CHAPTER 1

STATISTICAL MODELLING AND INFERENCE FOR COMPONENT FAILURE TIMES UNDER PREVENTIVE MAINTENANCE AND INDEPENDENT CENSORING

Bo Henry Lindqvist and Helge Langseth
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
E-mail: {bo,helgel}@math.ntnu.no

Consider the competing risks situation for a component which may be subject to either a failure or a preventive maintenance action, where the latter will prevent the failure. It is then reasonable to expect a dependence between the failure mechanism and the PM regime. The chapter reconsiders the so called repair alert model which is constructed for handling such cases. A main ingredient here is the repair alert function which characterizes the "alertness" of the maintenance crew. The main emphasis of the chapter is on statistical inference for the model, based on possibly right censored data. Both nonparametric and parametric inference is studied. The methods are applied to two different data sets.

1. Introduction

We consider the competing risks situation occurring when a potential component failure at some time X may be avoided by a preventive maintenance (PM) at time Z. The experienced event will in this case be at time $Y = \min(X, Z)$, and it will either be a failure or a PM. It is convenient to use the notation $\delta = I(Z < X)$ to denote the type of event, where $I(A)$ is the indicator function of the event A. Thus $\delta = 0$ means that the component fails and $\delta = 1$ means that it is preventively maintained.

The observable result is now the pair (Y, δ), rather than the underlying times X and Z, which will often be the times of interest. For example, knowing the distribution of X would be important as a basis for maintenance optimization. It is well known[13,4], however, that in a competing risks case as described here, the marginal distributions of X and Z are not identifiable from observation of (Y, δ) alone unless specific assumptions are made

on the dependence between X and Z. The most used assumption of this kind is to let X and Z be independent, in which case identifiability follows. This assumption is not reasonable in our application, however, since the maintenance crew is likely to have some information regarding the component's state during operation. This insight is used to perform maintenance in order to avoid component failures. We are thus in practice usually faced with a situation of dependent competing risks between X and Z.

Lindqvist, Støve and Langseth[10] suggested a model called *the repair alert model* for describing the joint behavior of failure times X and PM-times Z. This model is a special case of *random signs censoring*, Cooke[2,3], under which the marginal distribution of X is identifiable. Recall that Z is said to be a random signs censoring of X if the event $\{Z < X\}$ is stochastically independent of X, i.e. if the event of having a PM before failure is not influenced by the time X at which the component fails or would have failed without PM. The idea is that the component emits some kind of signal before failure, and that this signal is discovered with a probability which does not depend on the age of the component. The repair alert model extends this idea by defining in addition a repair alert function which describes the "alertness" of the maintenance crew as a function of time.

The main emphasis of the present chapter is on statistical inference for the repair alert model. It will be assumed that data are available for a sample of N independent observations of (Y, δ), which may be right censored. In the case of censoring we only know that Y is greater than the censoring time, but do not know the type of event (failure of PM) that would have been eventually experienced. Independent censoring will be assumed in this case. This assumption is reasonable in many cases and is needed to identify the distribution of (Y, δ) and hence the distribution of X under random signs censoring. The ability to handle censored data is important for practical applications, and this is the main motivation for the present chapter.

Two examples will be given: In the first example we reconsider the data given by Mendenhall and Hader[11]. These data are type I censored at a fixed time τ, but were for illustrative purposes analyzed in Lindqvist et al.[10] without taking these censorings into account.

The second example is based on data from the OREDA database[12] and are also considered by Langseth and Lindqvist[8]. The component failures can in this example be due to several different failure modes. We study one of the failure modes with respect to failure time X and PM-time Z, while treating failure and PM events for the other failure modes as censorings.

2. Notation, definitions and basic facts

We assume that (X, Z) is a pair of continuously distributed life variables, with the properties that $P(X = Z) = 0$ and $0 < P(Z < X) < 1$. The cumulative distribution functions of X and Z are, respectively, $F_X(t) = P(X \leq t)$ and $F_Z(t) = P(Z \leq t)$.

Now let (Y, δ) define a competing risk case between X and Z. Here $Y = \min(X, Z)$ and $\delta = I(Z < X)$. The distribution of (Y, δ) is characterized by the subdistribution functions of X and Z, defined respectively by $F_X^*(t) = P(X \leq t, X < Z) \equiv P(Y \leq t, \delta = 0)$ and $F_Z^*(t) = P(Z \leq t, Z < X) \equiv P(Y \leq t, \delta = 1)$. Note that the functions F_X^* and F_Z^* are nondecreasing with $F_X^*(0) = F_Z^*(0) = 0$ and $F_X^*(\infty) + F_Z^*(\infty) = 1$. Any pair of functions K_1, K_2 satisfying these conditions, will be referred to as a subdistribution pair.

