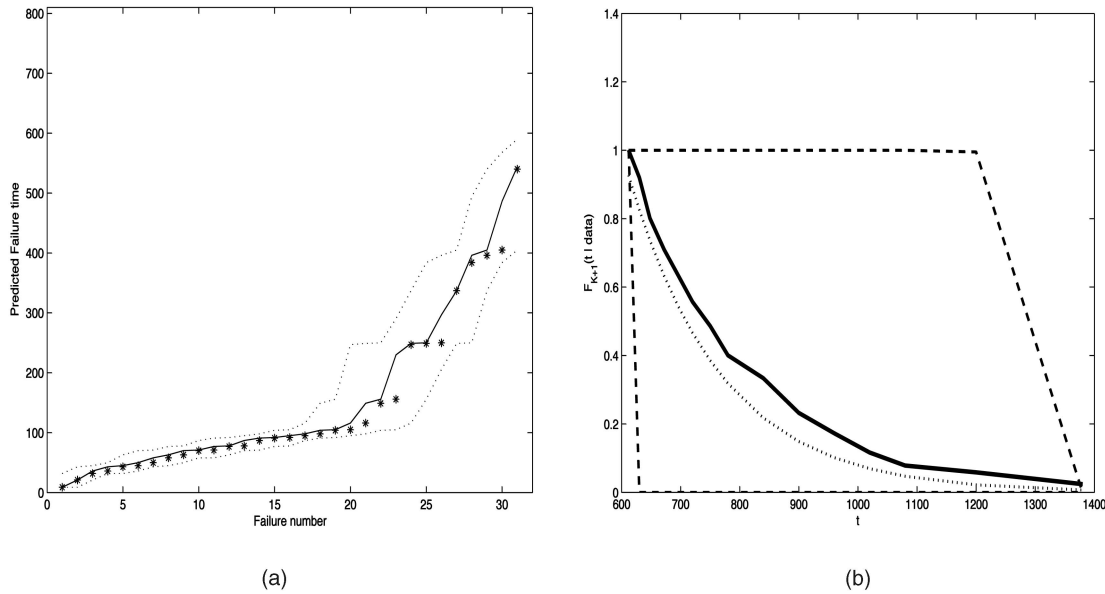(a)   (b)

Fig. 7. (a) Observed bug discovery times of the NTDS data, plotted with the posterior mean of the median discovery time (solid line) and the posterior mean of the 10th and 90th percentiles (dotted lines). (b) The posterior reliability function of the 32nd time to failure for the NTDS data given $N > 32$: pointwise posterior mean (solid line), pointwise posterior 2.5th and 97.5th percentiles (dashed lines), and reliability according to the MLE fit of the JM model (dotted line).

The approach presented in Sections 3 through 5 requires that the tester specify a prior distribution for N and a distribution function that is the mean of the prior for F, and we have suggested how these can be defined. The computation of the posterior distribution and presentation of results can be complicated, but we have successfully demonstrated its feasibility in the examples presented herein.

The strengths of this approach are the more realistic sampling scheme and a more robust fitting procedure. The disadvantages are a complicated computation process, and the problem of evaluating convergence of the many separate Markov chains in the simulation procedure. Other criticisms are the assumption that the distributions generating the order statistics are identical and that we have not modeled a priori dependence between F and N.

The former criticism is more difficult to address than the latter. If one allowed each bug to have a possibly different F, then there would be one datum available to learn about each. The success of the inference relies on being able to treat all the observations as coming from one distribution F. Any model that relaxed the identical distribution assumption and wanted to make use of the method described here would have to allow the data to be treated in this way. We also point out that our modeling of F allows it to be multimodal and, thus, describe a mixture of subpopulations of bugs.

Modeling dependence between N and F is more straightforward; for example, one could allow the expected value of the mean of F to depend on N. An obvious extension of the work is a fully Bayesian approach, where inference is also conducted on the parameters of the prior mean of F. The problem is that it is difficult to write the likelihood in terms of these parameters. Another extension is to fit count data rather than discovery times, by using data augmentation to sample times from a sample of F conditional on the counts.

