Biometrika (2004), **91,** 3, *pp.* 543–558

Printed in Great Britain

Inference based on the EM algorithm for the competing risks model with masked causes of failure

BY RADU V. CRAIU

Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, Canada

craiu@utstat.toronto.edu

AND THIERRY DUCHESNE

Département de mathématiques et de statistique, Université Laval, Québec, G1K 7P4, Canada

duchesne@mat.ulaval.ca

SUMMARY

In this paper we propose inference methods based on the EM algorithm for estimating the parameters of a weakly parameterised competing risks model with masked causes of failure and second-stage data. With a carefully chosen definition of complete data, the maximum likelihood estimation of the cause-specific hazard functions and of the masking probabilities is performed via an EM algorithm. Both the E- and M-steps can be solved in closed form under the full model and under some restricted models of interest. We illustrate the flexibility of the method by showing how grouped data and tests of common hypotheses in the literature on missing cause of death can be handled. The method is applied to a real dataset and the asymptotic and robustness properties of the estimators are investigated through simulation.

Some key words: Grouped data; Likelihood ratio test; Missing data; Piecewise-constant hazard; Proportional hazards; Robustness; SEM algorithm; Symmetry assumption; Time-varying masking probability.

1. INTRODUCTION

In many applications of survival analysis with competing causes of failure, the true cause of failure is not identified at the time of the initial data collection. However, it may be possible to restrict the diagnosis to a subset of all causes. In this case we say that the actual failure cause is masked by the restricted group (Flehinger et al., 1998). In practice one may be able to conduct a second-stage analysis, such as autopsy, in which the true cause can be uniquely determined. The second-stage analysis tends to be costly, so usually only a subset of the masked items are analysed further.

Some authors have derived semi- and nonparametric inference procedures in the case with two failure causes and no second-stage analysis, which often occurs in carcinogenicity bioassays: Dinse (1986) proposed nonparametric maximum likelihood estimators of prevalence and mortality; Goetghebeur & Ryan (1990) derived a modified log-rank test for comparing the survival of populations, which they later extended to proportional hazards regression (Goetghebeur & Ryan, 1995); Racine-Poon & Hoel (1984) considered

inference for this model when a probability of death from each missing cause is provided by a pathologist; and Kodell & Chen (1987) tackled the problem via the EM algorithm. In the case of a general number of failure causes and availability of second-stage analysis data, Flehinger et al. (1998, 2002) propose maximum likelihood estimation under a model with nonparametric proportional cause-specific hazards (Flehinger et al., 1998) and a model with completely parametric cause-specific hazards (Flehinger et al., 2002).

When one purpose of the analysis is the estimation of cumulative incidence functions or inference about the competing risks model, proportionality between the cause-specific hazard functions or their complete parametric specification are strong constraints on the model, and methods for testing these assumptions could be useful. In this paper, we propose an approach based on the EM algorithm which allows simple inference, even under very general models. The method can be used to conduct robust likelihood ratio tests, in particular of the proportional hazards assumption, the so-called symmetry assumption which is described in § 4, and some assumptions made in Dinse (1986) and Goetghebeur & Ryan (1995). Furthermore, the framework developed here easily accommodates an extension in which the masking probabilities vary with time, and makes inference with grouped data feasible with only minor modifications. Finally, the estimators of piecewise-constant cause-specific hazard functions along with their variance estimates can be used for the goodness-of-fit assessment of a fully parametric model or for exploratory data analysis.

2. COMPLETE AND OBSERVED-DATA LIKELIHOODS

2·1. *Data and notation*

Suppose that we observe n independent items and that each item can fail as a result of exactly one of J possible causes. The data collection is performed in two stages. In the first stage, for each item i ($i = 1, \ldots, n$), we observe one of three possible outcomes: i fails because of cause j_i at time t_i; i fails because of a cause that is not known precisely, but is known to belong to a group of failure causes $g_i \subset \{1, \ldots, J\}$; or i had still not failed by time t_i. For some of the items whose cause of failure was not determined uniquely at the first stage, a second-stage analysis is done and a unique cause of failure is determined. Therefore, some of the uncensored items will have a masking group instead of a failure cause, and all the items have a failure time or a censoring time. We define the complete data as the dataset that would be obtained if every item with a masked failure time were sent to a second-stage analysis. The observation for item i in this complete dataset would be $(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_{G+J}}, \delta_{i1}, \ldots, \delta_{iJ})$, 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 $g = \{j\}$. Also, 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} ($j = 1, \ldots, J$) are 0. In addition, $G + J$ is the total number of masking groups in the dataset, including the J groups consisting of the individual failure causes. The masking groups containing more than one cause will be called proper and in the remainder of this paper we will assume that $g_1, \ldots, g_G$ are proper.

2·2. *General model*

Our statistical model is made up of two parts, a competing risks part for the failure times and failure causes and a masking part for the masking group indicators. The competing risk part is the usual competing risks model found in the survival analysis literature

(Lawless, 2003, Ch. 9). Let T^* denote the failure time and J^* denote the failure cause. Define the counting processes $N_{ij}(t) = I(T_i^* \leqslant t, J^* = j, i \text{ uncensored})$, for $i = 1, \ldots, n$, $j = 1, \ldots, J$ and $t \in [0, \tau]$, and the at-risk indicators $Y_i(t) = I$ (i at risk at t), for $i = 1, \ldots, n$ and $t \in [0, \tau]$. Put $\bar{Y}(t) = \sum_i Y_i(t)$ and $\bar{N}_{.j}(t) = \sum_i N_{ij}(t)$. The cause-specific hazard functions are defined as

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{\text{pr}\,(t < T^* \leqslant t + h, J^* = j | T^* \geqslant t)}{h} \quad (j = 1, \ldots, J). \tag{1}$$

From equation (1) we obtain that $S(t) = \text{pr}\,(T^* > t) = \exp\{-\int_0^t \sum_{j=1}^J \lambda_j(u)\,du\}$. The cumulative incidence functions are $F_j(t) = \text{pr}\,(T^* \leqslant t, J^* = j) = \int_0^t \lambda_j(t) S(t)\,dt$. Regression can be handled by letting $\lambda_j(t)$ depend on a vector of covariates. This may be done with a proportional hazards model, either parametrically (Lawless, 2003, Ch. 9) or semi-parametrically (Goetghebeur & Ryan, 1995; Lawless, 2003, Ch. 9). Given a failure of cause j at time t, from the model assumptions it results that the $\{\gamma_{ig} : g \in \{g_1, \ldots, g_{G+J}\}, g \ni j\}$ have a multinomial distribution with total 1 and probabilities given by the masking probabilities

$$P_{g|j}(t) = \text{pr}\,(\text{cause masked to group } g \text{ at stage } 1 | T^* = t, J^* = j) \quad (j \in g). \tag{2}$$

Of obvious interest to practitioners and of prime importance in our EM algorithm are the diagnostic probabilities (Flehinger et al., 1998, 2002)

$$\pi_{j|g}(t) = \text{pr}\,(\text{actually failed of cause } j | \text{failed at time } t \text{ and failure cause masked in } g).$$

Using Bayes' rule one obtains

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

Let θ be the parameter vector specifying $\lambda_j(.)$ and $P_{g_m|j}(.)$, for $j = 1, \ldots, J$ and $m = 1, \ldots, G + J$. Denote by $\mathscr{G}_j = \{g : j \in g\}$ the set of all masking groups that include cause j and let $\mathscr{G}_j^* = \mathscr{G}_j \backslash \{j\}$. Using equations (1) and (2) we obtain the loglikelihood function under the complete data as follows:

$$\begin{aligned} l_C(\theta) = \sum_{i=1}^n \sum_{j=1}^J \int_0^\tau \Bigg(& \{\log \lambda_j(t)\,dN_{ij}(t) - Y_i(t)\lambda_j(t)\,dt\} \\ & + \Bigg[\Bigg(1 - \sum_{g \in \mathscr{G}_j^*} \gamma_{ig}\Bigg) \log \Bigg\{1 - \sum_{g \in \mathscr{G}_j^*} P_{g|j}(t)\Bigg\} \\ & + \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \log P_{g|j}(t) \Bigg] dN_{ij}(t) \Bigg). \end{aligned} \tag{4}$$

For right-censored observations, the term inside square brackets in equation (4) vanishes, and hence the γ_{ig} are not needed for right-censored observations.

