Utility Based Maintenance Analysis using a Random Sign Censoring Model

J. Andrés Christen\textsuperscript{a,*}, Fabrizio Ruggeri\textsuperscript{b}, Enrique Villa\textsuperscript{c}

\textsuperscript{a}CIMAT, A.P. 402, Guanajuato, GTO, 36000, Mexico.
\textsuperscript{b}CNR IMATI, Milano, Italy.
\textsuperscript{c}CIMAT, Guanajuato, Mexico.

Abstract
Industrial systems subject to failures are usually inspected when there are evident signs of an imminent failure. Maintenance is therefore performed at a random time, somehow dependent on the failure mechanism. A competing risk model, namely a Random Sign model, is considered to relate failure and maintenance times. We propose a novel Bayesian analysis of the model and apply it to actual data from a water pump in an oil refinery. The design of an optimal maintenance policy is then discussed under a formal decision theoretic approach, analyzing the goodness of the current maintenance policy and making decisions about the optimal maintenance time.

Keywords: Competing Risks, Bayesian Analysis, MCMC, Failure Times, Optimal Maintenance Time.

1. Introduction

Complex systems and components subject to failures are the object of preventive maintenance in order to avoid the more or less disruptive consequences of their failure. Examples range from nuclear power plants to heating systems and car tires. Both maintenance and failure have a cost, with the former being
less expensive in general. At the same time, policies based on too early main-
tenance could be unacceptable since there should be an excessive use of new
components well before failing of old ones and the continuity of services pro-
vided by system could not be guaranteed because of maintenance. Failure and
maintenance times are naturally intertwined and there is a quest for optimal
maintenance policies that act just before failures.

Different models have been proposed in the literature about optimal mainte-
nance policies, going back, for example, to Barlow and Hunter (1960). Here we
concentrate on a particular case, in which data are available as either failure or
maintenance time and maintenance is performed when some warnings denote a
possible incipient failure. Such situation arises quite naturally under a condition
based maintenance policy. As an example, described in D’Ippoliti and Ruggeri
(2008), a micrometer is inserted in a cylinder liner of a ship diesel engine to
to check its wear and decide about the liner replacement before its failure. Both
maintenance and failure are modelled by random variables, as in the Random
Sign model developed by Cooke (1996), which will be considered in this paper.
The model is a particular example of competing risks model, widely used in
reliability and survival analysis. A review about competing risks is provided by

The Random Sign model, illustrated in Section 2, considers the case of a
component whose failure time is subject to a possible right censoring, due to
maintenance; here censoring is supposed independent of the age $X$ at which
the component would expire but, given that the component is censored, the
censoring time may depend on $X$.

The Random Sign model seems suitable for analyzing data about an oil re-
finery water pump, considered in Pérez and Villa (2008). A limited number of
data are collected and the actual maintenance policy is unknown. At the same
time, company experts can provide opinions on the failure process of the pump
and the maintenance policy followed so far. Maintenance and repair costs can
be assessed as well, and combined with knowledge about the failure process
to develop optimal maintenance policies using a decision theoretic approach.
Therefore, a Bayesian analysis of the Random Sign model, novel in literature, is presented in Section 3, along with a utility based optimal maintenance policy; both are applied to the pump data in Section 4. Concluding remarks are presented in Section 5.

2. Random Sign model

Since Cox (1959), lack of identifiability of marginal distributions of some competing risk models has been discussed in the literature; we will not further discuss this problem and refer the interested reader to, for example, Bunea (2003). Cooke (1996) proposed the Random Sign censoring, which is probably the simplest model allowing for identifiability of marginal distributions. The model assumes that a component, which would fail at time $X$, could be subject to right censoring. The event that the lifetime could be censored is independent of the age $X$ but, once censoring occurs, then censoring time depends on $X$.

The typical situation modelled by Random Sign censoring is about censoring occurring when some warning about failing age $X$ is available. In particular, in our case study, censoring is due to maintenance and, therefore, an intervention occurs in advance to avoid failure once some warning denotes an incipient failure.

We present the definition of a Random Sign censoring model (see, for example, Bunea, Cooke and Lindqvist, 2003, p.359).

**Definition 1.** Given a random variable $X$, consider $Y = X - W\delta$, where $W$ is a random variable with $0 < W < X$ and $\delta = \{-1, 1\}$ is also a random variable independent of $X$. The variable $Z \equiv \min\{X, Y\}, 1(Y < X)$, with $1(\cdot)$ denoting the set function, is called a random sign censoring of $X$ by $Y$.

