Using EM and Data Augmentation for the Competing Risks Model

Radu V. Craiu
Department of Statistics
University of Toronto
100 St. George Street
Toronto, ON, M5S 3G3, Canada
craiu@utstat.toronto.edu

Thierry Duchesne
Département de mathématiques et de statistique
Pavillon Vachon
Université Laval
Québec, QC, G1K 7P4, Canada
duchesne@mat.ulaval.ca

February 17, 2004

We consider a survival analysis problem in which items are subject to failure from competing risks. For some of the items, the failure cause is known only to belong to a subset of the set of all possible causes, while for the remaining items the cause of death is known precisely. In this chapter we investigate two complementary analyses based on models in which the hazard rates are assumed piecewise constant. The approaches proposed rely on the EM algorithm and its Bayesian counterpart, the data augmentation (DA) algorithm. An example is used to illustrate the advantages of each analysis.

1 Introduction

In situations in which the survival data involve several different failure types, the analysis is performed using the theory of competing risks. In most medical and industrial applications the data includes the time of censoring or failure and an indicator of the failure cause for each item/patient. However, it is often the case, especially with modular systems, that for a certain subset of the items, the true cause of failure is not known exactly. Such items are said to have a *masked cause of failure*. While in some cases the failure can be isolated

down to a subset of causes, without any such additional information the masking group is considered to be the entire set of causes. In certain experiments, a second stage analysis can be conducted so that part of the items with a masked cause of failure are investigated and an exact diagnostic of the failure cause is obtained.

The literature on competing risks with masked causes of failure has grown greatly in the recent years. In the context of carcinogenecity studies, Racine-Poon and Hoel (1984) establish a non-parametric estimate of the survival function, while Dinse (1986) proposes non-parametric maximum likelihood estimators of prevalence and mortality. Several other authors also discuss the problem of missing cause-of-death in carcinogenicity studies (Kodell and Chen, 1987; Lagakos, 1982; Lagakos and Louis, 1988). Goetghebeur and Ryan (1990) and Dewanji (1992) construct a log-rank test to assess the difference between survival functions for subgroups of the population under study in the presence of covariates. Goetghebeur and Ryan (1995) subsequently generalize the approach to proportional cause-specific hazards regression models. Flehinger et al. (1998, 2002) consider the analysis of datasets in which there are second stage data. They propose maximum likelihood estimation using a model with non-parametric proportional cause-specific hazards (Flehinger et al., 1998) and a model with completely parametric cause-specific hazards (Flehinger et al., 2002). The literature regarding the Bayesian analyses of this problem is reviewed at the beginning of Section 4.

Proportionality between the cause-specific hazards or their complete parametric specification are assumptions which do not always mirror reality. In this chapter we propose two approaches, both based on piecewise constant hazards. We assume no proportionality between the hazards and only weak parametric assumptions are made, namely no particular shape is imposed on the hazards. The model is defined in Section 2. In Section 3 we briefly describe an EM-based approach which is analyzed in detail in Craiu and Duchesne (2004). The model described in Section 2 is the backbone of the Bayesian analysis presented in Section 4 which represents the main contribution of this chapter. While pro's and con's are discussed for each analysis, we hope that the illustration from Section 4 will emphasize the advantages of each approach as well as the potential for combining their strengths. Conclusions and further work are in Section 6.

2 The Model

We consider a situation in which n independent items are observed in the time interval $[0, T_{\max}]$ and each of them can fail of exactly one of J possible causes. The data are collected in two stages. In the first stage we observe for each item its failure time which may be censored if at time $T_{\max}$ the item was still functioning. For those items which have failed while in the study, we can observe one of the following two situations: (1) item i fails due to cause j at time t, (2) item i fails due to an unknown cause of failure which is known to belong to a group of failure causes $g(i) \subset \{1, \ldots, J\}$. The items that belong to the second situation have a masked failure cause. In the second stage, a subset of the masked items are sent for further analysis and the precise cause of failure is then determined. It is intuitive that the masking parameters shall be estimated using those items that are sent to the second stage of the experiment. In fact, if all the items were sent to the second stage, then all the information needed for estimation would be available and no missing data procedure would be necessary. Hence, we get a natural definition of the complete data as the dataset that we would obtain if every masked item with an uncensored failure time were sent to a second stage analysis. Suppose there are M masking groups in the dataset (including the groups consisting of the individual failure causes). The observation for item i in the complete dataset would be $(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_M}, \delta_{i1}, \ldots, \delta_{iJ})$, where γ_{ig} is the indicator that item i's failure cause was masked to group g at the first stage (if the failure cause is known to be j at the first stage, then we say that it is masked to $g = \{j\}$), δ_{ij} is the indicator that item i's actual failure cause is j (if an item is right-censored, then all the indicators δ_{ij}, $j = 1, \ldots, J$, take on value 0). The groups containing more than one cause are called *proper*.