Under masking some δ_{ij}, or equivalently $dN_{ij}(t_i)$, will be missing. Let $\mathscr{M}$ denote the set of items that have a masked failure cause in stage 1 and for which there is no stage-2 analysis. For any $i \in \mathscr{M}$, the γ_{ig} will be known so let g_i be the masking group for item i, that is $\gamma_{ig_i} = 1$. Since we assume a single cause of failure, the vector $\{\delta_{ij} : j \in g_i\}$ follows

a multinomial distribution with total 1 and probabilities given by $\pi_{j|g}(t_i)$, for $j \in g_i$. Hence, the loglikelihood of the missing data given the observed data can be written as $l_{\mathscr{M}}(\theta) = \sum_{i \in \mathscr{M}} \sum_{j \in g_i} \delta_{ij} \log \pi_{j|g_i}(t_i)$. Using the well-known identity on which the EM algorithm is based (Dempster et al., 1977; Wu, 1983) and denoting the expectation conditional on the observed data by $E(.|\text{OBS})$, we can write the observed loglikelihood as

$$\begin{aligned} l_{\text{OBS}}(\theta) &= Q(\theta|\theta') - H(\theta|\theta') = E_{\theta'}\{l_{\text{C}}(\theta)|\text{OBS}\} - E_{\theta'}\{l_{\mathscr{M}}(\theta)|\text{OBS}\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^{\tau} \Bigg([\log \lambda_j(t) E_{\theta'}\{dN_{ij}(t)|\text{OBS}\} - Y_i(t)\lambda_j(t)\,dt] \\ &\qquad + \left[\left(1 - \sum_{g \in \mathscr{G}_j^*} \gamma_{ig}\right) \log \left\{1 - \sum_{g \in \mathscr{G}_j^*} P_{g|j}(t)\right\} + \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \log P_{g|j}(t) \right] \\ &\qquad \times E_{\theta'}\{dN_{ij}(t)|\text{OBS}\} \Bigg) - \sum_{i \in \mathscr{M}} \sum_{j \in g_i} E_{\theta'}(\delta_{ij}|\text{OBS}) \log \pi_{j|g_i}(t_i), \end{aligned} \tag{5}$$

where $E_{\theta'}\{dN_{ij}(t)|\text{OBS}\} = 0$ $(t \neq t_i)$ and $E_{\theta'}\{dN_{ij}(t_i)|\text{OBS}\} = E_{\theta}(\delta_{ij}|\text{OBS})$, with

$$E_{\theta}(\delta_{ij}|\text{OBS}) = \begin{cases} 1, & \text{if cause of failure of } i \text{ known to be } j, \\ 0, & \text{if cause of failure of } i \text{ known not to be } j, \\ \pi_{j|g_i}(t_i), & \text{if cause of } i \text{ masked in } g_i \text{ and no stage-2 data for } i. \end{cases}$$

The model with loglikelihood (5) does not consider the selection process of the masked items sent to a second-stage analysis, so inferences based on it are only correct if data are missing at random (Rubin, 1987, p. 53); the data are missing at random when pr(item i is masked|OBS) does not depend on the missing δ_{ij}. This is a weak assumption as it allows the selection of items to be sent to stage 2 to be based on the observed times to failure and/or masking groups.

If all the masked items were sent to the second stage, then equation (4) would imply that $l_{\text{C}}(\theta)$ can be split into $l_{\text{C}}(\theta) = l_{\text{I}}(\lambda) + l_{\text{II}}(p)$, where

$$l_{\text{I}}(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^{\tau} \{\log \lambda_j(t)\,dN_{ij}(t) - Y_i(t)\lambda_j(t)\,dt\},$$

$$l_{\text{II}}(p) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left[\left(1 - \sum_{g \in \mathscr{G}_j^*} \gamma_{ig}\right) \log \left\{1 - \sum_{g \in \mathscr{G}_j^*} P_{g|j}(t)\right\} + \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \log P_{g|j}(t) \right] dN_{ij}(t), \tag{6}$$

λ represents the parameters of the competing risks part of the model and p represents the parameters of the masking part. With complete data the maximum likelihood estimates of λ and p are obtained independently. However, under masking the partition in (6) cannot be performed because the $E_{\theta}(\delta_{ij}|\text{OBS})$ are functions of both λ and p for any item i that has a masked failure cause. Simulations in § 5 suggest that inferences about the masking probabilities, $P_{g|j}(t)$, retain the robustness to the misspecification of the hazards even in the presence of missing data.

2·3. *Specific models*

Our preferred specification of the cause-specific hazard functions is the piecewise-constant model, i.e.

$$\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t), \tag{7}$$

where $0 = a_0 < a_1 < \ldots < a_K = \tau$, and $1_k(t)$ is the indicator that $t \in (a_{k-1}, a_k]$. While we have assumed, for ease of notation, that the cut-points $a_0, \ldots, a_K$ are the same for all failure causes, this is not necessary in order to carry out all the computations in closed form. Advantages of the piecewise-constant specification include closed-form expressions for the maximum likelihood estimators under complete data in both the unrestricted and proportional cause-specific hazards case, convergence of the EM algorithm under mild assumptions, and flexibility in the shape of the hazard functions. In this case the complete-data likelihood function is maximised by

$$\hat{\lambda}_{jk} = \frac{\sum_{i=1}^{n} \delta_{ij} 1_k(t_i)}{e_k}, \tag{8}$$

where $e_k = \sum_{i=1}^{n} \int_0^{t_i} 1_k(u)\, du$ denotes the total time lived by all items, i.e. the exposure, in the interval $(a_{k-1}, a_k]$; see Lawless (2003, Ch. 9) for the derivation of (8) and of the maximum likelihood estimators of the survivor and cumulative incidence functions.

Other specifications of $\lambda_j(t)$ that also work in the context of masked failure causes include the semiparametric proportional hazards model with $\lambda_j(t) = r_j \lambda(t)$, with $\lambda(t)$ arbitrary and $\sum_j r_j = 1$. Let $\Lambda(t) = \int_0^t \lambda(u)\, du$. In this case parameter estimators are $\hat{r}_j = \sum_i \delta_{ij} / \sum_i \sum_j \delta_{ij}$ and $\hat{\Lambda}(t) = \sum_j \int_0^t d\bar{N}_{\cdot j}(u) / \bar{Y}(u)$. A more general model that includes both the piecewise-constant and proportional hazards models is the piecewise-proportional model with $\lambda_j(t) = \sum_k r_{jk} \lambda(t) 1_k(t)$, with $\sum_j r_{jk} = 1$. Then the estimator of $\Lambda(t)$ is the same as in the proportional hazards model and $\hat{r}_{jk} = \sum_i \delta_{ij} 1_k(t_i) / \sum_i \sum_j \{\delta_{ij} 1_k(t_i)\}$. Note that in semiparametric cases the diagnostic probability estimator is

$$\hat{\pi}_{j|g_i}(t_i) = \frac{d\hat{\Lambda}_j(t_i) \hat{P}_{g_i|j}(t_i)}{\sum_{l \in g_i} d\hat{\Lambda}_l(t_i) \hat{P}_{g_i|l}(t_i)}.$$

Some authors, such as Flehinger et al. (2002), have also used a specification from a parametric family, such as the Weibull family $\lambda_j(t) = \alpha_j (t/\beta_j)^{\alpha_j - 1} / \beta_j$.

