Model Selection for the Competing-Risks Model With and Without Masking

Radu V. CRAIU

Department of Statistics
University of Toronto
Toronto, ON M5S 3G3, Canada
(*craiu@utstat.toronto.edu*)

Thomas C. M. LEE

Department of Statistics
Colorado State University
Fort Collins, CO 80523-1877
(*tlee@stat.colostate.edu*)

The competing-risks model is useful in settings in which individuals (or units) may die (or fail) because of various causes. It can also be the case that for some of the items, the cause of failure is known only up to a subgroup of all causes, in which case we say that the failure is group-masked. A widely used approach for competing-risks data with and without masking involves the specification of cause-specific hazard rates. Often, because of the availability of likelihood methods for estimation and testing, piecewise constant hazards are used. The piecewise constant rates also offer model flexibility and computational convenience. However, for such piecewise constant hazard models, the choice of the endpoints for each interval on which the hazards are constant is usually a subjective one. In this article we discuss and propose the use of model selection methods that are data-driven and automatic. We compare three model selection procedures based on the minimum description length principle, the Bayes information criterion, and the Akaike information criterion. A fast-splitting algorithm is the computational tool used to select among an enormous number of possible models. We test the effectiveness of the methods through numerical studies, including a real dataset with masked failure causes.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Code length; Competing risks; EM algorithm; Group masked cause; Minimum description length principle; Missing data; Model selection; Piecewise constant hazard.

1. INTRODUCTION

In many survival data studies, the individuals under study can experience any one of J types of failure. Consequently, each individual/item under study has associated with it J potential failure times, one time for each possible failure. Obviously, in practice only one of the potential failure times is observed, unless the item is right-censored, that is, it does not fail before the end of the study, in which case no failure time is observed. The competing-risks problem involves estimating failure rates for each type of failure. An additional complication arises when a subset of the individuals has a cause of failure known to belong only to a certain subset of all possible causes; in other words, their cause of failure is *group-masked*. In practice, one possibility is to conduct a second-stage analysis, such as autopsy, in which the true cause can be uniquely determined. In fact, inference is possible even if not all items are subjected to a second-stage analysis, as we discuss in Section 2.2.

Examples of failure data obtained under a competing-risks model are abundant in the literature and range from survival analysis studies in biostatistics to applications of reliability in engineering to risk models in actuarial science. For instance, Gaynor et al. (1993) and Barrett et al. (1989) discussed the importance of estimating the probability of death due to cancer relapse after treatment versus the probability of death due to treatment-related complications, providing an example where the competing risks are not acting independently. In reliability studies, Sun and Tiwari (1997) analyzed the failure times of small electrical appliances that may fail due to two competing risks, whereas Taylor (1994) used competing risks to model the probability distribution of the tensile strength of certain materials known to contain two or more subpopulations of flaw types. Lapidus, Braddock, Schwartz, Banco, and Jacobs (1994) presented a study of motorcycle fatalities in which 40% of the death certificates have missing information.

Parametric analyses of the competing-risks model were proposed by Hoel (1972), Moeschberger and David (1971), Lagakos (1977), and Prentice et al. (1978). Cause-specific hazard functions were used in nonparametric estimation by Nelson (1969), Aalen (1978), and Crowder (2001). Semiparametric methods based on proportional-hazards models were discussed by Holt (1978), Kalbfleisch and Prentice (2002, chap. 8), and Lawless (2003, chap. 9).

For the competing-risks model in which a subset of all items have masked causes of failure, some authors have derived semiparametric 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 and Ryan (1990) derived a modified log-rank test for comparing the survival of populations, which they later extended to proportional hazards regression (Goetghebeur and Ryan 1995); Racine-Poon and Hoel (1984) considered inference for this model when a probability of death from each missing cause is provided by a pathologist; and Kodell and 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, Reiser, and Yashchin (1998, 2002) proposed 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.

TECHNOMETRICS, NOVEMBER 2005, VOL. 47, NO. 4
DOI 10.1198/004017005000000238

2002). Craiu and Duchesne (2004a) proposed a semiparametric model with piecewise-constant cause-specific hazard functions that presents robust properties and can be used in most situations in which some second-stage data are available. The same approach can be used to integrate prior knowledge about the failure process using a Bayesian analysis in which the data augmentation algorithm is used for computation (Craiu and Duchesne 2004b).

The model with piecewise-constant cause-specific hazard functions achieves a good balance between flexibility and accuracy on one hand and computational feasibility on the other hand. In addition, although essentially nonparametric, these models allow for likelihood-based methods for estimation and testing (Craiu and Duchesne 2004a; Lawless 2003; He and Lawless 2003). Generally, the endpoints of the intervals that define the piecewise-constant hazard functions (henceforth called simply "the intervals") are chosen by each researcher based on past experience or intuition regarding the failure process.