The developments in this paper provide a useful stepping stone for facilitating further research on nonparametric software reliability. As with any Bayesian analysis,

we should mention the need to tailor the analysis to the intended application. A sensitivity analysis within any given application, so as to determine the extent to which the analysis is affected by the prior model, is important (see chapter 6 of [22]). One can present a Bayesian analysis with greater conviction when the inference is fairly stable over a class of "reasonable" prior distributions than when it is quite sensitive to the prior selected within such a class.

APPENDIX A

SAMPLING FROM $Z_c(t_i) - Z_c(t_{i-1})$

Let there be a Dirichlet process prior for F with measure $\beta(s)$. Our data consist of noncensored observations at $t_1, \ldots, t_k$ and $N - k$ right-censored observations at T^*. In [21], it is shown that the posterior distribution for F can be written in the form $F(t) = 1 - \exp(-Z(t))$, where $Z(t)$ is a beta-Stacy process. As is described in Section 3, $Z(t)$ can be written as the sum of independent increments $W_1, \ldots, W_k$ at the observed discovery times $t_1, \ldots, t_k$ and a continuous component $Z_c(t)$. In this appendix, we describe how to sample from $Z_c(t_i) - Z_c(t_{i-1})$, for $i = 1, \ldots, k$, the change in the continuous component between successive discovery times. The description is also valid for sampling from $Z_c(T^*) - Z_c(t_k)$. The beta-Stacy process has independent increments, and so these increases in the continuous process are independent.

Define $\alpha(s)$ to be such that $\beta(s) = \int_s^\infty d\alpha(u)$. In this case, we have defined $\beta(s) = B e^{-\theta s}$, from which $d\alpha(s) = B\theta e^{-\theta s} ds$. From [21], in our case, the continuous part of the posterior process $Z_c(t)$ has Levy measure

$$K(z, t) = \frac{\exp(-z[\beta(t) + N - i + 1])d\alpha(t)}{1 - e^{-z}}. \qquad (11)$$

In [23], it is shown that, by its Levy representation, we can approximate $Z_c(t_i) - Z_c(t_{i-1})$ as a compound Poisson
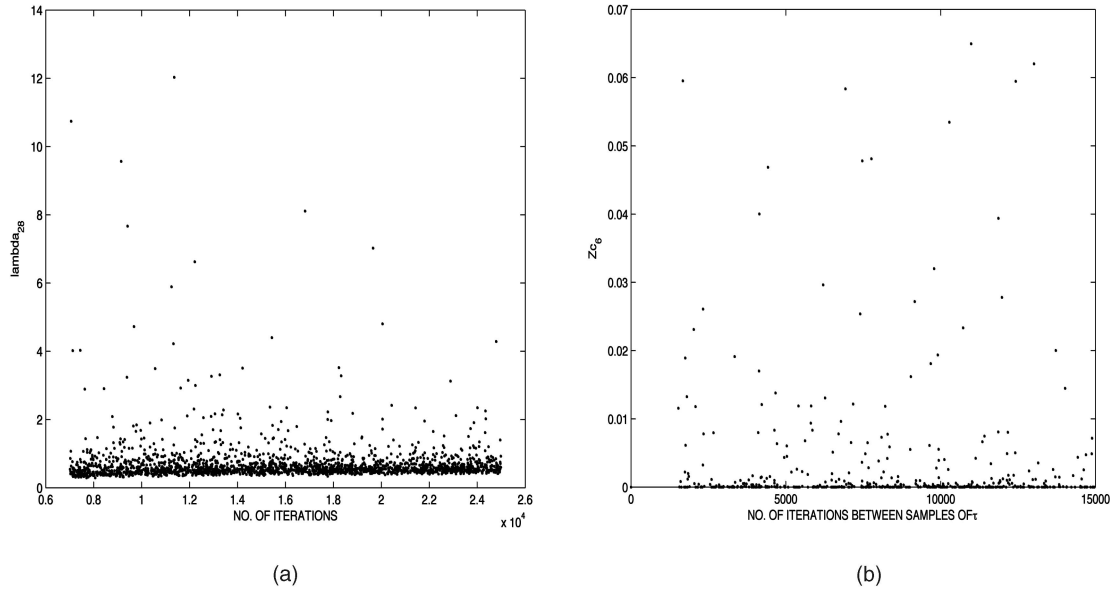
Fig. 8. (a) Sampled values of λ for $Z_c(t_{28}) - Z_c(t_{27})$ against the number of iterations in the MCMC chain used to produce that value. (b) Sampled values of $Z_c(t_6) - Z_c(t_5)$ against the number of iterations in the MCMC chain between samples of τ_j.