Bunea et al. (2003, p.359) present an alternative definition using the indicator function $1(X < Y)$; our definition above should lead to no confusion.

3. Bayesian analysis

We consider the Random Sign censoring model given in Definition 1 and denote maintenance and failure times by $Y$ and $X$, respectively. Unlike the more
standard notation used in Definition we take $\delta = 2\epsilon - 1$ so that $\epsilon = \{0,1\} \text{ (failure and maintenance, respectively).}$ Then $\epsilon$ and $X$ are independent and $X$ and $Y$ are related by

$$Y = X - (2\epsilon - 1)W,$$  \hspace{0.5cm} (1)

$0 < W < X$. Let $T = \min(X,Y)$ and note that $Z = (T, \epsilon)$. It follows that

$$f(T = t, \epsilon = 0) = f(X = t, \epsilon = 0) = f(X = t)f(\epsilon = 0),$$  \hspace{0.5cm} (2)

given the assumed independence of $\epsilon$ and $X$. In this case $W$ takes any value in $(0, t)$.

Conversely, censoring implies

$$f(T = t, \epsilon = 1) = f(Y = t, \epsilon = 1) = f(Y = t|\epsilon = 1)f(\epsilon = 1)$$

$$= f(\epsilon = 1) \int f(X = t + w, W = w)dw$$

$$= f(\epsilon = 1) \int f(X = t + w|W = w)f(W = w)dw,$$  \hspace{0.5cm} (3)

since $X = Y + W$, with $Y = t$ and $W > 0$. Distributions on $\epsilon$ and $X$ are to be specified for both cases, whereas the distribution on $W$ has to be specified only for the censoring case. In this case $W$ has the meaning of the “time elapsed between the censoring (maintenance) time and the failure time which would have been observed if no censoring had been performed”. Now we present a particular (parametric) proposal for these distributions.

### 3.1. Models for failure and censoring times

We consider a Gamma model $G(\alpha, \lambda)$ for the failure time $X$, with $\alpha, \lambda > 0$. This is a standard failure time model with sufficient flexibility, and adequate when light tails are expected. We now turn to defining the distribution for $W|X = x$. This distribution should have support in $[0, x]$. Moreover, since we expect $W$, the elapsed time between maintenance and (unobserved) failure, to be definitely closer to 0 rather than to its largest value $X = x$, we should have that this conditional distribution is decreasing. A reasonable assumption would be

$$f(W = w|X = x) = \frac{\alpha - 1}{x} \left( \frac{w}{x} \right)^{\alpha-2} f^{(w)}_{(0,x)},$$  \hspace{0.5cm} (4)
a truncated power law. We choose $1 < \alpha \leq 2$, so that it is non-increasing.

The shape of this distribution is not dependent on $x$ and is solely governed by the parameter $\alpha$.

Under such assumptions on the distributions of $X$ and $W|X$, it can be easily shown that, marginally, $W \sim G(\alpha - 1, \lambda)$, $Y|\epsilon = 1 \sim E(\lambda)$ and $f(X = x|W = w) = \lambda \exp\{-\lambda(x - w)\}I_{(w, \infty)}(x)$. As a special case, $\alpha = 2$ implies $X \sim G(2, \lambda)$, $W|X = x \sim U(0, x)$ and $W \sim E(\lambda)$. Regarding the censoring mechanism, we take $\epsilon \sim Ber(\theta)$. This model, arising from the Gamma model for failure times (restricting $\alpha \in (1, 2]$) and the truncated power law model for $W|X$, leads to a setting where all marginals are well defined, having known distributions, and is governed by the parameters $(\theta, \lambda, \alpha)$, providing reasonable flexibility for typical applications.

3.2. Prior choice

We consider the parameter vector $\eta = (\theta, \lambda, \alpha)$ and we take independent priors $\theta \sim Be(\alpha_1, \beta_1)$, $\lambda \sim G(\alpha_2, \beta_2)$ and $\alpha \sim U(1, 2)$.