The main contribution of this article is the exploration of more objective and data-specific criteria for automatically choosing the intervals. To the best of our knowledge, this is the first time such a study has been conducted. In particular, we focus on three widely used model selection criteria: the minimum description length (MDL) principle, the Bayesian information criterion (BIC), and a small-sample version of the Akaike information criterion (AICC). None of the three criteria is uniformly optimal. The AICC and MDL seem to perform better if the number of intervals (relative to the number of observations) is large, whereas the BIC is slightly better than the MDL in situations in which only a small number of intervals is needed, with AICC lagging behind in this case. However, for situations in which the statistician does not have any knowledge regarding the number of intervals, we recommend the MDL criterion because it performs the best on average.

In the next section we describe the data and the likelihood methods used for estimation. We also provide some theoretical justification of the importance of correctly selecting the endpoints for each interval. In Section 3 we discuss in detail the three criteria used for model selection and show how they can be applied to the competing-risks problem. We provide real examples and a simulation study in Section 4 to illustrate the efficiency of the model selection procedure. We close with conclusions and ideas for further improvements.

2. DATA AND MODELS

2.1 Competing Risks Without Masking

In the competing-risks model with unmasked data, assume that there are J possible failure causes and that N items are observed between time $t_0 = 0$ and $t_{\max}$, the time when the study is stopped. Each item i that has failed at a time $t_i \in [t_0, t_{\max}]$ corresponds to a pair (c_i, t_i) in which c_i is equal to the cause of failure; that is, $c_i = j$ if the jth cause is responsible for the failure. In general we refer to (c_i, t_i) as one realization of a bivariate random variable (C, T). For those items that have not failed during the observation period $[t_0, t_{\max}]$ (i.e., for those items that are right-censored), neither c_i nor t_i is observed. In our model selection procedure we use only the uncensored items, because the censored items do not contain any information regarding the cutpoints of the intervals. Practically, this implies that we ignore in the likelihood the terms involving censored items. It should be noted that such assumptions should not be carried over in the estimation phase of the study once the model is selected.

The dependence between T and C is usually specified using the cause-specific hazard rates,

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{\Pr(t < T \le t + h, C = j | T \ge t)}{h}, \qquad j = 1, \ldots, J. \tag{1}$$

From (1), it follows that the marginal hazard function for T is $\lambda(t) = \sum_{j=1}^{J} \lambda_j(t)$ and the marginal survivor function for T is $S(t) = \Pr(T > t) = \exp\{-\int_0^t \sum_{j=1}^{J} \lambda_j(u)\, du\}$. The cumulative incidence functions are $F_j(t) = \Pr(T \le t, C = j) = \int_0^t \lambda_j(t) S(t)\, dt$.

We define each cause-specific hazard rate to be a piecewise constant function; that is, we partition the interval $[0, t_{\max}]$ into K disjoint intervals $(a_{k-1}, a_k]$ so that

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

where $0 = a_0 < a_1 < \cdots < a_K = t_{\max}$ and $\mathbb{1}_k(t)$ is the indicator that $t \in (a_{k-1}, a_k]$. Note that in (2) we have made the implicit assumption that the cutpoints $a_0, \ldots, a_K$ are the same for all failure causes. This assumption is not necessary to carry out the model selection procedure presented here. However, using common cutpoints simplifies the notation and the understanding of the ideas. In addition, the model with equal intervals across causes encompasses the proportional hazards model with piecewise hazards in which $\lambda_j(t) = \sum_{k=1}^{K} r_{jk} \lambda(t) \mathbb{1}_k(t)$ and $\sum_j r_{jk} = 1$. For the remaining of the article we assume the model defined by (2).

The likelihood function is then proportional to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\left\{ \sum_{k=1}^{K} \lambda_{jk} \mathbb{1}_k(t_i) \right\}^{\delta_{ij}} \times \exp\left\{ - \int_0^{t_i} \sum_{k=1}^{K} \lambda_{jk} \mathbb{1}_k(u)\, du \right\} \right], \tag{3}$$

where $\boldsymbol{\theta}$ is the $(J \times K)$-dimensional vector of parameters $(\lambda_{11}, \ldots, \lambda_{JK})$. It is convenient to introduce, for each item i and for each cause j, the indicator δ_{ij}, which is equal to 1 if item i has failed due to cause j and equal to 0 otherwise. The maximum likelihood estimate for $\boldsymbol{\theta}$ can be obtained from (3) as

$$\hat{\lambda}_{jk} = \frac{\sum_{i=1}^{N} \delta_{ij} \mathbb{1}_k(t_i)}{e_k}, \tag{4}$$

where e_k is the exposure in the interval $(a_{k-1}, a_k]$, that is, the sum of all the time lived by each item in this interval. For instance, if item 1 has failure time $t_1 \in (a_{k-1}, a_k]$ and item 2 has failure time $t_2 > a_k$, then their contributions to e_k are $t_1 - a_{k-1}$ and $a_k - a_{k-1}$.