Here is a short example to set the notation straight. Suppose that we have two potential causes of failure, say causes 1 and 2. Let us assume that at the first stage we either identify the cause of failure directly (in which case we say that it is masked in group $\{1\}$ or $\{2\}$ accordingly) or we only know that failure is due to one of causes 1 or 2 (in which case we say that failure is masked in group $\{1, 2\}$). For item 1, we have failure at time 2.4 masked in group $\{1, 2\}$ at stage 1 with no second stage. Item 2 fails at time 6.3 of a cause masked in group $\{1, 2\}$ and it is found in a second stage analysis that failure was actually due to cause 2. Item 3 is right-censored at time 4.1, and item 4 fails at time 7.2 and its failure is diagnosed in stage 1 as being due to the first cause. These four observations would be coded

as

$$\begin{aligned}
(t_1, \gamma_{1\{1\}}, \gamma_{1\{2\}}, \gamma_{1\{12\}}, \delta_{11}, \delta_{12}) &= (2.4, 0, 0, 1, \cdot, \cdot) \\
(t_2, \gamma_{2\{1\}}, \gamma_{2\{2\}}, \gamma_{2\{1,2\}}, \delta_{21}, \delta_{22}) &= (6.3, 0, 0, 1, 0, 1) \\
(t_3, \gamma_{3\{1\}}, \gamma_{3\{2\}}, \gamma_{3\{1,2\}}, \delta_{31}, \delta_{32}) &= (4.1, \cdot, \cdot, \cdot, 0, 0) \\
(t_4, \gamma_{4\{1\}}, \gamma_{4\{2\}}, \gamma_{4\{1,2\}}, \delta_{41}, \delta_{42}) &= (7.2, 1, 0, 0, 1, 0)
\end{aligned}$$

where $\cdot$ represents missing data. We denote by M_2 all masked items which have not been sent to a second stage analysis and by G_j the set of all masking groups containing cause j. The number of elements in G_j is denoted L_j and we define $G_j^* = G_j \backslash \{j\}$.

The statistical model has a part involving the competing-risk aspect (failure times, hazard rates) and a part due to masking (masking probabilities). If T^* and J^* are random variables that represent the failure time and the cause of failure, respectively, then the cause specific hazards are

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{P[t < T^* \leq t + h, J^* = j | T^* \geq t]}{h}, \quad j = 1, \ldots, J. \tag{1}$$

In this chapter we suppose that the cause-specific hazard functions are piecewise constant, i.e., there exists a partition of the time interval $[0, T_{\max}]$ given by $0 = a_0 < a_1 < \cdots < a_K = T_{\max}$ such that, if $1_k(t)$ is the indicator that $t \in (a_{k-1}, a_k]$, then

$$\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t). \tag{2}$$

The choice of the same endpoints for the hazard intervals $(a_{k-1}, a_k]$ is justified because it allows testing for the proportionality of cause-specific hazards and symmetry, as shown in Craiu and Duchesne (2004). However, if no such tests are necessary the analysis described here can be carried on even if the intervals have different lengths for different cause-specific hazards. In such a situation the notation for the end points would have to include a second index, j, to show their dependence on the cause. Of ultimate interest are the *diagnostic probabilities*

$$\pi_{j|g(i)}(t_i) = P[\text{item } i \text{ failed of } j | \text{failed at } t_i \text{ and was masked in } g(i)],$$

for all masked items i and all causes $j \in g(i)$. In order to compute $\pi_{j|g}(t)$ we need the *masking probabilities*

$$p_{g|j} = P[\text{cause masked to group } g \text{ at stage 1}|\ \text{actual failure cause is } j], \quad j \in g.$$

With the Bayes' rule we obtain

$$\pi_{j|g}(t) = \frac{\lambda_j(t) p_{g|j}}{\sum_{l \in g} \lambda_l(t) p_{g|l}}. \tag{3}$$

If θ is the vector of parameters that contains λ_{jk}, $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $p_{g_m|j}$, $j = 1, \ldots, J$, $m = 1, \ldots, M$, then the log-likelihood function under complete data is

$$\begin{aligned} \log p_C(\theta) \quad &= \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ \left[\delta_{ij} \ln \sum_{k=1}^{K} \lambda_{jk} 1_k(t_i) - \sum_{k=1}^{K} \lambda_{jk} \int_0^{t_i} 1_k(u)\, du \right] \right. \\ &+ \quad \left. \delta_{ij} \left[(1 - \sum_{g \in G_j^*} \gamma_{ig}) \ln(1 - \sum_{g \in G_j^*} p_{g|j}) + \sum_{g \in G_j^*} \gamma_{ig} \ln p_{g|j} \right] \right\}. \end{aligned} \tag{4}$$

The likelihood (4) contains a competing-risk part which involves the failure times and failure causes (first line), and a masking part which involves the masking probabilities (second line). Under complete data, these two parts would be maximized separately making the maximum likelihood estimates of the masking probabilities robust to the specification of hazard intervals. One can notice that for right-censored observations the term on the second line of equation (4) vanishes and for such items there is no need to know γ_{ig}.