We next define the conditional distribution functions of X and Z respectively by $\tilde{F}_X(t) = P(X \leq t | X < Z)$ and $\tilde{F}_Z(t) = P(Z \leq t | Z < X)$. Note that $\tilde{F}_X(t) = F_X^*(t)/F_X^*(\infty)$, $\tilde{F}_Z(t) = F_Z^*(t)/F_Z^*(\infty)$.

For convenience we assume the existence of densities corresponding to each of the functions defined above, i.e. $f_X(t) = F_X'(t)$, $f_X^*(t) = F_X^{*\prime}(t)$, $\tilde{f}_X(t) = \tilde{F}_X'(t)$, and similarly for Z.

It follows by definition that the subdistribution functions F_X^* and F_Z^* are identifiable from observation of (Y, δ). In practice this means that if an infinite sample of (Y, δ) is available, then we can estimate the subdistribution functions without error. On the other hand, the marginal distribution functions F_X and F_Z are not identifiable in this manner from observation of (Y, δ)[13,4]. Thus, even with an infinite sample of (Y, δ) we are unable to estimate F_X and F_Z exactly.

3. The repair alert model

Definition 1: The pair (X, Z) of life variables satisfies the requirements of the repair alert model provided the following two conditions both hold:

(i) The event $\{Z < X\}$ is stochastically independent of X (i.e. Z is a random signs censoring of X).

(ii) There exists an increasing function G with $G(0) = 0$ such that for all $x > 0$,

$$P(Z \leq z | Z < X, X = x) = \frac{G(z)}{G(x)}, \quad 0 < z \leq x \, .$$

The function G is called *the cumulative repair alert function*. Its derivative g (which we shall assume exists) is called *the repair alert function*.

The repair alert model is, as already noted, a specialization of random signs censoring, obtained by introducing the repair alert function g. Part (ii) of the definition means that, given a potential failure at time $X = x$, and given that a PM will be performed before that time, the conditional density of the actual time Z of PM is proportional to g. The repair alert function is meant to reflect the reaction of the maintenance crew. Thus $g(t)$ ought to be large at times t for which failures are expected and the alert therefore should be high. Langseth and Lindqvist[7] simply used $g(t) = \lambda_X(t)$ where λ_X is the hazard rate of the failure time X.

It is seen that the repair alert model is completely determined by the marginal distribution function F_X of X, the cumulative repair alert function G, the probability $q \equiv P(Z < X)$, and the assumption that the event $\{Z < X\}$ is independent of X. In fact, given those ingredients it is straightforward to derive a valid joint distribution for (X, Z)[10].

From the definition we obtain the following expressions for the subdistribution- and conditional distribution functions[10]:

$$\tilde{F}_X(t) = F_X(t), \tag{1}$$

$$F_X^*(t) = (1-q)F_X(t), \tag{2}$$

$$\tilde{F}_Z(t) = F_X(t) + G(t)\int_t^\infty \frac{f_X(y)}{G(y)}\mathrm{d}y, \tag{3}$$

$$\tilde{f}_Z(t) = g(t)\int_t^\infty \frac{f_X(y)}{G(y)}\mathrm{d}y, \tag{4}$$

$$F_Z^*(t) = q\tilde{F}_Z(t). \tag{5}$$

It follows from (1)-(2) that the marginal distribution function F_X as well as q are identifiable under the repair alert model, being functions of the subdistribution function $F_X^*(t)$. Moreover, (1) and (3) imply the following relation between the conditional distribution functions $\tilde{F}_Z$ for Z and $\tilde{F}_X$ for X,

$$\tilde{F}_Z(t) > \tilde{F}_X(t) \text{ for all } t > 0. \tag{6}$$

This property can be used in a graphical check of plausibility of a repair alert model for a data set by plotting empirical estimators of $\tilde{F}_X$ and $\tilde{F}_Z$. Two examples are given in Figure 1.

The ordering (6) between $\tilde{F}_X$ and $\tilde{F}_Z$ holds whenever Z is a random signs censoring of X (Cooke[2]). In fact, Cooke[2] proved that this ordering is

also sufficient for the existence of a joint distribution of X and Z satisfying the requirements of random signs and having a given set of subsurvival functions consistent with $\tilde{F}_X$ and $\tilde{F}_Z$.