For the masking probabilities $P_{g|j}(t)$, previous work in the literature has extensively considered the time-fixed model $P_{g|j}(t) \equiv P_{g|j}$. In this case the maximum likelihood estimator of $P_{g|j}$ under complete data is given by

$$\hat{P}_{g|j} = \frac{\sum_{i=1}^{n} \delta_{ij} \gamma_{ig}}{\sum_{i=1}^{n} \delta_{ij}}. \tag{9}$$

We focus on this latter specification, though the framework of this paper allows for time-varying masking probabilities. For instance we can consider a piecewise-constant model of the form $P_{g|j}(t) = \sum_k P_{g|j}(k) 1_k(t)$, with maximum likelihood estimator

$$\hat{P}_{g|j}(k) = \frac{\sum_i \gamma_{ig} \delta_{ij} 1_k(t_i)}{\sum_i \delta_{ij} 1_k(t_i)}, \tag{10}$$

which is just an interval-based version of (9). A drawback of the model with time-varying masking probabilities is the large number of parameters involved, which makes it less appealing for applications with a small number of observations and/or a large number of failure causes and masking groups. In § 5, we use (10) to test for the hypothesis of time-fixed masking probabilities in an analysis of hard drive failures.

2·4. *Grouped data*

The model with piecewise-constant cause-specific hazards can handle grouped data with little modification. Suppose that we are given the same observations regarding the failure causes and the masking groups, but that we are no longer given exact failure times, but rather the number of items that failed or were censored in each interval $(a_{k-1}, a_k]$, for $k = 1, \ldots, K$. Let d_{jk} represent the number of failures of cause j in the interval $(a_{k-1}, a_k]$ and let l_k be the number of observations censored between a_{k-1} and a_k. Dykstra et al. (1995) give the loglikelihood for the competing risks part of the model with grouped data and, upon substitution in (4), we obtain

$$l_{\mathrm{C}}^{\mathrm{Grouped}}(\theta) = \sum_{k=1}^{K} \left\{ (n_k - d_{.k}) \log\left(1 - \sum_{j=1}^{J} \lambda_{jk}\right) + \sum_{j=1}^{J} d_{jk} \log \lambda_{jk} \right\} + \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} \left\{ \left(1 - \sum_{g \in \mathscr{G}_j^*} \gamma_{ig}\right) \log\left(1 - \sum_{g \in \mathscr{G}_j^*} P_{g|j}\right) + \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \log P_{g|j} \right\}, \tag{11}$$

where $n_k = \sum_{m=k}^{K} (l_m + \sum_{j=1}^{J} d_{jm})$ is the number of items still at risk, i.e. alive and uncensored, just before time a_k and $d_{.k} = \sum_{j=1}^{J} d_{jk}$. The missing data under masking will be the $d_{jk} = \sum_{i=1}^{n} \delta_{ij} 1_k(t_i)$. Hence $E(d_{jk}|\mathrm{OBS}) = \sum_{i=1}^{n} E(\delta_{ij}|\mathrm{OBS}) 1_k(t_i)$, with $E(\delta_{ij}|\mathrm{OBS})$ still given by (6). The observed-data likelihood is obtained by replacing $Q(\theta|\theta')$ based on $l_{\mathrm{C}}(\theta)$ in (5) by that based on $l_{\mathrm{C}}^{\mathrm{Grouped}}(\theta)$ given by (11).

3. EM ALGORITHM

3·1. *The algorithm*

The simplicity of the complete-data loglikelihood (4) and the fact that it is linear in the missing data make the EM algorithm a fitting tool for this problem. The observed-data loglikelihood (5) is maximised, once a starting point is chosen, by iterating between the following two steps.

E-step. Compute $E_{\hat{\theta}^{(l-1)}}(\delta_{ij}|\mathrm{OBS})$ using (3) and (6).

M-step. Maximise $E_{\theta'}(l_{\mathrm{C}}(\theta)|\mathrm{OBS})$ with respect to θ. For instance, in the piecewise-constant model with time-fixed probabilities, this amounts to setting

$$\hat{\lambda}_{jk}^{(l)} = \frac{\sum_{i=1}^{n} E_{\hat{\theta}^{(l-1)}}(\delta_{ij}|\mathrm{OBS}) 1_k(t_i)}{e_k}, \quad \hat{P}_{g|j}^{(l)} = \frac{\sum_{i=1}^{n} E_{\hat{\theta}^{(l-1)}}(\delta_{ij}|\mathrm{OBS}) \gamma_{ig}}{\sum_{i=1}^{n} E_{\hat{\theta}^{(l-1)}}(\delta_{ij}|\mathrm{OBS})}. \tag{12}$$

In § 4 we show that the E-step and M-step can be solved in closed form even when the parameters are restricted to some subspaces of interest.

In the case of grouped data, the E-step remains unchanged but in the M-step one must use $\hat{\lambda}_{jk}^{(l)} = \sum_{i=1}^{n} E_{\hat{\theta}^{(l-1)}}(\delta_{ij}|\mathrm{OBS}) 1_k(t_i)/n_k$. In the special case of grouped data with a single interval and time-fixed masking probabilities, Flehinger et al. (1996) obtain the maximum

likelihood estimators of θ under observed data in closed form. Indeed, it is easy to check that the EM algorithm of § 3·1 reaches a fixed point after a single M-step when using $\hat{\pi}_{j|g} = n_{g,j}/n_g^+$, where $n_{g,j}$ is the number of items whose cause is masked in group g at stage 1 and found to be cause j at stage 2, and n_g^+ denotes the number of items masked in g at stage 1 that went on to the second stage.

3·2. *Convergence of the algorithm*

We use the results of Wu (1983) to show that the algorithm of § 3·1 does converge to a stationary point in the case with piecewise constant hazards and masking probabilities.

Let $\Omega \subset R^d$ denote the parameter space of the model. The dimension d depends on the number of intervals on which the hazards and masking probabilities are assumed constant, the number of masking groups and the number of failure causes. For ease of notation, assume that the cut-points for the hazard and masking probability intervals are the same. Suppose we choose these cut-points so that, for each interval k and each failure cause j, there exists an i such that $j \in g_i$ and $1_k(t_i) = 1$.

LEMMA 1. *With this choice of* $\{a_k, 0 \leqslant k \leqslant K\}$, *for any* θ_0 *such that* $l_{\mathrm{OBS}}(\theta_0) > -\infty$, *the set* $\Omega_{\theta_0} = \{\theta \in \Omega : l_{\mathrm{OBS}}(\theta) \geqslant l_{\mathrm{OBS}}(\theta_0)\}$ *is compact in* R^d.

Proof. From (12), for any interval k and failure cause j and any masking group g which includes j, $1/n \leqslant P_{g|j}(k) \leqslant (n-1)/n$ and $1/e_k \leqslant \lambda_{jk} \leqslant \sum_{i=1}^n 1_{I_k}(t_i)/e_k$. Since the map $l_{\mathrm{OBS}}(\theta)$ is continuous, the set $\Omega_{\theta_0} = \{\theta \in \Omega : l_{\mathrm{OBS}}(\theta) \geqslant l_{\mathrm{OBS}}(\theta_0)\}$ is a closed subset of a compact set, and therefore it is also compact. □

Lemma 1 implies that the conditions (5)–(7) in Wu (1983) are satisfied. Also, the map $Q(\theta|\theta')$ is continuous in both θ and θ', which implies (Wu, 1983, Theorem 2) that the limit points of any instance $\theta^{(l)}$ of the EM algorithm are stationary points of l_{OBS}, and $l_{\mathrm{OBS}}(\theta^{(l)})$ converges monotonically to $l_{\mathrm{OBS}}(\theta^*)$ for some stationary point θ^*.