3 EM-based Analysis

The EM algorithm (Dempster et al., 1977) has become a classic among the methods designed to handle the maximization of intractable likelihood functions. The use of EM to maximize (4) is recommended since the log-likelihood is linear in the missing data $\{\delta_{ij} : i \in M_2, 1 \leq j \leq J\}$ and the maximization required in the M-step can be performed in closed form, as shown below.

The algorithm

For each $i \in M_2$ with uncensored failure time t_i and with a failure cause masked in $g(i)$, we have that

$$\mathrm{E}[\delta_{ij}|Y_{OBS}, \theta] = \hat{\pi}_{j|g(i)}(t_i) = \frac{\hat{\lambda}_j(t_i) \hat{p}_{g(i)|j}}{\sum_{l \in g_i} \hat{\lambda}_l(t_i) \hat{p}_{g(i)|l}}.$$

Since the complete data log-likelihood (4) is linear in the missing δ_{ij}, substitution of the missing δ_{ij} with $\mathrm{E}[\delta_{ij}|Y_{OBS},\theta]$ constitutes the E-step of the algorithm. In addition, if we let

$$e_k = \sum_{i=1}^{n} \int_0^{t_i} 1_k(u)\, du \tag{5}$$

denote the k-th interval *exposure*, i.e. the total time lived by all items in the interval $(a_{k-1}, a_k]$, then one easily obtains that (4) is maximized when

$$\hat{\lambda}_{jk} = \frac{\sum_{i=1}^{n} \delta_{ij}\ 1_k(t_i)}{e_k} \quad \text{and} \quad \hat{p}_{g|j} = \frac{\sum_{i=1}^{n} \delta_{ij}\ \gamma_{ig}}{\sum_{i=1}^{n} \delta_{ij}}.$$

Hence, once the starting points have been chosen, the algorithm iterates between the E-step described above and the M-step given by

$$\hat{\lambda}_{jk}^{(l)} = \frac{\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}^{(l-1)}}[\delta_{ij}|Y_{OBS}]\ 1_k(t_i)}{e_k} \quad \text{and} \quad \hat{p}_{g|j}^{(l)} = \frac{\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}^{(l-1)}}[\delta_{ij}|Y_{OBS}]\ \gamma_{ig}}{\sum_{i=1}^{n} \mathrm{E}_{\hat{\theta}^{(l-1)}}[\delta_{ij}|Y_{OBS}]}. \tag{6}$$

The algorithm can be easily extended to include time-varying masking probabilities $p_{g|j}(t)$ (see Craiu and Duchesne, 2004).

In all situations encountered with relatively large sample sizes and a 30-50% percentage of masked items sent to the second stage analysis, the algorithm converges in less than 10-20 iterations. Caution is required in situations where there are no data collected in the second stage and the cause-specific hazard rates are proportional. In such a case, the parameters are unidentifiable conditional on the observed data (Flehinger et al. 1998) but are identifiable given the complete data. If the parameters are identifiable only in the complete data model there is a ridge of local maxima in the likelihood surface and the EM algorithm will converge to one of the points on the ridge, depending on the starting point. The erratic behavior of the EM can be detected by using multiple starting points. Previous authors (Goetghebeur and Ryan, 1990; Dewanji, 1992; Lo, 1991) propose a working hypothesis of *symmetry* to reduce the number of parameters and obtain identifiability. The symmetry assumption assumes that the masking probabilities $p_{g|j}$ does not depend on the cause j, i.e., $p_{g|j} = p_g$ for any group g and any $j \in g$.

Craiu and Duschesne (2004) prove results regarding the convergence of the EM algorithm, develop inference methods such as likelihood ratio tests for the assumptions of symmetry and proportionality of hazards, and apply the supplementary EM (SEM) algorithm (Meng and Rubin, 1991) for the estimation of the asymptotic variance matrix of the maximum likelihood estimators.

However, even if the cause-specific hazards are not proportional, with little or non-existent second stage data, the information about the $p_{g|j}$'s is obtained via the hazard rate estimates, which are time dependent. As a result, if the intervals for the hazards are misspecified then the maximum likelihood estimates can be far from the true values. Equations (6) require that the hazard intervals are chosen so that for each interval $1 \leq k \leq K$ and for each failure cause $1 \leq j \leq J$, there exists an i such that $j \in g_i$ and $1_k(t_i) = 1$. In most cases this implies that the intervals for the piecewise hazards are fairly large leading naturally to misspecification. We expect that combining the previous approach with the Bayesian analysis proposed in the next section will remedy this problem since, due to the prior specifications, there are no restrictions on the number and size of intervals for each cause-specific hazard.