The choice of the hyperparameters is a complex aspect in Bayesian analysis and it is important to explore some features of the involved random variables, especially the observable ones, to specify their values. It is worth observing that $X \sim G(\alpha, \lambda)$ and $Y \sim E(\lambda)$ imply that $E(X) = \frac{\alpha}{\lambda}$ and $E(Y) = \frac{1}{\lambda}$, so that $\lambda = \frac{1}{E(Y)}$ and $\alpha = \frac{E(X)}{E(Y)}$. The choice of the hyperparameters $\alpha_2$ and $\beta_2$ could be performed noting that

$$E(X) = E[E(X|\alpha, \lambda)] = E(\alpha)E\left(\frac{1}{\lambda}\right) = E(\alpha)\frac{\beta_2}{\alpha_2 - 1}.$$  

(5)

Similarly, using $Var(X) = E[Var(X|\alpha, \lambda)] + Var[E(X|\alpha, \lambda)]$, it can be proved that

$$C_X^2 = \frac{1}{\alpha_2 - 2} \left\{ (\alpha_2 - 1)(\mu_\alpha^{-1} + C_\alpha^2) + 1 \right\},$$  

(6)

where $C_X$ and $C_\alpha$ are the coefficients of variation (standard deviation over the mean) of $X$ and $\alpha$, and $\mu_\alpha = E(\alpha)$. To ensure the existence of $E(X)$ and $C_X^2$ we need $\alpha_2 > 2$; furthermore, (5) and (6) imply $\alpha_2 = 1 + \frac{C_X^2 + 1}{C_X^2 - (\mu_\alpha^{-1} + C_\alpha^2)}$ and
\( \beta_2 = \mu_\alpha^{-1} E(X)(\alpha_2 - 1) \). Both \( \alpha_2 \) and \( \beta_2 \) are positive if

\[
C_X > \sqrt{\mu_\alpha^{-1} + C_\alpha^2}.
\]

The last condition is a property of our model, and for the uniform prior assumed for \( \alpha \) entails to \( \mu_\alpha = 3/2, \) \( C_\alpha^2 = 1/27 \) and thus \( C_X > 0.8388 \); the a priori coefficient of variation for \( X \) cannot be lower than 0.8388. If we elicit an a priori expected value \( m_X \) and a standard error \( s_X \) for \( X \) (e.g. arising from the manufacturer’s specifications for continuous operation of the system), we take \( C_X = s_X/m_X \) to obtain \( \alpha_2 \) and \( \beta_2 \) as above.

Regarding the choice of parameters for the Beta prior on \( \theta \), they could be chosen specifying some quantiles or by observing that \( E(\theta|\alpha_1, \beta_1) = \frac{\alpha_1}{\alpha_1 + \beta_1} \) and \( \text{Var}(\theta|\alpha_1, \beta_1) = \frac{\alpha_1 \beta_1}{(\alpha_1 + \beta_1)^2 (\alpha_1 + \beta_1 + 1)} \). When scarce prior information is available about \( \theta \), a default non-informative prior may be stated as \( \alpha_1 = \beta_1 = 0.5 \) (see Bernardo and Smith, 1994).

3.3. Likelihood

Suppose data are given by \( (t, \epsilon) = \{(t_i, \epsilon_i)\}_{i=1}^n \); therefore the likelihood would be

\[
f(t, \epsilon | \eta) = \prod_{i=1}^n f(t_i, \epsilon_i | \eta),
\]

where \( f(t_i, \epsilon_i | \eta) \) would be given by (2) and (3), depending on \( \epsilon = 0 \) or 1, respectively. The integrals involved in (3) are an evident drawback of the approach using the full likelihood (7). Therefore, we introduce latent variables \( w_i \), with the same meaning as in (1), for the censored data and deal with the conditional likelihood

\[
f(t, \epsilon | \eta, w) = \prod_{i=1}^n f(t_i | \epsilon_i, \alpha, \lambda, w_i) f(\epsilon_i | \theta)
\]

where \( w = \{w_i\}_{i: \epsilon_i = 1} \) and there is no dependence on \( w_i \) when \( \epsilon_i = 0 \).

We get \( f(\epsilon_i | \theta) = \theta^{\epsilon_i} (1 - \theta)^{1-\epsilon_i} \). For \( f(t_i | \alpha, \lambda, w_i, \epsilon_i = 0) = f_X(t_i | \epsilon_i = 0) = \frac{\lambda^\alpha}{\Gamma(\alpha)^i \tau^{\alpha-1} \exp{-\lambda t_i}} \), we get the Gamma model, that does not depend on \( w_i \). It follows that \( \int f(t_i | \alpha, \lambda, w_i, \epsilon_i = 1) f(w_i | \alpha, \lambda, \epsilon_i = 1) dw_i = f_Y(t_i | \epsilon_i = 1) = \)
Therefore, the likelihood is

$$f(t, \xi | \eta) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \left\{ \frac{(\lambda t_i)^{\alpha-1}}{\Gamma(\alpha)} \right\}^{1-\epsilon_i} I_{(0,\infty)}^\alpha(\theta^\epsilon, (1-\theta)^{1-\epsilon_i}).$$  \hspace{1cm} (9)