Our simulations show that the algorithm converges quickly, in typically less than 20 steps, when the tolerance in estimate changes is 10^{-8} and even when the number of items is in the thousands. There was no evidence of multimodality when starting the algorithm at different initial values.

3·3. *Absence of second-stage data*

If no item is sent to a second-stage analysis, the algorithm proposed here still works as long as the hazards are not proportional. In this case the convergence of the EM algorithm is extremely slow, requiring thousands of iterations. This is not surprising, given the direct connection between the amount of missing information and the convergence rate of EM (Dempster et al., 1977; Meng & Rubin, 1991).

Under proportionality of the hazards and no second-stage data, the parameters of the model under the observed data are unidentifiable (Flehinger et al., 1998). However, the EM algorithm still converges since the parameters are identifiable given the complete data. Dempster et al. (1977) state that, when the parameters are identifiable only in the complete-data model, there is a ridge of local maxima in the observed-data likelihood surface and the EM algorithm will converge to one of the points on the ridge, depending on the starting point. The erratic behaviour of EM can be detected by using multiple starting points. In

such situations, we suggest using one of the alternative approaches used in the literature to bypass the issue of over-parametrisation, such as the symmetry assumption (Miyakawa, 1984; Dinse, 1986; Goetghebeur & Ryan, 1990, 1995; Lo, 1991).

4. INFERENCE PROCEDURES

4·1. *Likelihood ratio tests*

Since we can calculate the values of the observed-data loglikelihood via (5), tests based on the likelihood ratio statistic are readily obtained. Hypotheses of interest that can be easily tested in this way include the time-fixed masking probabilities assumption, $P_{g|j}(t) = P_{g|j}$, the symmetry assumption, $P_{g|j} = P_g$, the equal stage 1 identifiability assumption, $P_{\{1\}|1} = \ldots = P_{\{J\}|J}$, and the proportionality of cause-specific hazards assumption, $\lambda_j = \phi_j \lambda_1$. We illustrate the method by deriving tests of the hypotheses of symmetry and of proportional cause-specific hazards.

The symmetry assumption states that the conditional probabilities $P_{g|j}$ are independent of j. Formally, we have $H: P_{g_i|j} = P_i$ for all $i \in \{1, \ldots, G\}$, and $j \in g_i$. Note that, when the number of possible failure causes is $J = 2$, this symmetry assumption and assumption (ii) of Goetghebeur & Ryan (1990, 1995) are equivalent. Under H_0, complete data, and from (6), $l_{\mathrm{II}}(p)$ is expressed as

$$l_{\mathrm{II,sym}}(p) = \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} \left\{ \left(1 - \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \right) \log \left(1 - \sum_{g \in \mathscr{G}_j^*} P_g \right) + \sum_{g \in \mathscr{G}_j^*} \gamma_{ig} \log P_g \right\}. \quad (13)$$

The maximisation of (13) can be done in closed form if the proper masking groups are nested, that is when $g_1 \subset g_2 \subset \ldots \subset g_G$. The calculations for the general situation are shown in the Appendix and the solutions are found using first $P_G = m_{P_G} / (\sum_{l=1}^{G} m_{g_l} + \sum_{j=1}^{J} n_j)$ and then, recursively for each $1 \leqslant k \leqslant G-1$,

$$P_{G-k} = \frac{m_{P_{G-k}} (1 - \sum_{h=0}^{k-1} P_{G-h})}{\sum_{l=1}^{G-k} m_{g_l} + \sum_{j \in g_{G-k}} n_j},$$

where m_{g_k} represents the number of items whose failure causes are masked in the group g_k at stage 1, $m_{g_k} = \sum_{i=1}^{N} \sum_{j \in g_k} \delta_{ij} \gamma_{ig_k}$, and $n_j = \sum_{i=1}^{N} \delta_{ij} (1 - \sum_{\{k: g_k \in G_j^*\}} \gamma_{ig_k})$ is the number of items which have failed of cause j and the cause was not masked in stage 1. The nested-groups condition is satisfied in the important case of masking with only one masking group (Racine-Poon & Hoel, 1984; Dinse, 1986; Goetghebeur & Ryan, 1990, 1995). In situations where the groups are not nested analytical solution to the maximisation problem (13) is still possible, as can be seen in the simulation study presented in § 5. If a closed-form solution cannot be found, a numerical maximisation method, such as Newton–Raphson, can be implemented.

To perform the likelihood ratio test, we compute the likelihood ratio statistic $R^{\mathrm{SYM}} = 2\{l_{\mathrm{OBS}}(\hat{\theta}) - l_{\mathrm{OBS}}(\hat{\theta}^{\mathrm{SYM}})\}$, where $\hat{\theta}^{\mathrm{SYM}}$ is the maximum likelihood estimate of the parameters under the symmetry assumption. If we denote by G_k the number of causes in the masking group g_k then standard first-order large-sample theory gives an asymptotic distribution for R^{SYM} that is chi-squared on $\sum_{k=1}^{G} G_k - G$ degrees of freedom, and this asymptotic distribution proves to be quite accurate in the simulation study of § 5. We note that the asymptotic theory is applied conditional on the fixed cut-points of the hazards intervals.

Another assumption of interest under masking is that of proportionality between the cause-specific hazard functions. Under the piecewise-constant model, this hypothesis can be stated as $H_0: \lambda_{jk} = \phi_j \lambda_{1k}$, for $j = 2, \ldots, J$ and $k = 1, \ldots, K$, where $\phi_j > 0$, for $j = 2, \ldots, J$, are positive constants. Maximum likelihood estimation under complete data and the constraint given by H_0 can be performed analytically and the calculations are outlined in the Appendix. This leads to a new M-step where the estimators of $P_{g|j}$ are the same as in (12), but with

$$\hat{\lambda}_{1k}^{(l)\mathrm{PH}} = \frac{E_{\hat{\theta}^{(l-1)\mathrm{PH}}}(\sum_{i=1}^{n} \delta_{i1} | \mathrm{OBS})}{N} \times \frac{E_{\hat{\theta}^{(l-1)\mathrm{PH}}}(\sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} 1_k(t_i) | \mathrm{OBS})}{e_k} \quad (k = 1, \ldots, K), \tag{14}$$

$$\hat{\phi}_j^{(l)} = \frac{E_{\hat{\theta}^{(l-1)\mathrm{PH}}}(\sum_{i=1}^{n} \delta_{ij} | \mathrm{OBS})}{E_{\hat{\theta}^{(l-1)\mathrm{PH}}}(\sum_{i=1}^{n} \delta_{i1} | \mathrm{OBS})} \quad (j = 2, \ldots, J). \tag{15}$$

The likelihood ratio statistic is $R^{\mathrm{PH}} = 2\{l_{\mathrm{OBS}}(\hat{\theta}) - l_{\mathrm{OBS}}(\hat{\theta}^{\mathrm{PH}})\}$, where $\hat{\theta}^{\mathrm{PH}}$ denotes the maximum likelihood estimators of the parameters under the proportional hazards assumption. The limit distribution of R^{PH} is chi-squared on $(J-1)(K-1)$ degrees of freedom. We confirm the accuracy of this approximation and the robustness to misspecification of the competing risks model through simulation in § 5.