4 Bayesian Analysis

Most of the Bayesian inferences presented in the literature of competing hazards allow parametric models for the hazard rates. Reiser et al. (1995) assume that the component lifetimes are exponentially distributed, Kuo and Yang (2000) consider also Weibull distributed lifetimes, while Basu et al. (2003) incorporate in their analysis all commonly used parametric distributions. In recent years the non-parametric Bayesian analysis of survival models has spurred a lot of work. Following the ideas of beta and gamma processes devised by Hjort (1990), Kalbfleisch (1978), Dykstra and Laud (1981), statisticians have increased the complexity of the prior elicitation for the hazards rates in the competing risks models. We refer the reader to Arjas and Gasbarra (1994), Walker and Mallick (1997), Gasbarra and Karia (2000), Salinas-Torres et al. (2002), Nieto-Barajas and Walker (2002), Ibrahim et al. (2001).

While the EM-based inference offers simplicity and robustness to the misspecification of the hazards rate when there are enough second stage data, it can also produce the wrong estimates when there is not enough information (sparse second stage data, large percentage of masked items, etc). It is therefore important to be able to incorporate in the model the knowledge accumulated from past similar experiments. In addition, the performance may be improved with a more flexible choice of the intervals for the hazards.

In the following we construct a Bayesian analysis structured on the model (4) that uses the work of Nieto-Barajas and Walker (2002) to define the prior distribution on the hazard rates. More precisely, their discrete gamma process is used to model piecewise constant hazard rates as we adapt their method to the context of competing proportional cause-

specific hazards.

Prior Distributions

As before, assume that for each cause $j \in \{1, \ldots, J\}$ we define K intervals on which the j-th hazard is constant and equal to $\lambda_{jk},\ 1 \leq k \leq K$. If we consider these intervals to be shorter, then it is likely that the values of the hazards in two successive pieces are not independent. We follow Nieto-Barajas and Walker (2002) and assume a latent process u_{jk} so that for each cause j, there is a Markovian dependence summarized by the graph

$$\lambda_{j1} \to u_{j1} \to \lambda_{j2} \to \ldots \to u_{jK-1} \to \lambda_{jK}.$$

Adding the latent variables u_{jk} allows one to model and control the dependence between values taken by one cause-specific hazard rate on adjacent intervals. Such dependence is important in situations in which we choose the intervals without a good knowledge of the underlying process (as is usually the case in practice). Alternatively, one may interpret the u_{jk}'s as virtual failures of a process identical in nature to the one under study; this point of view is attractive as it allows an intuitive interpretation of the model.

Formally, take the following conditional distributions

$$\begin{aligned} \lambda_{j1} &\sim \text{Gamma}(\alpha_{j1}, \beta_{j1}), \\ u_{jk}|\lambda_{jk} &\sim \text{Poisson}(c_{jk}\lambda_{jk}) \\ \lambda_{j,k+1}|u_{jk} &\sim \text{Gamma}(\alpha_{jk+1} + u_{jk}, \beta_{jk+1} + c_{jk}) \end{aligned} \tag{7}$$

with $\alpha_{jK+1} = \beta_{jK+1} = 0$ for all $1 \leq j \leq J$ and $1 \leq k \leq K$. The c_{jk} regulates the smoothing of the hazard λ_j so that if $c_{jk} = 0$ then λ_{jk} and λ_{jk+1} are independent. In general $10 \leq c_{jk} \leq 20$ is enough to produce smoother hazards, while taking $c = 0$ will result in approximately the same inference as the EM based one. The choice of the c_{jk}'s has to be done in connection with the width of the intervals, e.g. a succession of larger intervals requires a smaller value of the smoothing parameter. One can also let the data decide by considering the c's as part of the parameter vector and assigning exponential priors to them as suggested in Nieto-Barajas and Walker (2002). In the absence of prior information the α_{jk} and β_{jk} are recommended to be small. If we have prior information regarding the process λ_j, say we know $\text{E}[\lambda_{jk}] = \psi_{jk}$ then we can choose the α_{jk}, β_{jk} and c_{jk} such that

$$\frac{\alpha_{jk+1}}{\beta_{jk+1}} = \frac{\psi_{jk+1} - \xi_{jk+1}\psi_{jk}}{1 - \xi_{jk+1}}$$

where $\xi_{jk+1} = c_{jk}/(\beta_{jk+1} + c_{jk})$. We refer to Nieto-Barajas and Walker (2002) for other properties of the gamma process prior.

A natural conjugate prior assigned to the masking probabilities is

$$(p_{g_1|j}, p_{g_2|j}, \ldots, p_{g_{L_j}|j}) \sim \text{Dirichlet}(\eta_{1j}, \ldots, \eta_{L_j j}) \tag{8}$$

for all $1 \leq j \leq J$ causes. Lack of information on the masking probabilities will produce $\eta_{ij} =$ constant for all $1 \leq i \leq L_j$ and all causes j, while prior information can be included as $\text{E}[p_{g_i|j}] = \eta_{ij} / \sum_{h=1}^{L_j} \eta_{hj}$.