### 3.4. Posterior distribution

Based on the priors and the likelihood given in Sections 3.2 and 3.3 respectively, then the posterior distribution is given by

$$f(\eta | t, \xi) \propto f(t, \xi | \eta) f(\eta)$$

$$\propto \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \left\{ \frac{\lambda^{\alpha-1} t_i^{\alpha-1}}{\Gamma(\alpha)} \right\}^{1-\epsilon_i} \theta^\epsilon(1-\theta)^{1-\epsilon_i} \lambda^{\alpha-1} e^{-\lambda t_i + \beta_2 \theta \alpha_2 - 1} \theta^{\beta_1 - 1}$$

$$\propto \lambda^{\alpha_2 + (n - \sum \epsilon_i) \alpha + \sum \epsilon_i - 1} e^{-\lambda (\beta_2 + \sum t_i)} \prod_{i=1}^{n} \left( \frac{t_i^{\alpha}}{\Gamma(\alpha)} \right)^{1-\epsilon_i} \theta^{\alpha_1 + \sum \epsilon_i - 1} (1-\theta)^{\beta_1 + n - \sum \epsilon_i - 1} \beta e^{(\epsilon)(1-\theta)} f(\lambda)(\alpha)^{(1-2)} I(\lambda)(\alpha)^{(1-2)}.$$  \hspace{1cm} (10)

For simplicity, we avoid writing the indices, $i = 1, \ldots, n$, over $\sum \epsilon_i$ and $\sum t_i$.

All marginal posterior distributions cannot be obtained in closed form. However, by straightforward marginalizations, we see that

$$[\theta | t, \xi] = [\theta | \xi] \sim Be \left( \alpha_1 + \sum \epsilon_i, \beta_1 + n - \sum \epsilon_i \right),$$

$$f(\alpha | t, \xi) \sim \frac{\Gamma(\alpha_2 + (n - \sum \epsilon_i) \alpha + \sum \epsilon_i) \left\{ \frac{\Pi_{i=1}^{n} t_i^{1-\epsilon_i}}{\Gamma(\alpha)^{n-\sum \epsilon_i}} \right\}^{\alpha}}{\beta e^{(\epsilon)(1-\theta)} f(\lambda)(\alpha)^{(1-2)} I(\lambda)(\alpha)^{(1-2)}}$$

$$\lambda | \alpha, t, \xi \sim G \left( \alpha_2 + (n - \sum \epsilon_i) \alpha + \sum \epsilon_i, \beta_2 + \sum t_i \right).$$

A simple MCMC (Markov chain Monte Carlo) algorithm is devised to sample from the posterior marginal distribution of $\alpha, f(\alpha | t, \xi)$. Using an independent proposal taking the same form for $f(\alpha | t, \xi)$ but removing the $\Gamma$ terms, we obtain a distribution proportional to $e^{\alpha D} I_{(1,2)}^{(\alpha)}$, where $D = -(n - \sum \epsilon_i) \log(\beta_2 + \sum t_i) + \sum (1 - \epsilon_i) \log(t_i)$. This distribution may be easily normalized and its cdf inverted to obtain a straightforward algorithm to simulate from the proposal; the resulting Metropolis-Hastings (MH) ratio is the ratio of the remaining $\Gamma$ terms. This MH setting is quite efficient. Once removing the burn-in and using a suitable thinning, we obtain a quasi-independent sample for the marginal posterior of
\( \alpha \). This sample is in turn “plugged” in the posterior conditional distribution \([\lambda|\alpha, t, \xi]\) to sample from the posterior marginal of \( \lambda \) (and accordingly from the joint posterior of \( \alpha \) and \( \lambda \)). \( \theta \) is independent of \( \alpha \) and \( \lambda \) given the data and may be easily sampled.

3.5. Predictive distribution

The Bayesian approach allows for simple, but very useful, predictions of future observations. In particular we are interested in predicting the next failure time, if the next observation will be censored or not, and the elapsed time between maintenance and failure if the former occurs first.