As a strengthening of Cooke's result it was shown in Lindqvist et al[10]. that whenever (6) holds, there is an essentially unique repair alert model having a given set of subsurvival functions for X and Z consistent with $\tilde{F}_X$ and $\tilde{F}_Z$. A precise formulation of this is given by the following result:

Theorem 2: Let K_1, K_2 be a subdistribution pair such that K_2 is differentiable. Suppose furthermore that

$$\frac{K_1(t)}{K_1(\infty)} < \frac{K_2(t)}{K_2(\infty)} \text{ for all } t > 0.$$

Then there exists a pair (X, Z) of life variables which satisfy the requirements of the repair alert model and which are such that

$$F_X^*(t) = K_1(t),\ F_Z^*(t) = K_2(t) \text{ for all } t \geq 0.$$

Moreover, for any such pair (X, Z) we have $F_X(t) = K_1(t)/K_1(\infty)$, $q = K_2(\infty)$, while the cumulative repair alert function G is uniquely (modulo a multiplicative constant) given by

$$G(t) = \exp\left\{\int_{t_0}^{t} \frac{\tilde{f}_Z(w)}{\tilde{F}_Z(w) - F_X(w)} \mathrm{d}w\right\} \tag{7}$$

$$= \exp\left\{\int_{\tilde{F}_Z(t_0)}^{\tilde{F}_Z(t)} \frac{\mathrm{d}y}{y - F_X(\tilde{F}_Z^{-1}(y))}\right\} \tag{8}$$

for all $t > 0$, where $t_0 > 0$ is a fixed, arbitrary constant.

The theorem is proved in Lindqvist et al.[10] Note that the expression (7) for G is obtained from equations (3)-(4) which imply that

$$\frac{\tilde{f}_Z(t)}{\tilde{F}_Z(t) - F_X(t)} = \frac{g(t)}{G(t)}.$$

A simple example of a cumulative repair alert function is $G(t) = t^\beta$ where $\beta > 0$ is a parameter. Then $g(t) = \beta t^{\beta-1}$ so $\beta = 1$ means a constant repair alert function, while $\beta < 1$ and $\beta > 1$ correspond to, respectively, a decreasing and increasing repair alert function. It follows, furthermore, that for this repair alert function we have

$$E(Z|Z < X) = \int_0^\infty (1 - \tilde{F}_Z(z))\mathrm{d}z = \frac{\beta}{\beta + 1} E(X). \tag{9}$$

Thus cost efficient PM performance corresponds to large values of β, since this implies that PM can be expected to be close to the potential failure time.

4. Statistical inference in the repair alert model

4.1. *Independent censoring*

Let (Y, δ) be the result of a competing risk case between failure time X and PM-time Z of a component. Suppose now that observations of (Y, δ) may be right censored by a random variable C which is independent of (X, Z) and hence of (Y, δ). Then by considering the competing risk case between Y and C it follows by independence that the marginal distributions of Y and C are identifiable. However, in order to identify the underlying repair alert model we need to have identifiability of the distribution of the pair (Y, δ). Fortunately, this is the case.

To see this, note first that the probabilities $P(y \leq Y \leq y+\mathrm{d}y, \delta = 0, Y < C)$ and $P(y \leq Y \leq y + \mathrm{d}y, \delta = 1, Y < C)$ are identifiable from observation of the competing risk case between X, Z and C. But these probabilities can be written as, respectively, $f^*_X(y)P(C > y)\mathrm{d}y$ and $f^*_Z(y)P(C > y)\mathrm{d}y$ by independence of Y and C. Thus, assuming that $P(C > y) > 0$ for all y, the subdistribution functions of X and Z are identifiable since the distribution of C is. Hence the underlying repair alert model can be identified as well.

4.2. *Data sets and preliminary graphical model checking*

Let there be N independent right censored observations as described in the previous subsection. By extending the notation of Bedford and Cooke[1], Section 9.5, we may let these observations be represented on the form $x_1, \ldots, x_m, z_1, \ldots, z_n, c_1, \ldots, c_r$, which are, respectively, the observed times to failure, the observed times to PM, and the observed times to censoring.

For practical illustration we use two data sets. The first one, from Mendenhall and Hader[11], gives failure times for ARC-1 VHF communication transmitter-receivers of a single commercial airline. They will later be referred to as the VHF-data. Failed units were removed from the aircraft for maintenance. However, in some cases the apparent failures were unconfirmed upon arrival at the maintenance center, as the unit exhibited satisfactory operation when tested there. Thus, the failure times can be divided into two groups, unconfirmed, Z, and confirmed failures, X. There are $m = 218$ observations of X, and $n = 107$ observations of Z. The data were

censored at time $\tau = 630$, and there are $r = 44$ such censored observations. This gives a total of $N = 369$ observations in the dataset.