process $\sum_{j=1}^{\eta} \tau_j$, where, for small ϵ, the number of jumps η is Poisson with mean

$$\lambda = \int_0^{\infty} \int_0^{\infty} \frac{e^{-z\beta(s)}\alpha(s)I(z > \epsilon, t_{i-1} < s < t_i)}{1 - e^{-z}} \, dz \, ds \quad (12)$$

and the jumps τ_j are independent and come from a density

$$g(z) \propto \frac{I(z > \epsilon)\int_{t_{i-1}}^{t_i} \exp(-z[\beta(s) + N - i + 1])d\alpha(s)}{1 - e^{-z}}. \quad (13)$$

A sample from $Z_c(t_1)$ is obtained by this definition if we let $t_0 = 0$. The approximation converges to $Z_c(t_i) - Z_c(t_{i-1})$ as $\epsilon \to 0$; practically, we can take ϵ to be small. In [23], the error in this approximation as a function of ϵ is discussed, which is of the order $E(\epsilon) < \epsilon\alpha(t_{i-1}, t_i)$. In this paper, following some exploration, we found that there was no difference in results for any $\epsilon < 10^{-5}$; thus, we set, conservatively, $\epsilon = 10^{-7}$.

In order to sample the compound Poisson process $\sum_{j=1}^{\eta} \tau_j$, we first sample λ, then sample η from the Poisson with mean λ, and then independently sample $\tau_1, \ldots, \tau_{\eta}$ from $g(z)$. Summing the τ_j obtains the approximate sample from $Z_c(t_i) - Z_c(t_{i-1})$. This is independently repeated for $i = 1, \ldots, k$ and also for $Z_c(T^*) - Z_c(t_k)$. For the latter case, (11), (12), and (13) are still valid with t_{i-1} replaced by t_k, t_i replaced by T^* and $N - i + 1$ (in (11) and (13)) replaced by $N - k + 1$. In [23], it is shown how to draw a sample from λ and from $g(z)$. For our specific case, one would do the following (which follows exactly the Appendix of that paper):

1. **Sampling λ for $Z_c(t_i) - Z_c(t_{i-1})$.** This is done by a Gibbs sampler on variables (z, s). Take initial values of the sampler to be $z^{(0)} > \epsilon$ and $t_{i-1} < s^{(0)} < t_i$; for sampling $Z_c(T^*) - Z_c(t_k)$, we have $t_k < s^{(0)} < T^*$.

Iteration $l + 1$ of the sampler that draws $(z^{(l+1)}, s^{(l+1)})$ given $(z^{(l)}, s^{(l)})$ is then:

a. Sample $z^{(l+1)}$ from an exponential distribution with mean $[\beta(s^{(l)}) + N - i + 1]^{-1}$ restricted to (ϵ, ∞). For sampling $Z_c(T^*) - Z_c(t_k)$, the mean is $[\beta(s^{(l)}) + N - k + 1]^{-1}$.

b. Sample $s^{(l+1)}$ by a Metropolis move. Propose a point s^* from the density $f(s) \propto \alpha(s)$ restricted to $t_{i-1} < s < t_i$. Let $s^{(l+1)} = s^*$ with probability the minimum of 1 and

$$\exp\left(-z^{(l+1)}\left[\beta(s^*) - \beta(s^{(l)})\right]\right);$$

otherwise, $s^{(l+1)} = s^{(l)}$.

c. Once many samples have been taken, the chain has converged, and L samples from the stationary distribution are taken, then

$$\lambda \approx \frac{1}{L}\sum_{l=1}^{L} \frac{1}{1 - \exp(-z^{(l)})} \times \int_{t_{i-1}}^{t_i} \frac{\exp(-\epsilon\beta(s))}{\beta(s)} d\alpha(s),$$

where the integral here is one-dimensional and should be sufficiently well approximated by the easier to compute $\int_{t_{i-1}}^{t_i} d\alpha(s)/\beta(s) - \epsilon\alpha(t_{i-1}, t_i)$. For sampling $Z_c(T^*) - Z_c(t_k)$, we replace t_{i-1} by t_k and t_i by T^*.