Considering \( f(\epsilon_{n+1}|t, \xi) = \int f(\epsilon_{n+1}|\theta)f(\theta|\xi)d\theta \), it follows that

\[
\begin{align*}
    f(\epsilon_{n+1}|t, \xi) &= \begin{cases} \\
        \frac{\beta_1 + n - \sum \epsilon_i}{\alpha_1 + \beta_1 + n} & \epsilon_{n+1} = 0 \\
        \frac{\alpha_1 + \sum \epsilon_i}{\alpha_1 + \beta_1 + n} & \epsilon_{n+1} = 1.
    \end{cases}
\end{align*}
\]

(12)

Since \( w_{n+1}|\lambda, \alpha \sim \mathcal{G}(\alpha - 1, \lambda) \), the predictive distribution of the next elapsed time between maintenance and failure is given by

\[
    f(w_{n+1}|t, \xi) = \int \frac{\lambda^{\alpha-1}}{\Gamma(\alpha - 1)} w_{n+1}^{\alpha-2} e^{-\lambda w_{n+1}} f(\lambda, \alpha|t, \xi) d\lambda d\alpha.
\]

(13)

The above integral does not have a closed form. However, we may easily generate samples from this predictive distribution by simulating \( w_{n+1}^{(j)} \sim \mathcal{G}(\alpha^{(j)} - 1, \lambda^{(j)}) \), where \( \{\lambda^{(j)}, \alpha^{(j)}\}_{j=1}^M \) is a sample from the posterior distribution of \((\lambda, \alpha)\) (i.e. “plugging in” the simulated samples for \( \lambda \) and \( \alpha \) described in the previous Section). Similarly, we may simulate from \( X_{n+1}|\epsilon_{n+1}, t, \xi, X > t \), i.e. the unconditional and conditional posterior predictive distributions of \( X_{n+1} \).

The above quantities are important to predict if, without any external intervention on the maintenance policy, maintenance will occur before failure and how close these events will appear. Therefore high values of both \( f(\epsilon_{n+1} = 1|t, \xi) \) and \( f(w_{n+1} \approx 0|\epsilon_{n+1} = 1, t, \xi) \) are desirable. The situation of an external intervention on the next maintenance time, based upon past data, will be discussed in Section 3.6.
3.6. Optimal policies

We are now considering the search for optimal policies, under two different sets of conditions, to decide upon the next maintenance time. According to a widely accepted principle in Bayesian decision analysis, we choose the policy which minimizes the expected loss under adequate loss function(s) and the posterior distributions obtained earlier.

First, we would like to compare the current maintenance policy based on early warnings with the one suggesting an intervention at some time $t$. We assume there is a repair cost, $C_R$, and a maintenance cost, $C_M$, such that $C_M < C_R$, and we want to compare the expected cost under the two policies. Current maintenance practices lead to an expected cost

$$E_C = C_R P(\epsilon_{n+1} = 0|\xi, \epsilon) + C_M P(\epsilon_{n+1} = 1|\xi, \epsilon), \quad (14)$$

where $P(\epsilon_{n+1} = 0|\xi, \epsilon)$ and $P(\epsilon_{n+1} = 1|\xi, \epsilon)$ are given by (12). The expected cost of a maintenance performed at time $t$ is given by

$$E_t = C_R P(X_{n+1} \leq t|\xi, \epsilon) + C_M P(X_{n+1} > t|\xi, \epsilon), \quad (15)$$

where the predictive posterior probabilities can be easily obtained, as described in Section 3.5. It can be shown that $E_C = E_t$ for the critical time $t = t^c$ such that $P(X_{n+1} \leq t^c|\xi, \epsilon) = P(\epsilon_{n+1} = 0|\xi, \epsilon)$. The result holds regardless of the values of $C_R$ and $C_M$. Therefore, maintenance at any time $t < t^c$ should improve upon the current maintenance policy under the expected cost criterion.

Most approaches to maintenance policies follow a non-formal decision theoretic approach (see for example, Ferreira, Teixeira de Almeida and Cavalcante, 2009; Rezg, Dellagi and Chelbi, 2008; Bris, Châtelet and Yalaoui, 2003; to mention some recent references) and we expect that a great deal of clarity and optimality will be gained by following a coherent framework for rational decision making (see DeGroot, 1970). This strategy would required a loss function $L(t, X_{n+1})$ measuring the loss of a maintenance policy at time $t$ and a failure time $X_{n+1}$. Besides the costs $C_M$ and $C_R$, we would also like to consider the “down time” of the system due either to maintenance, $d_M$, or failure, $d_R$ (as
a simplification we consider both times as non-random). Maintenance policies may be very complex and may consider many characteristics of the problem. We follow this simple approach using only the costs and down times and this will be satisfactory to tackle the example presented in Section[J]. We also believe that many maintenance problems in simple systems in industry will have similar characteristics and thus prone to be analyzed with this model. As prescribed by the theory, after defining a loss function, we would use the posterior predictive distribution of \( X_{n+1} \) to find the expected loss function \( L^*(t) \) for any \( t \) and the optimal maintenance policy would be \( t^* = \arg \min L^*(t) \).