The second dataset was prepared by Langseth and Lindqvist[8]. These data are failure times of a single mechanical component taken from the OREDA database[12], and will be referred to as the OREDA data. The component under study could fail due to several different failure modes. In the present data study we will focus on failures of type "2", and treat the other failures as independent censorings of the failure times. The component failures are either "critical", X, or "non-critical" (degraded or incipient), Z. We will only use data starting from the tenth event, that is, after the first critical failure was repaired. This gives us $m = 12$ observations of X, $n = 29$ observations of Z, and $r = 37$ censored observations, a total of $N = 78$ cases. The resulting data are given in Table 1.

Table 1. OREDA data. The first line contains the observed failure times x_i; the second line contains the observed PM times z_j; the two last lines contain the censoring times c_k.

x_i:	1,1,5,8,10,11,11,13,25,80,85,117
z_j:	1,1,1,1,1,1,1,1,3,3,3,3,4,5,7,8,10,12,12,14,17,18,24,24,28,28,28,32,36
c_k:	1,1,2,2,2,2,2,2,3,3,4,4,4,5,6,6,6,7,7,7,10,12,12,12,12,13,19,30,31,32, 32,47,49,61,65,76,97

Suppose we want to fit a repair alert model to the data. By (6) we need to have $\tilde{F}_Z(t) > \tilde{F}_X(t)$ for all $t > 0$. For a graphical verification of this, we use nonparametric estimators of the conditional distribution functions $\tilde{F}_X$ and $\tilde{F}_Z$ as derived by Lawless[9], Section 9.2. We then start by computing the Kaplan-Meier estimator $\hat{S}(t)$ based on the right censored sample of Y's, i.e. the union $\{y_i\}$ of the x_i and the z_j with the c_k being censorings. This leads to

$$\hat{S}(t) = \prod_{i:y_i<t} \frac{R(y_i) - d(y_i)}{R(y_i)};\ t > 0,$$

where $R(t)$ is the total number of units (components) which are at risk just before t, while $d(t)$ is the number of observed events (failure or PM) at time t. The subdistribution functions can next be estimated using equation

(9.2.5) in Lawless[9], which in our notation can be written as

$$\hat{F}_X^*(t) = \sum_{i:x_i \le t} \frac{\hat{S}(x_i)}{R(x_i)}, \tag{10}$$

$$\hat{F}_Z^*(t) = \sum_{j:z_j \le t} \frac{\hat{S}(z_j)}{R(z_j)}. \tag{11}$$

Recall that the conditional distribution functions are given by $\tilde{F}_X(t) = F_X^*(t)/F_X^*(\infty)$ and likewise for $\tilde{F}_Z$. Moreover, recall that $F_X^*(\infty)+F_Z^*(\infty) = 1$. The natural estimators of $\tilde{F}_X$ and $\tilde{F}_Z$ are obtained by dividing (10)-(11) by $\hat{F}_X^*(\infty)$ and $\hat{F}_Z^*(\infty)$, respectively. However, the estimates $\hat{F}_X^*(\infty)$ and $\hat{F}_Z^*(\infty)$ do not necessarily add to 1, so Lawless[9] suggests to normalize them to have sum 1. Thus, defining

$$\hat{q} = \frac{\hat{F}_Z^*(\infty)}{\hat{F}_X^*(\infty) + \hat{F}_Z^*(\infty)}, \tag{12}$$

we obtain the estimators

$$\hat{\tilde{F}}_X(t) = \hat{F}_X^*(t)/(1-\hat{q}), \tag{13}$$

$$\hat{\tilde{F}}_Z(t) = \hat{F}_Z^*(t)/\hat{q}. \tag{14}$$

Figure 1 shows the plots of $\hat{\tilde{F}}_X(t)$ and $\hat{\tilde{F}}_Z(t)$ obtained in this way for the two datasets. The required inequality (6) is apparently satisfied for the estimated functions, and we conclude that it is indeed meaningful to fit repair alert models to both datasets. Formal tests for investigations of this kind are considered by Dewan et al.[6]

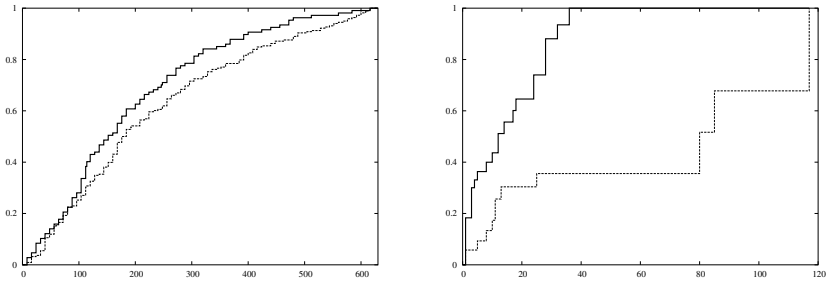

Fig. 1. Empirical subdistribution functions $\hat{\tilde{F}}_Z(t)$ (thick line) and $\hat{\tilde{F}}_X(t)$ (thin line) for the VHF data (left panel) and OREDA data (right panel).