Data Augmentation Algorithm

There are two sets of latent variables in the model. For each item $i \in M_2$, there are J unobserved random variables $(\delta_{i1}, \ldots, \delta_{iJ})$. In addition the prior (7) introduces $K-1$ additional latent variables, $(u_{j1}, \ldots, u_{jK-1})$, for each cause j. In the initialization step we need to input initial guesses for all the latent variables. For the set of δ's one can use the output from the EM algorithm described in the previous section. Although the δ_{ij}^{EM} computed in the E-step are not integers, we can choose for each $i \in M_2$ the j_0 with the largest δ_{ij}^{EM} and assign $\delta_{ij_0}^{(0)} = 1$, $\delta_{ij}^{(0)} = 0$, $j \neq j_0$. In our applications we use $u_{jk}^{(0)} = 1$ for all j, k.

The data augmentation algorithm (Tanner and Wong, 1987) consists in the following steps at iteration t:

Masking probabilities For each $j \in \{1, \ldots, J\}$ sample

$$(p_{g_1|j}^{(t)}, \ldots, p_{g_{L_j}|j}^{(t)}) \sim \text{Dirichlet}\left(\eta_{1j} + \sum_{i=1}^{N} \gamma_{ig_1} \delta_{ij}^{(t-1)}, \ldots, \eta_{L_j j} + \sum_{i=1}^{N} \gamma_{ig_{L_j}} \delta_{ij}^{(t-1)}\right)$$

Hazard rates For each $j \in \{1, \ldots, J\}$

$$\lambda_{j1}^{(t)} \sim \text{Gamma}(\alpha_{j1} + u_{j1}^{(t-1)} + n_{j1}, \beta_{j1} + c_{j1} + e_1^{(t-1)})$$

$$\lambda_{jk}^{(t)} \sim \text{Gamma}(\alpha_{jk} + u_{jk-1}^{(t-1)} + u_{jk}^{(t-1)} + n_{jk}^{(t-1)}, \beta_{jk} + c_{jk-1} + c_{jk} + e_k^{(t-1)})$$

where $n_{jk}^{(t-1)}$ is the number of items which fail in the k-th interval due to cause j, $e_k^{(t)}$ is defined by equation (5), and $c_{jK} = u_{jK} = 0$. The supraindex $t-1$ means that these numbers are estimated using the latent variables imputed at step $t-1$.

Latent variables For each item $i \in M_2$

$$(\delta_{i1}^{(t)}, \ldots, \delta_{iJ}^{(t)}) \sim \text{Multin}\left(1, \frac{p_{g(i)|1}^{(t)}\lambda_1^{(t)}(t_i)}{\sum_{j\in g(i)} p_{g(i)|j}^{(t)}\lambda_j^{(t)}(t_i)}, \ldots, \frac{p_{g(i)|J}^{(t)}\lambda_J^{(t)}(t_i)}{\sum_{j\in g(i)} p_{g(i)|j}^{(t)}\lambda_j^{(t)}(t_i)}\right)$$

For each cause $j \in \{1, \ldots, J\}$ and for each interval $k \in \{1, ..., K\}$

$$\Pr(u_{jk}^{(t)} = u) \propto \frac{[c_{jk}(c_{jk} + \beta_{jk+1})\lambda_{jk}^{(t)}\lambda_{jk+1}^{(t)}]^u}{\Gamma(u+1)\Gamma(\alpha_{jk}+u)}.$$

The Proportional Hazards case

The assumption of proportional cause-specific hazards, denoted here A_{PH}, is recurrent in the literature of competing risks. However, tests to assess the correctness of such a hypothesis are rare. Craiu and Duchesne (2004) develop a likelihood ratio test for the hypothesis A_{PH}. In the present context one needs first to construct a data augmentation algorithm to sample from the parameter subspace defined by the constraints

$$A_{PH}: \ \lambda_{jk} = \phi_j \lambda_{1k}$$

for all $2 \leq j \leq J$ and all $1 \leq k \leq K$. The masking part of the model as well as the prior specification of λ_1 remain the same. For each $j \geq 2$ the prior distribution of ϕ_j is Gamma(ν_j, χ_j). The DA algorithm for the unrestricted model changes in that only the $\{u_{1k}: \ 1 \leq k \leq K-1\}$ is imputed and the hazard rates step becomes:

Hazard rates

$$\begin{aligned}
\lambda_{11}^{(t)} &\sim \text{Gamma}(\alpha_{11} + u_{11}^{(t-1)} + n_{11}, \beta_{11} + c_{11} + e_1^{(t-1)}) \\
\lambda_{1k}^{(t)} &\sim \text{Gamma}(\alpha_{1k} + u_{1k-1}^{(t-1)} + u_{1k}^{(t-1)} + n_{1k}^{(t-1)}, \beta_{1k} + c_{1k-1} + c_{1k} + e_k^{(t-1)}) \\
\phi_j^{(t)} &\sim \text{Gamma}(\nu_j + \sum_{i=1}^{N} \delta_{ij}, \chi_j + \sum_{k=1}^{K} \lambda_{1k}^{(t)} e_k).
\end{aligned}$$