It is clear that, if provided with a maintenance time \( t \) and the actual next failure time \( X_{n+1} \), the cost incurred would be

\[
C(t, X_{n+1}) = C_M 1(X_{n+1} > t) + C_R 1(X_{n+1} \leq t).
\]

In the same case, the time elapsed between the start of the system plus the down time due to maintenance or failure may be computed as

\[
R(t, X_{n+1}) = (t + d_M) 1(X_{n+1} > t) + (X_{n+1} + d_R) 1(X_{n+1} \leq t);
\]

this is known as the “renovation time” of the system. The overall cost per unit of time should be \( C(t, X_{n+1})/R(t, X_{n+1}) \). This is basically the age based approach to maintenance. Traditionally, the average total cost and average renovation times are used to obtain an objective function; that is \( E\{C(t, X_{n+1})\}/E\{R(t, X_{n+1})\} \) which is in turn is minimized (see, for example, Ferreira et al., 2009; Cheng and Chen, 2008; Bris, Châtelet and Yalaoui, 2003; Lawless, 1998; Nguyen and Murthy, 1981, and Barlow and Hunter, 1960). The decision theoretic approach would require to minimize the expected cost per time unit, that is \( E\{C(t, X_{n+1})/R(t, X_{n+1})\} \) which in general is not equal to \( E\{C(t, X_{n+1})\}/E\{R(t, X_{n+1})\} \), although commonly the latter is easier to calculate than the former. Fortunately, in our case the former expectation is relatively simple to compute. Indeed, the same decision is reached if we divide by \( C_R \) and then our stipulated loss function is

\[
L(t, X_{n+1}) = (t + d_M)^{-1} \frac{C_M}{C_R} 1(X_{n+1} > t) + (X_{n+1} + d_R)^{-1} 1(X_{n+1} \leq t). \tag{16}
\]
We integrate (16) with respect to the posterior predictive distribution and we get the posterior predictive expected loss is

\[ L^*(t) = \frac{C_M}{C_R} (t + d_M)^{-1} S(t) + g(t) \]

where \( S(t) = 1 - \int_0^t f_{X_{n+1}}(x|t, \epsilon) dx \) is the posterior predictive survival function and \( g(t) = \int_0^t (x + d_R)^{-1} f_{X_{n+1}}(x|t, \epsilon) dx \). By taking the derivative of the expected loss, it may be seen that \( L^*(t) \) is decreasing (increasing) if \((t + d_M) \lambda(t) < (>) (\frac{C_R}{C_M} \frac{t + d_M}{t + d_R} - 1)^{-1}\),

where \( \lambda(t) = \frac{f_{X_{n+1}}(t|t, \epsilon)}{S(t)} \); that is, the posterior predictive hazard function. Assuming \( \lambda(t) \) is nondecreasing (therefore justifying the need for a maintenance policy), \((t + d_M) \lambda(t)\) is monotonically increasing, then the optimal maintenance time \( t^* \) must be such that \((t^* + d_M) \lambda(t^*) = (\frac{C_R}{C_M} \frac{t^* + d_M}{t^* + d_R} - 1)^{-1}\). It is interesting to analyze cases where \( d_M \ll d_R \), resulting in a non-negligible term \( \frac{t + d_M}{t + d_R} \). This term acts as a “discount” on the cost ratio \( \frac{C_R}{C_M} \) leading to shorter maintenance policies. However, from now on, we will only consider the case \( d_M \approx d_R = d \) and therefore \((t^* + d) \lambda(t^*) \approx (\frac{C_R}{C_M} - 1)^{-1}\).