4·2. *Asymptotic variance estimators*

Different methods for calculating the asymptotic variance-covariance matrix for EM estimates have been devised by, among others, Louis (1982), Meng & Rubin (1991), Oakes (1999) and Meng (2001). Since our EM algorithm can generally be performed with both the E-step and the M-step in closed form, we prefer to use the supplemented EM, SEM, algorithm (Meng & Rubin, 1991) because it uses only the code for EM itself and simple code for standard matrix operations. Any iteration of an EM algorithm defines a mapping $\theta^{(l+1)} = M(\theta^{(l)})$. Meng & Rubin (1991) showed that, with

$$(\mathrm{DM})_{ij} \equiv \left(\frac{\partial M_j(\theta)}{\partial \theta_i} \right) \Bigg|_{\theta = \hat{\theta}_{\mathrm{MLE}}},$$

then $I_{\mathrm{om}} = \mathrm{DM}\, I_{\mathrm{oc}}$, where I_{om} and I_{oc} are the missing and complete information matrices, respectively. In most applications, the Jacobian of the map M cannot be obtained analytically; however, following Meng & Rubin (1991), we determine the elements of DM by running a series of restricted EM algorithms. The procedure is attractive as each of these 'forced' EM algorithms is identical in form to the one designed for the original problem but uses different initial values for $\theta^{(0)}$. The desired variance matrix $V = \mathrm{var}(\hat{\theta})$ is given by $V = I_{\mathrm{oc}}^{-1} + I_{\mathrm{oc}}^{-1} \mathrm{DM} (I - \mathrm{DM})^{-1}$, where I is the identity matrix. In our analyses the SEM algorithm shows stability. The stopping rule of each of the constrained EM algorithms which make up the SEM procedure uses, as recommended by Meng & Rubin (1991), the square root of the tolerance used in the original EM algorithm.

Once an asymptotic variance matrix is obtained, Wald-type confidence intervals can be constructed. For example, for the piecewise-constant hazards, we assume asymptotic

normality of $\log \hat{\lambda}_{jk}$ and use $\hat{\lambda}_{jk} \exp(\pm 1{\cdot}96\hat{\sigma}_{jk}/\hat{\lambda}_{jk})$, where $\hat{\sigma}^2_{jk}$ is the SEM variance estimate of $\hat{\lambda}_{jk}$. Asymptotic variances for the cumulative incidence or survivor functions are obtained via the delta rule. For instance, for cumulative incidence functions, we use equation (9.2.11) of Lawless (2003), with $\mathrm{diag}(\hat{v})$ replaced by V above.

5. APPLICATIONS

5·1. *Simulation study*

Simulation results are obtained with data generated under four different models with piecewise-constant hazards, $M_1, \ldots, M_4$, and four models with Weibull distributed hazard rates, $W_1, \ldots, W_4$. Each model is used to simulate 100 datasets, each containing 1000 items with observed failure times. The models M_i and W_i have the following common features: a 30% probability that a masked item is sent to stage two, three failure causes and three proper masking groups $g_1 = \{1, 2\}$, $g_2 = \{1, 3\}$ and $g_3 = \{1, 2, 3\}$. The models M_i share the same hazard intervals, namely $[a_0, a_1] = [0, 5]$, $(a_1, a_2] = (5, 10]$ and $(a_2, a_3) = (10, \infty)$. We use different masking probabilities $\{P_{g_k|j} : 1 \leqslant k \leqslant 3, 1 \leqslant j \leqslant 3\}$ and different values for the hazard rates $\{\lambda_{jk} : 1 \leqslant j \leqslant 3, 1 \leqslant k \leqslant 3\}$. In fact, we choose these parameters so that we can inspect the properties of the likelihood ratio tests proposed in § 4. Define the matrices $P = (P_{ij} = P_{g_i|j})_{1 \leqslant i,j \leqslant 3}$ and $Q = (Q_{ij} = P_{g_i|j})_{1 \leqslant i,j \leqslant 3}$ by

$$P = \begin{array}{c} \\ \{1,2\} \\ \{1,3\} \\ \{1,2,3\} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0{\cdot}2 & 0{\cdot}4 & 0 \\ 0{\cdot}2 & 0 & 0{\cdot}3 \\ 0{\cdot}2 & 0{\cdot}4 & 0{\cdot}4 \end{pmatrix} \end{array}, \quad Q = \begin{pmatrix} 0{\cdot}3 & 0{\cdot}3 & 0 \\ 0{\cdot}2 & 0 & 0{\cdot}2 \\ 0{\cdot}2 & 0{\cdot}2 & 0{\cdot}2 \end{pmatrix}.$$

Models M_1 and M_4 use the masking model given by P while models M_2 and M_3 use that given by Q and satisfy symmetry. The cause-specific hazard rates λ_j $(j = 1, 2, 3)$ are piecewise constant and are defined by the vectors $\tilde{\lambda}_1 = (0{\cdot}003,\ 0{\cdot}02,\ 0{\cdot}012)$, $\tilde{\lambda}_2 = (0{\cdot}006,\ 0{\cdot}04,\ 0{\cdot}024)$ and $\tilde{\lambda}_3 = (0{\cdot}015,\ 0{\cdot}01,\ 0{\cdot}006)$ for models M_1 and M_2, and for models M_3 and M_4 we take $\tilde{\lambda}_1 = (0{\cdot}003,\ 0{\cdot}02,\ 0{\cdot}012)$, $\tilde{\lambda}_2 = (0{\cdot}0045,\ 0{\cdot}01,\ 0{\cdot}03)$ and $\tilde{\lambda}_3 = (0{\cdot}001,\ 0{\cdot}04,\ 0{\cdot}01)$. For each $i = 1, \ldots, 4$, models W_i and M_i have the same masking probabilities. For models W_1 and W_2, for which the proportional hazards assumption holds, $\lambda_1 \sim \mathrm{Wei}(3, 12)$, $\lambda_2 \sim \mathrm{Wei}(3, 10)$ and $\lambda_3 \sim \mathrm{Wei}(3, 10)$, where $\lambda \sim \mathrm{Wei}(a, b)$ has density $f(t) = (a/b)(t/b)^{a-1} \exp\{-(t/b)^a\}$. For models W_3 and W_4 we use $\lambda_1 \sim \mathrm{Wei}(0{\cdot}9, 12)$, $\lambda_2 \sim \mathrm{Wei}(1{\cdot}5, 10)$ and $\lambda_3 \sim \mathrm{Wei}(2, 10)$.

For inference we use four models with piecewise-constant hazards. In each case the lower bound of the first interval is zero and the upper bound of the last interval is ∞. For data generated using models M_i $(i = 1, 2, 3, 4)$ we consider two possible choices for the remaining cut-points. Both choices correspond to misspecified hazard rates. In model PCF_1 we use two intervals with the separating cut-point equal to the sample median of the failure times, while model PCF_2 has four intervals with the cut-points equal to, respectively, the 25th, 50th and 75th percentiles of the failure times. Similarly, for data generated using models W_1 to W_4 we also consider intervals based on sample quantiles. Model WEIF_1 is with three intervals constructed with the 33rd and 67th percentiles of the failure times, and model WEIF_2 has four intervals defined by the 25th, 50th and 75th percentiles of the failure times.