4.3. *Nonparametric estimation*

In this subsection we suggest simple nonparametric estimators of q, F_X and G for the repair alert model. First, a natural estimator for q will be the $\hat{q}$ defined by (12). For the VHF-data this equals $\hat{q} = 0.33$, and the OREDA data gives $\hat{q} = 0.71$. Next, by (1) we may estimate F_X by $\hat{\tilde{F}}_X$ given in (13) and depicted in Figure 1. It remains therefore to estimate G.

Following Lindqvist et al.[10], we start from the definition of $G(t)$ in (8), repeated here for ease of reference,

$$G(t) = \exp\left\{ \int_{\tilde{F}_Z(t_0)}^{\tilde{F}_Z(t)} \frac{dy}{y - F_X(\tilde{F}_Z^{-1}(y))} \right\}.$$

We then proceed by substituting the estimator $\hat{\tilde{F}}_X(t)$ for $F_X(t)$. It follows from (10) and (13) that $\hat{\tilde{F}}_X$ is constant on intervals $[x_\ell, x_{\ell+1})$, with value $\hat{\tilde{F}}_X(x_\ell)$. Thus

$$\hat{\tilde{F}}_X(\tilde{F}_Z^{-1}(y)) = \hat{\tilde{F}}_X(x_\ell) \text{ for } \tilde{F}_Z(x_\ell) \le y < \tilde{F}_Z(x_{\ell+1}),\ \ell = 1, \ldots, m-1.$$

By selecting $t_0 = x_1$ and $t = x_i$ in (8), we obtain

$$\begin{aligned} \int_{\tilde{F}_Z(x_1)}^{\tilde{F}_Z(x_i)} \frac{\mathrm{d}y}{y - \hat{\tilde{F}}_X(\tilde{F}_Z^{-1}(y))} &= \sum_{\ell=1}^{i-1} \int_{\tilde{F}_Z(x_\ell)}^{\tilde{F}_Z(x_{\ell+1})} \frac{\mathrm{d}y}{y - \hat{\tilde{F}}_X(x_\ell)} \\ &= \sum_{\ell=1}^{i-1} \log \frac{\tilde{F}_Z(x_{\ell+1}) - \hat{\tilde{F}}_X(x_\ell)}{\tilde{F}_Z(x_\ell) - \hat{\tilde{F}}_X(x_\ell)}. \end{aligned}$$

Since $G(t)$ is only determined modulo a constant, we can define $\hat{G}(x_1) = 1$. Finally, substituting $\hat{\tilde{F}}_Z(t)$ from (14) for $\tilde{F}_Z(t)$, we obtain the nonparametric estimator for $G(t)$ defined at the points $t = x_i$:

$$\hat{G}(x_i) = \prod_{\ell=1}^{i-1} \frac{\hat{\tilde{F}}_Z(x_{\ell+1}) - \hat{\tilde{F}}_X(x_\ell)}{\hat{\tilde{F}}_Z(x_\ell) - \hat{\tilde{F}}_X(x_\ell)}. \tag{15}$$

We have tacitly assumed that $\hat{\tilde{F}}_Z(t) > \hat{F}_X\left(\hat{\tilde{F}}_Z^{-1}(t)\right)$ for all ℓ in this development. This assumption is theoretically justified by (6), but in practice it may still happen that $\hat{\tilde{F}}_Z(x_\ell) \le \hat{F}_X\left(\hat{\tilde{F}}_Z^{-1}(x_\ell)\right)$ for some ℓ. In this case we suggest to put the corresponding factor of (15) equal to 1.

Figure 2 shows the described estimator of G for the two data sets with $\log \hat{G}(x_i)$ plotted against $\log x_i$. The motivation for these plots is to check

whether the parametrization $G(t) = t^\beta$ is plausible. In that case we will have $\log G(t) = \beta \log t$, so we would expect plots of $\log \hat{G}(x_i)$ against $\log x_i$ to be approximately a straight line with slope β. This is roughly true in Figure 2. Based on the plots, we may estimate the slopes of the curves to be around 5 (VHF-data), and .7 (OREDA-data). These estimates are therefore our first guesses of β.