For example, it is possible to understand how the proposed method works when considering a simple exponential failure time model \( \mathcal{E}(\gamma) \). In that case the hazard function is \( \lambda(t) = \gamma \) and therefore \( t^* = \gamma^{-1} (\frac{C_R}{C_M} - 1)^{-1} \), considering that the down time \( d \) is negligible in relation to \( \gamma^{-1} \). We see then that \( t^* \) is the expected failure time multiplied by the correcting factor \( (\frac{C_R}{C_M} - 1)^{-1} \). In Figure 1 we present a plot of this correcting factor as a function of the ratio \( \frac{C_R}{C_M} \). When the cost of repair is only 20% more than the cost of maintenance we should wait 5 times the expected failure time. Maintenance is recommended at the expected failure time when the cost of repair is twice the cost of maintenance. For cases when \( C_R \) is more than, say, 10\( C_M \), the correcting factor is basically \( \frac{C_M}{C_R} \). If for example \( C_R = 100 C_M \) (i.e. very costly repair), then we should have a maintenance at 1% of the expected failure time.
4. Example

We analyze operation data of a water pump (Worthington make, 57 l/min at 1,760 RPM). The pump operates 24 hours a day and it stops only for servicing or failure. The pump operates at the PEMEX Salamanca refinery, in a Diesel processing plant, in Guanajuato, Mexico. Maintenance was done using previous experience with an heuristic approach. It is not clear whether 1) maintenance had some effect on the operation performance of the pump and 2) which maintenance policy was followed.

Table 1 shows the $n = 34$ failure and maintenance times which were collected for the water pump. The label 0 denotes failure times whereas 1 is used when a maintenance was performed before a failure; i.e. we are depicting realizations of the variable $Z$ in Definition 1.
4.1. Results

It was established that a priori the pump was expected to work on average for approximately one week, with one week as standard deviation. Applying the method explained in Section 3.2, this prior opinion lead to the choices $\alpha_2 = 6$ and $\beta_2 = 1,680$. The default non-informative prior for $\theta$ was used (Beta with parameters $\alpha_1 = \beta_1 = 0.5$); it is difficult to exploit prior information regarding the probability of which comes first between maintenance and failure, since the maintenance policy, if any, was not disclosed to us.

In Figure 2 we present the marginal posterior distributions of $\theta$, $\lambda$ and $\alpha$ and also the posterior predictive distribution of failure times, see Figure 2(d). The last distribution represents our current state of knowledge, based on all available data (censored or not) and our prior information, about the distribution of failure times of the system. The critical time $t^c \approx 57$ hours, explained in Section 3.6, is also depicted in Figure 2(d). Thus any maintenance policy before 57hr will improve on current practices.

Regarding an optimal maintenance policy, we consider the loss function presented in (16). We approximated the function $(t + d)\lambda(t) = \frac{f(t)}{S(t)}$ using the predictive sample for $X_{n+1|t,\epsilon}$. A kernel density estimation is used (implemented in a standard function in R; R Development Core Team, 2008) to estimate $f(t)$ at a fine grid of points $t_i$; the survival function is estimated with $S(t_i) \approx \frac{\left|\left\{j: X_{n+1}^{(j)}>t_i, j=1,2,\ldots,M\right\}\right|}{M}$, where $X_{n+1}^{(j)}$ are the simulated values of $X_{n+1|t,\epsilon}$. $(t_i, (t_i + d)\frac{f(t_i)}{S(t_i)})$ may be seen in Figure 3.
Figure 2: Marginal posterior distributions, with their prior densities in grey, for (a) $\theta$, (b) $\lambda$ and (c) $\alpha$. The solid line density in (c) is obtained by numerically normalizing (11). (d) Posterior predictive distribution for the failure time, that is $f_{X_{n+1}}(x|\xi)$; the critical time $t^* \approx 57$ hours, $P(X_{n+1} \leq t^*|\xi) = P(\epsilon_{n+1} = 0|\xi) = 0.81$, is depicted with a tick mark. All histograms are obtained using the sampling procedure described in Section 3.4.
A record of maintenance and failure repair costs was obtained by the involved engineers. Only the average costs are relevant, as far as our loss function is concerned, leading to an average ratio \( \frac{C_R}{C_M} = 4.46 \). The down time was considered as \( d = 1 \) hour. With these settings, we obtain the optimal maintenance policy \( t^* \approx 8 \) hours, see Figure 3.