2. **Sampling from $g(z)$.** A sample z from $g(z)$ can be obtained by a Gibbs sampling scheme over variables (z, u, s). Initial values are $z^{(0)} > \epsilon$, $0 < u^{(0)} < 1$, and $t_{i-1} < s^{(0)} < t_i$; for sampling $Z_c(T^*) - Z_c(t_k)$ we have $t_k < s^{(0)} < T^*$. To draw $(z^{(l+1)}, u^{(l+1)}, s^{(l+1)})$ given $(z^{(l)}, u^{(l)}, s^{(l)})$:

a. Sample $z^{(l+1)}$ from an exponential distribution with mean $[\beta(s^{(l)}) + N - i + 1]^{-1}$ restricted to (ϵ, ∞) if $u^{(l)} \leq 1$, or restricted to

$$(\epsilon, -\log(1 - (u^{(l)})^{-1}),$$

if $u^{(l)} > 1$. For sampling $Z_c(T^*) - Z_c(t_k)$, the mean is

$$[\beta(s^{(l)}) + N - k + 1]^{-1}.$$

b. Sample $u^{(l+1)}$ from a uniform on

$$(0, (1 - \exp(-z^{(l+1)}))^{-1}).$$

c. Sample $s^{(l+1)}$ exactly as it was sampled above in the generation of λ.

d. Repeat until convergence has been reached, at which point $z^{(l)}$ is a sample from $g(z)$.

e. To obtain all η independent samples from $g(z)$, it is not necessary to start this Gibbs sampler again, rather just continue the chain and draw the remaining samples at suitably spaced intervals.

APPENDIX B

DETERMINING CHAIN LENGTH FOR SAMPLING OF $Z_c(t_i) - Z_c(t_{i-1})$

We use some data simulated from the Jelinski-Moranda model to illustrate how chain length is determined. Fig. 8a shows sampled values of λ (see Appendix A for its definition) for the sampling of $Z_c(t_{28}) - Z_c(t_{27})$, which show no relationship between MCMC chain length and the sampled values. In Fig. 8b, the value of $Z_c(t_6) - Z_c(t_5)$ is plotted against the number of iterations in the MCMC chain between samples of τ_j. Cases where $\eta = 0$ and hence no τ_j are sampled are omitted. Again no relationship is seen between the computed values and the chain length.

Given these plots, we adopt a chain of length 20,000 for λ and 10,000 iterations between samples of τ_j. This proved sufficient for the NTDS data. For the simulated Weibull data described in Section 5, the same analysis led to a chain of length 100,000 for λ and 10,000 for τ_j.

REFERENCES

[1] Z. Jelinski and P. Moranda, "Software Reliability Research," *Statistical Computer Performance Evaluation,* W. Freiberger, ed., Academic, 1972.

[2] D.R. Miller, "Exponential Order Statistic Models of Software Reliability Growth," *IEEE Trans. Software Eng.,* vol. 12, pp. 12-24, Dec. 1986.

[3] P.M. Nagel, F.W. Scholz, and J.A. Skrivan, "Software Reliability: Additional Investigations into Modeling with Replicated Experiments," NASA Contractor Report 172378, NASA Langley Research Center, 1984.

[4] G. Campbell and K.O. Ott, "Statistical Evaluation of Major Human Errors during the Development of New Technological Systems," *Nuclear Science Eng.,* vol. 71, pp. 267-279, 1979.

[5] H. Joe, "Statistical Inference for General Order Statistics and Nonhomogeneous Poisson Process Software Reliability Models," *IEEE Trans. Software Eng.,* vol. 15, no. 11, pp. 1485-1490, Nov. 1989.