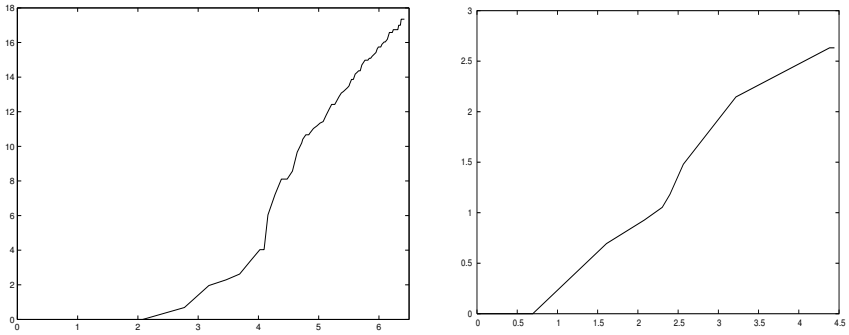

Fig. 2. Nonparametric estimate $\log \hat{G}(x_i)$ plotted against $\log x_i$ for the VHF data (left pane) and the OREDA data (right pane).

4.4. *Parametric estimation*

In this section we assume the special parametric model where X is exponentially distributed with $f_X(x) = \lambda e^{-\lambda x}$, while $G(t) = t^\beta$. For notational simplicity, we recall the definition of the incomplete Gamma function, $\Gamma(\psi, t) = \int_t^\infty w^{\psi-1} e^{-w} \mathrm{d}w$. Note that the integral converges for all real ψ when $t > 0$, and for all $\psi > 0$ when $t = 0$.

Following Crowder[4], the contributions to the likelihood from an uncensored observation is given by the subdensity function at the observed time. Thus, (1)-(5) imply that the likelihood contribution from an x_i is $f_X^*(x_i) = (1-q) f_X(x_i) = (1-q)\lambda e^{-\lambda x_i}$; the contribution from a z_j is $f_Z^*(z_j) = q \cdot g(z_j) \int_{z_j}^\infty [f_X(t)/G(t)] \mathrm{d}t = q\lambda\beta(\lambda z_i)^{\beta-1}\Gamma(1-\beta, \lambda z_i)$, and finally the contribution from a censoring c_k is $P(\min(X, Z) > c_k) = 1 - (F_X^*(c_k) + F_Z^*(c_k)) = e^{-\lambda c_k} - q(\lambda c_k)^\beta \cdot \Gamma(1-\beta, \lambda c_k)$.

The total likelihood for the data is obtained as the product for each

data point. Taking the logarithm we obtain the log-likelihood function

$$\begin{aligned} l(\lambda,\beta,q) = {} & m\log(1-q) + n\log q + (n+m)\log\lambda + n\log\beta \\ & -\lambda\sum_{i=1}^{m} x_i + (\beta-1)\sum_{i=1}^{n}\log(\lambda z_i) + \sum_{i=1}^{n}\log[\Gamma(1-\beta,\lambda z_i)] \\ & +\sum_{k=1}^{r}\log[\exp(-\lambda c_k) - q(\lambda c_k)^{\beta}\cdot\Gamma(1-\beta,\lambda c_k)]. \end{aligned} \tag{16}$$

Maximum likelihood estimates of the parameters λ, β and q can be found by maximizing (16), which needs to be done numerically. It turns out that the EM-algorithm[5] is useful here. The general idea is to augment the data artificially in order to obtain a more tractable likelihood function, for which there may exist simple expressions for the maximum likelihood estimators. This is the so called M-step (maximization step) of the EM-algorithm. The M-step alternates in an iterative manner with the E-step (expectation step), in which we compute the conditional expectation of the augmented likelihood function conditional on the observed data.

During the M-step we shall assume that we have always observed X and δ, while Z was observed only when $\delta = 1$. Furthermore, we assume that none of these observations are censored by C. It is practical to change slightly the meaning of the x_i and z_j. We now assume that there are N triples (x_i, z_i, δ_i). Here $\delta_i = 0$ if $x_i < z_i$, in which case we observe only x_i, δ_i, and $\delta_i = 1$ if $z_i < x_i$, in which case we observe the whole triple (x_i, z_i, δ_i). The augmented likelihood now becomes

$$L_A(\lambda,\beta,q) = \prod_{i=1}^{N}\left\{\lambda e^{-\lambda x_i}(1-q)^{1-\delta_i}q^{\delta_i}\left(\frac{\beta z_i^{\beta-1}}{x_i^{\beta}}\right)^{\delta_i}\right\}$$