When maintenance was performed, this was done at an average time of approximately 12 hours of operation. In this case a lower interval is suggested (8 hours). The expected time of the predictive distribution for \( X_{n+1} \) is approximately 34 hours. Although the underline model is not exponential, we see that \( 34/(4.46 - 1) = 9.8 \) is an approximate value for the optimal value \( t^* \), see Section 3.6. The optimal maintenance time, \( t^* = 8 \), is less than the average observed maintenance time (see Table 1). This may be explained as the result of the high repair to maintenance cost ratio \( (C_R/C_M = 4.46) \) reported by the engineers, thus making shorter maintenance times more cost-effective. However, if the cost ratio is reduced to 2, a longer optimal maintenance is obtained (c. 32 hours). It is also interesting to see from Figure 3 that our optimal maintenance policy of \( t^* = 8 \) hours coincides with the shortest maintenance time; thus a short maintenance time is not in complete contradiction with the engineer’s practices.

Note how Figure 2(d) clearly depicts what is currently known about continuous operation time for the pump. Moreover, in Figure 4 we plot the posterior predictive survival function (of \( X_{n+1} \)), along with the Kaplan-Meier estimator with some error bands, for comparisons. From Figure 4 we see that the predictive survival function is more optimistic when compared to the Kaplan-Meier estimator. Partially, this should be a result of the prior information included in the analysis, which is in turn rather optimistic.

Note that our optimal policies are based on the predictive distribution of failure. Predictive distributions are seldom used in the Bayesian analysis of reliability data. Usually, loss functions \( L(\theta, a) \) are considered, with \( \theta \) parameter and \( a \) action, and their posterior expected loss is used to determine the optimal action, namely maintenance time in our case. We prefer to take an approach
Figure 3: Relevant part of the estimated function $(t+d)\lambda(t)$, with down time $d = 1$ hour (black curve). The nonlinear left axis represents the ratio $\frac{CR}{CM}$. The assumed repair-to-maintenance cost ratio $\frac{CR}{CM} = 4.46$ is shown with the horizontal grey line. The optimal maintenance time $t^* \approx 8$ hours is depicted with a long tick mark in the x axis. The data (times $< 30$ hours) are plotted adding some jitter, with bullets for the failure times and circles for the maintenance times.
Figure 4: Posterior predictive survival function (black) and the Kaplan-Meier estimator with error bands (grey), for the water pump data analyzed. We used the `Surv` function of the `survival` R package with default settings.
based on observable quantities, i.e. failure times, instead of parameters. We
are aware of the difficulties in obtaining and using reliable information from
experts. As discussed earlier, we tried to elicit information used to specify
priors by using some physical properties of the models; we did the same about
the loss function, specified in terms of failure and maintenance times, for which
cost estimation was rather straightforward as well as the use of the posterior
predictive distribution of failure times.

5. Conclusions

We have performed a Bayesian analysis of the Random Sign censoring model,
which, despite its simple formulation, proved to be very useful in describing the
problem at hand, in which gaining insight on failure and maintenance mecha-
nisms is the goal. We chose a rather flexible, Gamma model for the failure time,
and a reasonable model for the time elapsed between maintenance and failure,
given the latter. The proposed model and priors allow either for closed form
computations or for simple MCMC simulations. Being parameter elicitation
one of the most critical issues in practical Bayesian analysis, our choice of mod-
els and priors allows for a rather user-friendly choice of the hyperparameters,
which are often related to properties of observable quantities, like the failure
time. This is a novel approach to competing risks and Random Sign modeling,
representing a relatively simple alternative that works and performs well in the
analyzed example.

Moreover, an integrated decision theoretical approach to designing a main-
tenance policy is presented. Only the expected ratio of repair vs. maintenance
cost is needed (also the repair and maintenance down times) to establish
the renovation time loss function to obtain the optimal maintenance time. A
simple numerical procedure is developed to minimize the loss function. All
software is available in a simple R program and may be downloaded from
\url{www.cimat.mx/~jac/software.html}.

It seems that predictive distributions are not widely used in (Bayesian) relia-
bility studies; as an example, posterior expected losses are more commonly used than predictive posterior expected losses. In our case study, the loss function for maintenance policies is based on the predictive distribution, with a clear interpretation in terms of the proportion of time the system is back up after an event (either maintenance or failure). We think that decisions should be based upon observable quantities, whose behavior is expressed through predictive distributions, rather than on distributions on parameters.

6. Acknowledgments

This work was partially funded by a joint CNR (Italy) and CONACYT (Mexico) grant.

References


Bris, R., Châtelet, E. and Yalaoui, F. (2003), “New method to minimize the preventive maintenance cost of seriesparallel systems”


Crowder, M.J. (2001), Classical competing risks, Chapman & Hall, Boca Raton, FL, USA.