We denote A_{PC} the general model with piecewise constant hazards. In assessing the validity of A_{PH} of interest is the Bayes factor

$$B_{PH} = \frac{p(Y_{OBS}|A_{PC})}{p(Y_{OBS}|A_{PH})} \tag{9}$$

It is well known (for example, Kass and Raftery, 1995; Meng and Wong, 1996; Chen et al. 2000) that

$$p(Y_{OBS}|A_{PH}) = \int p(Y_{OBS}|\theta_{PH}, A_{PH}) p(\theta_{PH}|A_{PH}) d\theta_{PH}$$

is just the normalizing constant of the posterior density $p(\theta_{PH}|Y_{OBS}, A_{PH})$. As a result, the estimation of (9) is equivalent to the estimation of a ratio of two normalizing constants.

The latter problem has been intensively studied in the last years, particularly in situations in which only samples (independent or dependent) from the two distributions of interest are available to the analyst. A simple but often highly variable solution is based on importance sampling (Geweke, 1989). Meng and Wong (1996) develop bridge sampling as a generalization of importance sampling that exploits optimally the overlap between the supports of the distributions. Recently, Gelman and Meng (1998) introduced path sampling as the limit of an infinite sequence of bridge samplers. While the theory of bridge sampling has been developed for situations in which independent realizations from each distribution are available, subsequent applications and studies (Servidea, 2002) have shown that the method also works well with dependent samples. In the context of the present analysis, we have reasonable confidence that the two models have a significant overlap since the two parameter spaces share the subset of masking probabilities. It is worth adding that the unnormalized posterior density can be computed at any point because $p(Y_{OBS}|\theta_{PH}, A_{PH})$ and $p(Y_{OBS}|\theta, A_{PC})$ can be expressed in closed form (Craiu and Duchesne, 2004).

5 Example

To assess the importance of transferring information between adjacent intervals, we consider a simulation example in which the hazards rates are Weibull distributed and there is no proportionality among them. There are 300 observations with times of failure between 0 and 15. Only 20% of the masked items are sent to a second stage analysis. There are three possible causes of failure and there are three masking groups: $g_1 = \{1, 2\}$, $g_2 = \{1, 3\}$ and $g_3 = \{1, 2, 3\}$. The probability to have the cause masked are: 60% for cause 1, 80% for cause 2 and 70% for cause 3. We implement the Bayesian analysis with 10 or 20 intervals. In the absence of prior information we take $c_{jk} = C$, $\alpha_{jk} = \beta_{jk} = 0.001$ for all causes j and all intervals k.