which by taking the logarithm gives the augmented log-likelihood,

$$\begin{aligned} l_A(\lambda,\beta,q) = {} & N\log\lambda - \lambda\sum_{i=1}^{N} x_i + N\log(1-q) - \log(1-q)\sum_{i=1}^{N}\delta_i \\ & +\log q\sum_{i=1}^{N}\delta_i + \log\beta\sum_{i=1}^{N}\delta_i + (\beta-1)\sum_{i=1}^{N}\delta_i\log z_i - \beta\sum_{i=1}^{N}\delta_i\log x_i. \end{aligned} \tag{17}$$

We consider the M-step first, where we find the maximum likelihood estimators from (17). Then we consider the E-step where we replace the unobserved terms of the log-likelihood by their expected values, conditional on the observed data and the current parameter estimates.

M-step: Maximization of (17) gives us explicit expressions for the maximum likelihood estimators:

$$\hat{q} = \frac{\sum_{i=1}^{N} \delta_i}{N}, \quad \hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}, \quad \hat{\beta} = \frac{\sum_{i=1}^{N} \delta_i}{\sum_{i=1}^{N} \delta_i \log(x_i/z_i)}. \tag{18}$$

E-step: In this step we compute the expected value of (17) given our data $x_1, \ldots, x_m, z_1, \ldots, z_n, c_1, \ldots, c_r$. For the observations where the failure time x_i is observed, we have $\delta_i = 0$ and the value of z_i is not needed. For the observations where the PM-time z_i is observed we have $\delta_i = 1$ while x_i is not observed. Hence, we need to replace the corresponding x_i and $\log x_i$ in (17) by their conditional expectations. These are computed by first noting that the conditional density of X given $\{Z < X, Z = z\}$ is

$$f(x|Z < X, Z = z) = \frac{x^{-\beta} e^{-\lambda x}}{\lambda^{\beta-1}\Gamma(1-\beta, \lambda z)} \text{ for } x > z,$$

and hence that $E[X|Z < X, Z = z] = \frac{\Gamma(2-\beta,\lambda z)}{\lambda \cdot \Gamma(1-\beta,\lambda z)}$ and $E[\log(X)|Z < X, Z = z] = \int_{\lambda z}^{\infty} \log(w) w^{-\beta} \exp(-w) \mathrm{d}w / \Gamma(1-\beta, \lambda z) - \log(\lambda)$.

Finally, we consider the observations where the censoring time $C = c$ is observed. In this case we do not observe δ_i, and hence from (17) we need to compute $E[X| \min(X, Z) > c]$, $E[\delta| \min(X, Z) > c]$, $E[\delta \log X| \min(X, Z) > c]$, and $E[\delta \log X| \min(X, Z) > c]$. After some algebraic manipulations, we find that

$$\begin{aligned}
P(Z < X| \min(X, Z) > c) &= \frac{P(Z > c|Z < X) \cdot P(Z < X)}{P(X > c, Z > c)} \\
&= \frac{q\beta \int_{\lambda c}^{\infty} u^{\beta-1}\Gamma(1-\beta, u)\mathrm{d}u}{\exp(-\lambda c) - q(\lambda c)^{\beta} \cdot \Gamma(1-\beta, \lambda c)}, \\
f(x| \min(X, Z) > c, Z < X) &= \frac{f_X(x) P(Z > c|X = x, Z < X)}{\int_c^{\infty} f_X(u) P(Z > c|X = u, Z < X)\mathrm{d}u} \\
&= \frac{\lambda \exp(-\lambda x) \left[1 - (c/x)^{\beta}\right]}{\exp(-\lambda c) - (\lambda c)^{\beta} \Gamma(1-\beta, \lambda c)} \text{ for } x > c, \\
f(x| \min(X, Z) > c, X < Z) &= \frac{f_X(x)}{P(X > c)} = \lambda \exp\left(-\lambda(x - c)\right) \text{ for } x > c, \\
f(z| \min(X, Z) > c, Z < X) &= \frac{g(z) \int_z^{\infty} [f_X(x)/G(x)]\, \mathrm{d}x}{\int_c^{\infty} g(z) \int_z^{\infty} [f_X(x)/G(x)]\, \mathrm{d}x\, \mathrm{d}z} \\
&= \frac{z^{\beta-1}\Gamma(1-\beta, \lambda z)}{\int_c^{\infty} z^{\beta-1}\Gamma(1-\beta, \lambda z)\, \mathrm{d}z} \text{ for } z > c.
\end{aligned}$$

From these conditional densities we find the desired expectations. The EM-algorithm now proceeds by using the augmented dataset to re-estimate the parameters using (18), use these new estimators to generate a new augmented database, and so on until convergence.