Note that we do not use nested masking groups. Hence we cannot use the equations derived in § 4·1 to test for symmetry. However, for the three masking groups $g_1 = \{1, 2, 3\}$, $g_2 = \{1, 2\}$ and $g_3 = \{1, 3\}$, and with the notation of § 4·1, we have that under the symmetry assumption the M-step results in

$$P_1 = \frac{m_{g_{123}}}{n_1 + n_2 + n_3 + m_{g_{12}} + m_{g_{13}} + m_{g_{123}}}, \quad P_3 = \frac{B - (B^2 - 4AC)^{\frac{1}{2}}}{2A},$$

$$P_2 = \frac{m_{g_{12}}}{m_{g_{123}}/P_1 - n_3/(1 - P_1 - P_3)},$$

where $A = n_3 + m_{g_{13}} + n_1$,

$$B = n_1^2 + n_1 n_3 + 3n_1 m_{g_{13}} + m_{g_{12}} n_1 + n_2 n_1 + m_{g_{12}} m_{g_{13}} + 2n_2 m_{g_{13}} + 2m_{g_{13}}^2 + 2n_3 m_{g_{13}} + n_2 n_3,$$

$$C = 2m_{g_{13}}^2 n_2 + m_{g_{13}}^3 + m_{g_{13}} n_2^2 + m_{g_{13}} n_1^2 + m_{g_{13}} n_1 m_{g_{12}} + m_{g_{13}} n_2 m_{g_{12}} + m_{g_{13}} n_3 n_1 + 2m_{g_{13}}^2 n_1 + m_{g_{13}} n_2 n_3 + n_3 m_{g_{13}}^2 + m_{g_{13}}^2 m_{g_{12}} + 2m_{g_{13}} n_2 n_1.$$

If we denote the hypothesis of symmetry by A_{SYM} and the proportional hazards hypothesis by A_{PH}, then the accuracy of the likelihood ratio test of A_{SYM} and A_{PH} for each model is summarised in Table 1. The numbers inside each cell represent the observed number of rejections, at the 0·05 level, out of 100 replicates. The results suggest that the tests are robust to the choice of the competing risks part of the model. Plots, not shown, of the empirical cumulative distribution function constructed using the 100 observed-data likelihood ratio statistics together with the cumulative distribution function of the chi-squared distribution with the number of degrees of freedom given in § 4·1, indicate that the chi-squared distribution is a good approximation.

To verify the impact of the missing data on the precision of our estimates we performed comparisons for each parameter under the eight different models M_i, W_i $(i = 1, \ldots, 4)$. The estimates of the masking probabilities are calculated for each of these models 100 times. We restrict the presentation of findings to one of the parameters, as the results are representative of all. Table 2 contains the results concerning $P_{\{1,2,3\}|1}$. The Monte Carlo average column presents the average estimate along with the estimator's standard error, SE_{SEM}, calculated using the SEM algorithm. The standard error of the SE_{SEM} estimates, which was computed using 100 replicates of the experiment, is uniformly less than 0·002. For comparison, in the fourth column, the true value of the parameter and the Monte Carlo

Table 1. *Number of rejections, out of* 100 *samples, of* A_{SYM} *and* A_{PH} *at the* 5% *level for data generated under models* M_i *and* W_i*, with cut-points defined through* PCF_1 *or* PCF_2 *for* $i = 1, \ldots, 4$*, using the likelihood ratio test of* § 4·1

Model	A_{PH}	A_{SYM}	Model	A_{PH}	A_{SYM}
$M_1 \backslash W_1$	True	False	$M_3 \backslash W_3$	False	True
PCF_1	6\2	100\100	PCF_1	100\100	3\8
PCF_2	7\7	100\100	PCF_2	100\100	2\8
$M_2 \backslash W_2$	True	True	$M_4 \backslash W_4$	False	False
PCF_1	8\3	6\9	PCF_1	100\100	100\100
PCF_2	7\5	6\10	PCF_2	100\100	100\100

Table 2. *Mean estimates and standard errors of the masking probability* $P_{\{1,2,3\}|1}$ *when data were generated using models* M_i *and* W_i $(i=1,\ldots,4)$ *and for choices of cut-points specified by* PCF_1, PCF_2 *and* WEIF_1, WEIF_2, *respectively*

Model	Monte Carlo average (SE_{SEM})		True value (SE_{MC})
	PCF_1	PCF_2	
M_1	0·200 (0·031)	0·200 (0·031)	0·200 (0·030)
M_2	0·195 (0·031)	0·195 (0·033)	0·200 (0·028)
M_3	0·196 (0·034)	0·199 (0·031)	0·200 (0·031)
M_4	0·201 (0·035)	0·201 (0·035)	0·200 (0·032)
	WEIF_1	WEIF_2	
W_1	0·195 (0·042)	0·195 (0·040)	0·200 (0·046)
W_2	0·196 (0·048)	0·196 (0·044)	0·200 (0·050)
W_3	0·200 (0·033)	0·201 (0·033)	0·200 (0·038)
W_4	0·196 (0·031)	0·196 (0·031)	0·200 (0·031)

SE_{SEM}, average of the standard error estimates obtained via the SEM algorithm; SE_{MC}, Monte Carlo standard error of $\hat{P}_{\{1,2,3\}|1}$ obtained from 1000 replicates of the data.

estimate of the standard error, SE_{MC}, are also shown. The latter is obtained by calculating the sample standard error for the estimates obtained from 1000 replicates of the data and can be considered a good approximation of the true standard error of the estimator. Table 2 shows that even if the hazard functions are misspecified the estimates of the masking probabilities are still close to the true values. Figure 1 illustrates, for each Weibull

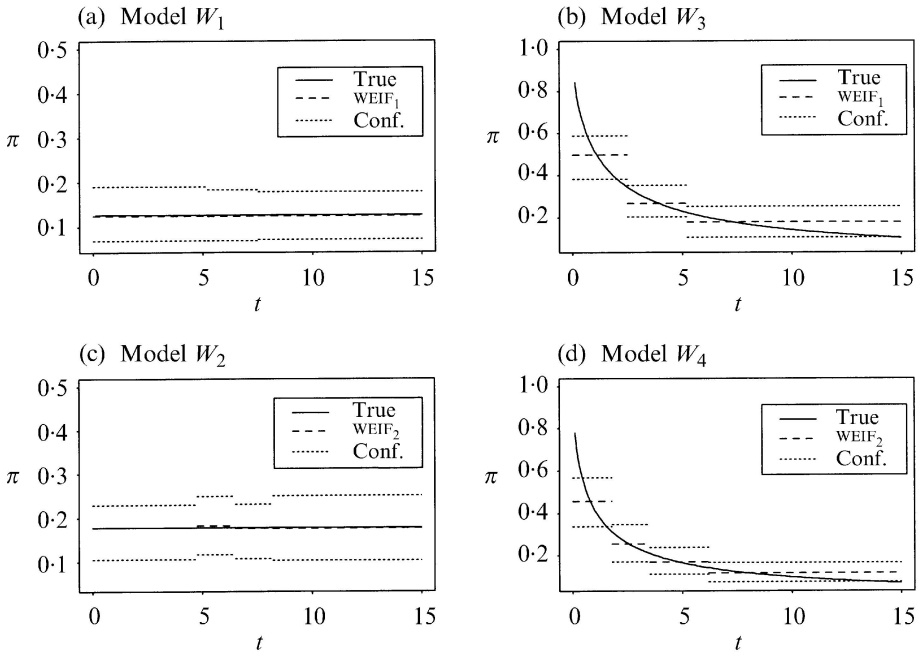

Fig. 1: Simulation study. Plot of the true curve of $\pi_{1|\{1,2,3\}}(t)$ for models (a) W_1, (b) W_3, and (c) W_2, (d) W_4, along with the estimates from WEIF_1 and, respectively, WEIF_2; the dotted lines represent the 5th and 95th Monte Carlo percentiles.

model considered and for both sets of cut-points, the true curve of $\pi_{1|\{1,2,3\}}(t)$ along with the Monte Carlo average, the 5th and the 95th percentiles. Again, a sample of 100 estimates has been obtained by generating 100 datasets from the same underlying population. It is seen that the true curve is always between the 5% and 95% Monte Carlo quantiles except for a short interval close to the origin where the hazard rate decreases sharply, a feature that our models cannot capture. For models W_1 and W_2 the proportional hazards assumption is true so the value of $\pi_{1|\{1,2,3\}}$ is constant in time. It appears that increasing the number of interval cut-points decreases the bias of the estimators. We incur the classical trade-off for this reduction in bias, namely a higher variance.