[6] M. Barghout, A.A. Abdel-Ghaly, and B. Littlewood, "A Non-Parametric Order Statistics Software Reliability Model," *Software Testing, Verification and Reliability,* vol. 8, pp. 113-132, 1998.

[7] J. Bunge and M. Fitzpatrick, "Estimating the Number of Species: A Review," *J. Am. Statistical Assoc.,* vol. 88, no. 421, pp. 364-373, 1993.

[8] B. Littlewood and J.L. Verall, "A Bayesian Reliability Growth Model for Computer Software," *J. Royal Statistical Soc., Series C,* vol. 22, pp. 332-346, 1973.

[9] A.E. Raftery, "Inference and Prediction for a General Order Statistics Model with Unknown Population Size," *J. Am. Statistical Assoc.,* vol. 82, pp. 1163-1168, 1987.

[10] T.A. Mazzuchi and R. Soyer, "A Bayes Empirical-Bayes Model for Software Reliability" *IEEE Trans. Reliability,* vol. 37, no. 2, pp. 248-254, June 1988.

[11] A. Sofer and D.R. Miller, "A Nonparametric Software-Reliability Growth Model," *IEEE Trans. Reliability,* vol. 40, no. 3, pp. 329-337, Aug. 1991.

[12] M.A. El-Aroui and J.L. Soler, "A Bayes Nonparametric Framework for Software-Reliability Analysis," *IEEE Trans. Reliability,* vol. 45, no. 4, pp. 652-660, Dec. 1996.

[13] J.F. Lawless and C. Nadeau, "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics,* vol. 37, pp. 158-168, 1995.

[14] H.A. David and H.N. Nagaraja, *Order Statistics,* third ed. Wiley, 2003.

[15] B. Littlewood and J.L. Verall, "On the Likelihood Function of a Debugging Model for Computer Software," *IEEE Trans. Software Reliability,* vol. 30, pp. 145-148, 1981.

[16] F.J. Samaniego and S.P. Wilson, "Estimation Problems with the Jelinski-Moranda Software Reliability Model," Technical Report 05/01, Dept. of Statistics, Trinity College, Dublin, 2005.

[17] M.T. Rodríguez and M.P. Wiper, "Bayesian Inference for a Software Reliability Model Using Metrics Information," *Safety and Reliability: Towards A Safer World, Proc. European Safety and Reliability Conf. (ESREL '01),* E. Zio, M. Demichela, and N. Piccinini, eds., pp. 1999-2006, 2001.

[18] S. Campodónico and N.D. Singpurwalla, "Inference and Predictions from Poisson Point Processes Incorporating Expert Knowledge," *J. Am. Statistical. Assoc.,* vol. 90, pp. 220-226, 1995.

[19] T.S. Ferguson, "A Bayesian Analysis of Some Non-Parametric Problems," *Annals of Statistics,* vol. 1, pp. 209-230, 1973.

[20] S.G. Walker, P. Damien, P.W. Laud, and A.F.M. Smith, "Bayesian Nonparametric Inference for Random Distributions and Related Functions," *J. Royal Statistical Soc. B,* vol. 61, pp. 485-527, 1999.

[21] S.G. Walker and P. Muliere, "Beta-Stacy Processes and a Generalization of the Pólya-Urn Scheme" *Annals of Statistics,* vol. 25, pp. 1762-1780, 1997.

[22] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis.* Chapman and Hall, 1995.

[23] S.G. Walker and P. Damien, "A Full Bayesian Nonparametric Analysis Involving a Neutral to the Right Process," *Scandinavian J. Statistics,* vol. 25, pp. 669-680, 1998.

Simon P. Wilson received the PhD degree in stochastic modeling from George Washington University in 1993. He is a senior lecturer in the Department of Statistics, Trinity College, Dublin. He is a fellow of the Royal Statistical Society and an elected member of the International Statistical Institute.

Francisco J. Samaniego received the PhD degree in mathematics from UCLA in 1971. He is a professor of statistics at the University of California, Davis. He is a fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the Royal Statistical Society.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.