The resulting estimates for the VHF and OREDA-data are given in Tables 2 and 3, respectively. We also include bounds for approximate 95% confidence intervals based on standard log-likelihood theory. We see that β appears to be larger for the VHF data than for the OREDA data. This is in correspondence with Figure 1, where the sub-distribution functions for the VHF data are closer together than those of the OREDA data. It is also interesting to not that $\hat{\beta} = 1.00$ for the OREDA data. This corresponds to choosing $g(t)$ proportional to the hazard rate of the failure times, as in the class of models investigated by Langseth and Lindqvist[7]. Finally, we note that the confidence interval for β in the VHF data extends all the way to infinity. The meaning of using $\beta = \infty$ should be seen in relation to (9), and indicates that when an unconfirmed failure (Z) was observed, it occurred immediately before a failure (X) would have been realized.

Table 2. Maximum likelihood estimates and approximate 95% confidence intervals for parametric repair alert model for the VHF-data.

Parameter	Estimate	Lower bound	Upper bound
λ	$3.10 \cdot 10^{-3}$	$2.73 \cdot 10^{-3}$	$3.51 \cdot 10^{-3}$
β	4.44	2.08	∞
q	0.318	0.270	0.369

Table 3. Maximum likelihood estimates and approximate 95% confidence intervals for parametric repair alert model for the OREDA-data.

Parameter	Estimate	Lower bound	Upper bound
λ	$1.80 \cdot 10^{-2}$	$1.04 \cdot 10^{-2}$	$2.86 \cdot 10^{-2}$
β	1.00	.553	2.74
q	0.621	0.461	0.771

5. Concluding remarks

In this chapter we have considered the repair alert model, which describes a specific dependence structure between failures (X) and preventive maintenance (Z). We extend our previous work[10] by including external censoring

in the model. The use of our model is exemplified by analyzing two different datasets: The VHF data are type I censored at a given time τ but have previously been analyzed without taking this into account. The OREDA dataset describes several different failure modes, and we handle this situation by focusing on one particular failure mode and consider all other failure modes as external censorings.

References

1. T. Bedford and R. M. Cooke, *Probabilistic risk analysis: Foundations and methods* (Cambridge University Press, Cambridge, 2001).
2. R. M. Cooke, The total time on test statistics and age-dependent censoring, *Statistics and Probability Letters*, **18**, 307–312 (1993).
3. R. M. Cooke, The design of reliability databases, Part I and II, *Reliability Engineering and System Safety*, **51**, 137–146 and 209–223 (1996).
4. M. J. Crowder, *Classical Competing Risks* (Chapman and Hall/CRC, Boca Raton, 2001).
5. A. P. Dempster, N. M. Laird, and D. B Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Ser. B*, **39**, 1–38 (1977).
6. I. Dewan, J. V. Deshpande, and S. B. Kulathinal, On testing dependence between time to failure and cause of failure via conditional probabilities, *Scandinavian Journal of Statistics*, **31**, 79–91 (2003).
7. H. Langseth and B. H. Lindqvist, A maintenance model for components exposed to several failure mechanisms and imperfect repair, in *Mathematical and Statistical Methods in Reliability, Series on Quality, Reliability and Engineering Statistics, Vol. 7*, Eds. B. H. Lindqvist and K. A. Doksum (World Scientific Publishing, Singapore, 2003), pp. 415-430.
8. H. Langseth and B. H. Lindqvist, Competing risks for repairable systems: A data study. To appear in *Journal of Statistical Planning and Inference,* Special Issue on Competing Risks (2005).
9. J. F. Lawless, *Statistical models and methods for lifetime data, 2nd ed.* (Wiley-Interscience, Hoboken, N.J., 2003).
10. B. H. Lindqvist, B. Støve and H. Langseth, Modelling of dependence between critical failure and preventive maintenance: The repair alert model. To appear in *Journal of Statistical Planning and Inference,* Special Issue on Competing Risks (2005).
11. W. Mendenhall and R. J. Hader, Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data, *Biometrika*, **45**, 504–520 (1958).
12. OREDA, Offshore Reliability Data Handbook, 4th ed. (Distributed by Det Norske Veritas, P.O. Box 300, N-1322 Høvik, Norway, http://www.oreda.com/, 2001).
13. A. Tsiatis, A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences, USA* **72**, 20-22 (1975).