5·2. *Example*: *Hard-drive reliability*

Flehinger et al. (2002) use a model with Weibull cause-specific hazards to model the reliability of 10 000 hard-drives subject to three possible causes of failure. All the hard-drives were put under test at time 0, and, during the observation period of 4 years, 172 of them failed. Many of the drives had a failure cause that was masked to either $\{1, 3\}$ or $\{1, 2, 3\}$. We use the techniques developed in this paper with the piecewise-constant hazards specification as a model assessment procedure. Table 3 gives the cause-specific hazard estimates obtained thereby. The estimates of the masking probabilities we obtain are identical to the third decimal place to those of Flehinger et al. (2002). Figure 2(a) shows that the Weibull parametric model and the piecewise-constant model are in good agreement for all three failure causes. The confidence limits were calculated with the method of § 4·2. One may argue that the piecewise-constant model seems to be pointing towards bathtub-shaped hazard function for cause 1, but we would need more data to confirm this statement. Figure 2(b) shows the estimates of the cause-specific survivor functions $S_j(t) = \exp\{-\int_0^t \lambda_j(u)\,du\}$, in close agreement with Fig. 1 of Flehinger et al. (2002).

Our methods allow us to test some of the assumptions made by Flehinger et al. (2002). Using the likelihood ratio tests of § 4·1 we obtain likelihood ratio statistics of 5·83 on 5 degrees of freedom and 12·5 on 10 degrees of freedom for the tests of time-fixed masking probabilties against piecewise-constant masking probabilities over the intervals $[0, 2)$ and $[2, 4)$, and $[0, 2)$, $[2, 3)$ and $[3, 4)$, respectively. These observations correspond to respective p-values of 0·32 and 0·25 and hence neither test rejects the time-fixed masking probability assumption at the 0·05 level. We also obtained likelihood ratio statistics of 30·53 on 4 degrees of freedom, $p = 0{\cdot}000004$, and of 7·34 on 2 degrees of freedom, $p = 0{\cdot}0254$, for the proportional hazards and symmetry assumptions, respectively. Flehinger et al. (2002) also rejected the proportional hazards assumption with their Weibull model, with a statistic of 32·2 on 2 degrees of freedom.

Table 3: *Hard-drive data. Cause-specific hazard estimates for causes* $j = 1, 2, 3$, *with standard errors in brackets*

Interval	$j = 1$	$j = 2$	$j = 3$
$(0, 1]$	0·00204 (0·0005)	0·00095 (0·0003)	0·00032 (0·0002)
$(1, 2]$	0·0012 (0·0004)	0·00026 (0·0002)	0·0021 (0·0005)
$(2, 3]$	0·00083 (0·0004)	0·00053 (0·0003)	0·0033 (0·0006)
$(3, 4]$	0·0013 (0·0004)	0·00067 (0·0003)	0·0039 (0·0007)

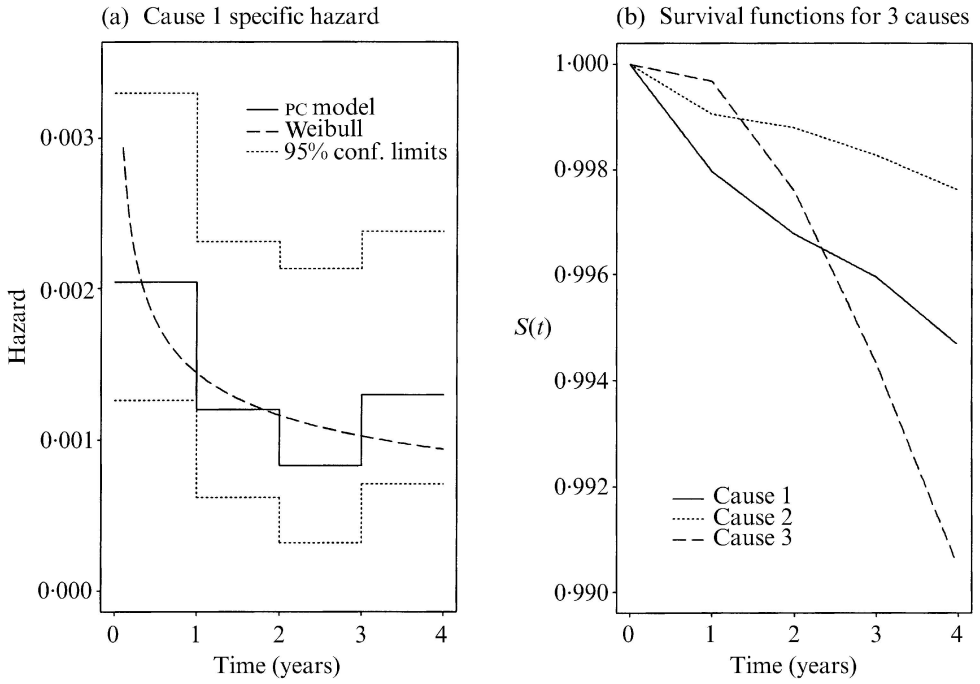

Fig. 2: Hard-drive data. (a) Cause 1 hazard rate estimate based on Weibull model, dashed lines, piecewise-constant model, solid line, along with 95% confidence intervals, dotted lines; (b) estimates of cause-specific survival functions for the three causes.

6. Discussion

The techniques described in this paper can be used on their own or as a complement to the methods of Goetghebeur & Ryan (1990, 1995) or Flehinger et al. (1998, 2002) to assess their goodness of fit or to guide with model selection. Our method is flexible enough to handle cases with only one-stage data or grouped data, and to allow the masking probabilities to depend on time. Moreover, simulations suggest that likelihood based inference is robust to misspecification of the competing risk part of the model.

Problems such as how to select optimally the items to be sent to a stage-2 analysis, or how to analyse data when first-stage diagnosis may actually be wrong, have received little attention in the literature and could represent interesting applications and/or generalisations of the results presented in this paper.

Acknowledgement

We thank David Andrews, Jerry Lawless, Xiao-Li Meng, the editor, an associate editor and two anonymous referees for their constructive comments. This research was supported by individual operating grants to each author by the Natural Sciences and Engineering Research Council of Canada, an individual Connaught grant to R. V. Craiu and an individual grant to T. Duchesne from the Fonds québécois de la recherche sur la nature et les technologies.

Appendix

Details of the M*-steps*

In the M-step with symmetry constraint assume that the proper masking groups are nested, that is $g_1 \subset g_2 \subset \ldots \subset g_M = \{1, 2, \ldots, J\}$. For each $k \in \{2, \ldots, M\}$, define $A_k = g_k \backslash g_{k-1}$. We use the notation from § 4·1. The M-step requires the solution to the following system of equations:

$$\frac{m_{P_k}}{P_k} = \sum_{j \in g_k} \frac{n_j}{1 - \sum_{\{h: g_h \in \mathscr{G}_j^*\}} P_h}, \tag{A1}$$

for all $k = 1, \ldots, M$ and all $j \in g_k$.