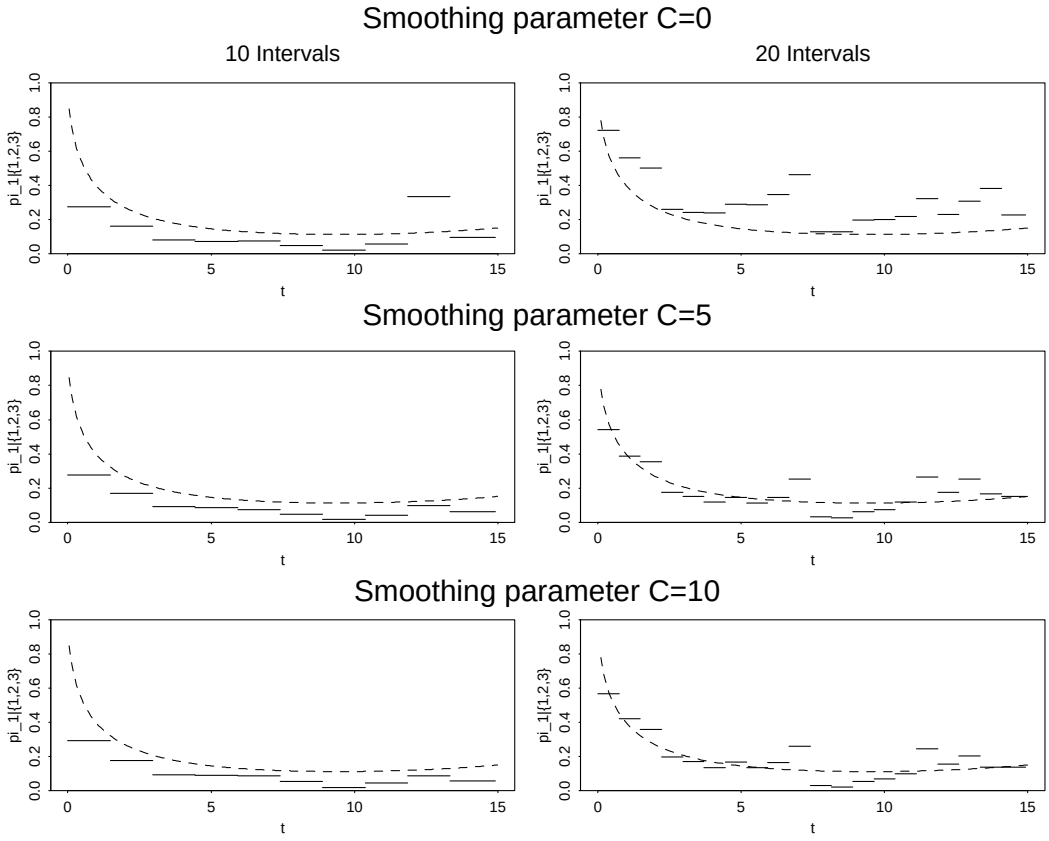

Figure 1: Plot of the posterior mean of the diagnostic probability $\pi_{1|\{1,2,3\}}$ against time as the number of intervals and the value of the smoothing parameter C vary. The true curve is represented by the dashed line, and the estimates within each interval are rendered with the solid lines.

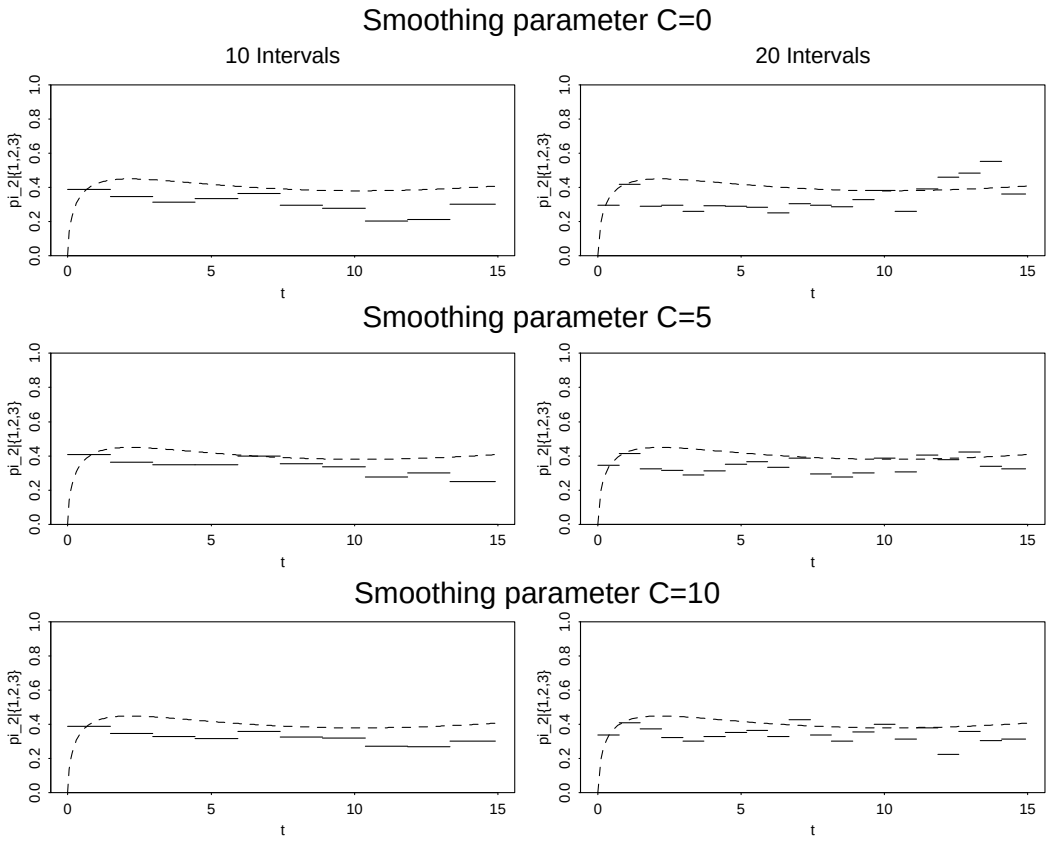

Figure 2: Plot of the posterior mean of the diagnostic probability $\pi_{2|\{1,2,3\}}$ against time as the number of intervals and the value of the smoothing parameter C vary. The true curve is represented by the dashed line, and the estimates within each interval are rendered with the solid lines.

The DA algorithm has been used to generate 4000 iterations out of which we used the last 2000 for estimation. The convergence assessment has been done following the ideas of Gelman and Rubin (1992) using Andrew Gelman's *itsim* function in S-plus applied to 4 parallel chains. The simulation lasted approximately one hour. Figures 1 and 2 illustrate the effect of increasing the "smoothing parameter" C from 0 to 10 when the number of intervals is relatively moderate (10-20). Under consideration are the posterior means of the estimators for the diagnostic probability $\hat{\pi}_{1|\{1,2,3\}}(t)$ and $\hat{\pi}_{2|\{1,2,3\}}(t)$. Each plot shows in solid line the true value, and in dotted line the piecewise constant estimator. The $C = 0$ value corresponds roughly to the EM-based inference. It is seen here that if we increase the number of intervals, the EM estimator is too rough due to the lack of sufficient data in some of the intervals. Raising the value of the smoothing parameter noticeably increases the precision of the estimate. It can also be seen from the plot that the difference between the estimators obtained for $C = 5$ and $C = 10$ is quite small.