We first show by mathematical induction that the solution to (A1) also satisfies

$$\sum_{j \in g_k} \frac{n_j}{1 - \sum_{\{h: g_h \in \mathscr{G}_j^*\}} P_h} = \frac{\sum_{l=1}^{k-1} m_{g_l} + \sum_{j \in g_k} n_j}{1 - \sum_{h=k}^{M} P_h}, \tag{A2}$$

for any $2 \leqslant k \leqslant M$. For $k = 2$, (A2) results from (A1) because, as a result of the nesting,

$$\frac{m_{g_1}}{P_1} = \sum_{j \in g_1} \frac{n_j}{1 - \sum_{h=1}^{M} P_h} \Rightarrow \sum_{j \in g_1} \frac{n_j}{1 - \sum_{h=1}^{M} P_h} = \frac{\sum_{j \in g_1} n_j + m_{g_1}}{1 - \sum_{h=2}^{M} P_h},$$

and then

$$\frac{m_{g_2}}{P_2} = \sum_{j \in g_2} \frac{n_j}{1 - \sum_{h=1}^{M} P_h} = \frac{\sum_{j \in g_1} n_j}{1 - \sum_{h=1}^{M} P_h} + \frac{\sum_{j \in A_2} n_j}{1 - \sum_{h=2}^{M} P_h} = \frac{m_{g_1} + \sum_{j \in g_1} n_j}{1 - \sum_{h=2}^{M} P_h}.$$

Assume that (A2) is true for all $2 \leqslant h \leqslant k$ with $k < M$; then (A1) for $k + 1$ gives

$$\frac{m_{k+1}}{P_{k+1}} = \sum_{j \in g_{k+1}} \frac{n_j}{1 - \sum_{\{h: g_h \in \mathscr{G}_j^*\}} P_h} = \sum_{j \in g_k} \frac{n_j}{1 - \sum_{\{h: g_h \in \mathscr{G}_j^*\}} P_h} + \sum_{j \in A_{k+1}} \frac{n_j}{1 - \sum_{\{h: g_h \in \mathscr{G}_j^*\}} P_h} \Rightarrow$$

$$\frac{m_{k+1}}{P_{k+1}} = \frac{\sum_{l=1}^{k} m_{g_l} + \sum_{j \in g_k} n_j + \sum_{j \in A_{k+1}} n_j}{1 - \sum_{h=k+1}^{M} P_h}.$$

This proves (A2) for all $1 \leqslant k \leqslant M$. A solution to the system (A1) is obtained iteratively by first replacing $k = M$ in (A1) and using (A2) to obtain

$$P_M = \frac{m_{P_M}}{\sum_{l=1}^{M} m_{g_l} + \sum_{j=1}^{J} n_j},$$

and then, using simple manipulations of (A1) and (A2), to obtain that, for each $1 \leqslant k \leqslant M - 1$,

$$P_{M-k} = \frac{m_{P_{M-k}}(1 - \sum_{h=0}^{k-1} P_{M-h})}{\sum_{l=1}^{M-k} m_{g_l} + \sum_{j \in g_{M-k}} n_j}.$$

In the M-step with proportional hazards constraint the maximisation step is changed only for the hazard rates λ_{jk}. Substituting $\lambda_{jk} = \phi_j \lambda_{1k}$ in $l_C(\theta)$ given by (4), we obtain the following set of score equations:

$$\sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\delta_{ij} 1_k(t_i)}{\lambda_{1k}} = e_k \sum_{j=1}^{J} \phi_j \quad (\text{for all } k \in \{1, \ldots, K\}),$$

where $e_k = \sum_{i=1}^{n} \int_0^{t_i} 1_k(u)\, du = 0$, and

$$\sum_{i=1}^{n} \delta_{ij}/\phi_j = \sum_{h=1}^{K} \lambda_{1h} e_h \quad (\text{for all } j \in \{1, \ldots, J\}). \tag{A3}$$

If $u_k = \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} 1_k(t_i)$, $v_j = \sum_{i=1}^{n} \delta_{ij}$ and $x_h = \lambda_{1h} e_h$, for all $h \in \{1, \ldots, K\}$, then

$$\sum_{j=1}^{J} v_j = \sum_{k=1}^{K} u_k = N$$

and

$$\frac{x_k}{u_k} = \frac{1}{1+\sum_{j=2}^{J}\phi_j} \quad (\text{for all } k \in \{1,\ldots,K\}) \Rightarrow \frac{\sum_{k=1}^{K} x_k}{\sum_{k=1}^{K} u_k} = \frac{1}{1+\sum_{j=2}^{J}\phi_j}. \tag{A4}$$

Equations (A3) and (A4) yield

$$\frac{v_j}{N} = \frac{\phi_j}{1+\sum_{j=1}^{J}\phi_j} \quad (\text{for all } j \in \{2,\ldots,J\}) \Rightarrow \frac{N-v_1}{N} = \frac{\sum_{j=1}^{J}\phi_j}{1+\sum_{j=1}^{J}\phi_j} \Rightarrow 1-\frac{v_1}{N} = 1-\frac{x_1}{u_1}.$$

Therefore,

$$\hat{\lambda}_{1k} = \frac{v_1 u_k}{e_k} = \frac{\sum_{i=1}^{N}\delta_{i1}}{N}\,\frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\delta_{ij}1_k(t_i)}{e_k} \quad (\text{for all } k \in \{1,\ldots,K\}),$$

$$\hat{\phi}_j = \frac{v_j}{\sum_{k=1}^{K} x_k} = \frac{\sum_{i=1}^{N}\delta_{ij}}{\sum_{i=1}^{N}\delta_{i1}} \quad (\text{for all } j \in \{1,\ldots,J\}).$$

REFERENCES

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

DINSE, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause of death data. *J. Am. Statist. Assoc.* **81**, 328–35.

DYKSTRA, R., KOCHAR, S. & ROBERTSON, T. (1995). Likelihood based inference for cause specific hazard rates under order restriction. *J. Mult. Anal.* **54**, 163–74.

FLEHINGER, B. J., REISER, B. & YASHCHIN, E. (1996). Inference about defects in the presence of masking. *Technometrics* **38**, 247–55.

FLEHINGER, B. J., REISER, B. & YASHCHIN, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika* **85**, 151–64.

FLEHINGER, B. J., REISER, B. & YASHCHIN, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Anal.* **8**, 177–203.

GOETGHEBEUR, E. & RYAN, L. (1990). A modified log rank test for competing risks with missing failure types. *Biometrika* **77**, 171–64.

GOETGHEBEUR, E. & RYAN, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821–33.

KODELL, R. L. & CHEN, J. J. (1987). Handling cause of death in equivocal cases using the EM algorithm (with Rejoinder). *Commun. Statist.* A **16**, 2565–85.

LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. New York: Wiley.

LO, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *J. Mult. Anal.* **29**, 217–35.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226–33.

MENG, X. L. (2001). Discussion of 'Optimization transfer using surrogate objective functions', by K. Lange, D. Hunter and I. Yang. *J. Comp. Graph. Statist.* **9**, 35–43.

MENG, X. L. & RUBIN, D. B. (1991). Using EM to obtain asymptotic variance: The SEM algorithm. *J. Am. Statist. Assoc.* **86**, 899–909.

MIYAKAWA, M. (1984). Analyses of incomplete data in competing risk model. *IEEE Trans. Reliab.* **33**, 293–6.

OAKES, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Statist. Soc.* B **61**, 479–82.

RACINE-POON, A. H. & HOEL, D. G. (1984). Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics* **40**, 1151–8.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

WU, C. F. J. (1983). On the convergence of the EM algorithm. *Ann. Statist.* **11**, 95–103.

[*Received February* 2003. *Revised October* 2003]