The Bayes factor (9) can be calculated following the iterative construction of Meng and Wong (1996). This calculation is possible since one can compute the observed likelihood in any point as shown in Craiu and Duchesne (2004). With any of the above values for C, (9) ranges between 25 and 40 and shows no support for A_{PH}.

6 Discussion and Further Work

The two methods presented are complementary and should be used together to increase the strength of the analysis. While the EM analysis produces robust inference of the masking probabilities and can be used to test for the symmetry and proportional hazards assumptions, it can also be used to determine the posterior modes for some or all of the model parameters as suggested in Gelman et al. (2003, Ch. 12). The Bayesian analysis is particularly useful in producing more sensible estimates of the hazard rates when the data are sparse. In addition, the calculation of the posterior variance of the diagnostic probabilities is more straightforward once it is possible to sample from the posterior distribution of the parameters.

Within the Bayesian framework, it may be of interest to produce an automatic sequential design procedure to help the experimenter decide which masked items should be sent to the second stage analysis so that a certain given utility function is maximized.

We would like to enrich the class of possible models by relaxing the condition of piecewise linearity of the hazards. However, the computation complexity increases rapidly once we give

up linearity and needs further investigation.

7 References

1. Arjas, E., and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica* **4**, 505-524.

2. Base, S., Sen, A., and Bannered, M. (2003). Bayesian analysis of competing risks with partially masked cause of failure. *Applied Statistics* **52**, 77-93.

3. Craiu, R. V., and Duchesne, T. (2004). Inference based on the EM algorithm for the competing risk model with masked causes of failure. *Biometrika*. To appear.

4. Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

5. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

6. Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika* **79**, 855–857.

7. Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *Journal of the American Statistical Association* **81**, 328–335.

8. Dykstra, R. L., and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Annals of Statistics* **9**, 356–367.

9. Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika* **85**, 151–164.

10. Flehinger, B. J., Reiser, B., and Yashchin, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Analysis* **8**, 177–203.

11. Gasbarra, D., and Karia, S. R. (2000). Analysis of competing risks by using bayesian smoothing. *Scandinavian Journal of Statistics* **27**, 605-617.

12. Gelman, A., and Meng, X. L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

13. Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

14. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. New York: Chapman & Hall.

15. Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339.

16. Goetghebeur, E., and Ryan, L. (1990). A modified log rank test for competing risks with missing failure types. *Biometrika* **77**, 151–164.

17. Goetghebeur, E., and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821–833.

18. Hjort, N. L. (1990). Nonparametric Bayes Estimators based on beta processes for life history data. *Annals of Statistics* **18**, 1259–1294.

19. Ibrahim, J. G., Chen, M. H., Sinha, D. (2001). *Bayesian Survival Analysis*, New York: Springer-Verlag.

20. Kalbfleisch, J. D. (1978). Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society B* **40**, 214–221.

21. Kass, R. E., and Raftery, A. E. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association* **90**, 773–795.

22. Kodell, R. L., and Chen, J. J. (1987). Handling cause of death in equivocal cases using the EM algorithm (with rejoinder). *Communications in Statistics: Theory and Methods* **16**, 2565–2585.

23. Kuo L., and Yang T. Y. (2000). Bayesian reliability modeling for masked system lifetime data. *Statistics and Probability Letters*, **47**, 229-241.

24. Lagakos, S. W. (1982). An evaluation of some two-sample tests used to analyze animal carcinogenicity experiments. *Utilitas Mathematicae*, **21B**, 239–260.

25. Lagakos, S. W., and Louis, T. A. (1988). Use of tumour lethality to interpret tumorigenicity experiments lacking cause-of-death data. *Applied Statistics*, **37**, 169–179.

26. Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *Journal of Multivariate Analysis*, **39**, 217–235.

27. Meng, X. L., and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

28. Meng, X. L., and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

29. Nieto-Barajas, L. E., and Walker, S. G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.

30. Racine-Poon, A. H., and Hoel, D. G. (1984). Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics* **40**, 1151–1158.

31. Reiser, B., Guttman, I., Lin, D. K. J., Guess, F. M., and Usher, J. S. (1995). Bayesian inference for masked system lifetime data. *Applied Statistics* **44**, 79–90.

32. Salinas-Torres, V. H., Pereira, C. A. B., and Tiwari, R. C. (2002). Bayesian nonparametric estimation in a series system or a competing-risk model. *Journal of Nonparametric Statistics* **14**, 449–458.

33. Servidea J. D. (2002). Bridge Sampling with Dependent Random Draws: Techniques and Strategy. Ph.D. thesis, Department of Statistics, University of Chicago.

34. Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

35. Walker, S. G., and Mallick, B. K. (1997). Hierarchical generalized linear models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society B* **59**, 845–860.