It is clear from (4) that the choice of the cutpoints is crucial for the estimation of each parameter λ_{jk}. To better understand the impact of misspecification of the a_k's on the estimates, consider the following simple example in which the data are

generated under a model, M_0, with two competing risks each having constant cause-specific hazards, λ_1 and λ_2, on the interval $[0, t_{\max}]$. However, suppose that we fail to choose this model and instead work with a model M_1 in which the cutpoints are $0 = a_0 < a_1 < a_2 = t_{\max}$, so that for each cause $j = 1, 2$, the cause-specific hazard is

$$\lambda_j(t) = \lambda_{j1}\mathbb{1}_{(0,a_1]}(t) + \lambda_{j2}\mathbb{1}_{(a_1,t_{\max}]}(t). \quad (5)$$

Denote by n_{jk} the number of items that died of cause j in the interval $(a_{k-1}, a_k]$ for each $k = 1, 2$.

Under the true model, M_0, $\hat{\lambda}_1 = \frac{n_{11}+n_{12}}{e_1+e_2}$, and under model M_1, $\hat{\lambda}_{11} = \frac{n_{11}}{e_1}$. We prove the following result in the Appendix.

Lemma 1. As $N \to \infty$, the following hold:

a. $\hat{\lambda}_{11}$ and $\hat{\lambda}_1$ converge almost surely to λ_1.
b. The variance of $\hat{\lambda}_{11}$ is larger than the variance of $\hat{\lambda}_1$.

This holds even for moderate values of N. For example, with a sample size $N = 50$, simulations show that the variance of $\hat{\lambda}_1$ is 25% smaller than the variance of $\hat{\lambda}_{11}$ when $a_1 = t_{\max}/2$. The situation discussed earlier describes a type of error that results only in variance inflation. But if the original "true model" has piecewise cause-specific hazards with more than one interval, then it is likely that one of the true endpoints will be included inside one of the assumed (misspecified) intervals. In such a case, calculations similar to the foregoing show that the estimates are asymptotically biased and less efficient than the estimates obtained under the true model. We emphasize that the asymptotic results are obtained under the assumption that the size of the sample increases but the true model as well as the specified model intervals remain constant. We must note that in practice it is usually the case that the piecewise-constant hazards are just an approximation to the true ones. However, Lemma 1 shows that it is possible to increase the efficiency of the estimators for the flat segments of the true hazards if the interval endpoints are properly selected. Simulations in Section 4.1 indeed reflect the result of the lemma.

2.2 Competing Risks With Masking

The simple competing-risks model presented earlier rapidly becomes more complicated once some of the items have unknown failure causes. In particular, here we consider the case when one can narrow down the cause of failure to a group of possible causes—in other words, the item's failure is *group-masked*. In addition, we assume that some of the items with a masked failure cause are sent to a second-stage analysis to determine the exact reason for failure.

Therefore, in the case of masked data, for each item i, there are three possible occurrences: (1) i fails because of cause j_i at time t_i; (2) 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 (3) i had still not failed by time t_i. Therefore, some of the items will have a masking group instead of a failure cause, and all of the items will have a failure time. If G is the number of proper groups (i.e., groups that contain more than one element), then the observation for item i is $(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_{G+J}}, \delta_{i1}, \ldots, \delta_{iJ})$, 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. Obviously, if the item is masked in the initial stage and is not sent for further analysis, then the indicators δ_{ij} are not known, and we denote by $\mathcal{M}$ the set of all such items.

As a result of masking, in addition to the parameters λ_{jk}, one must consider the *masking probabilities*,

$$P_{g|j} = \Pr(\text{cause masked to group } g \text{ at stage } 1 | C = j), \quad j \in g. \quad (6)$$

Of eventual interest to practitioners are the diagnostic probabilities (Flehinger et al. 1998, 2002),

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

Using Bayes's rule, we obtain

$$\pi_{j|g}(t) = \frac{\lambda_j(t)P_{g|j}}{\sum_{l\in g}\lambda_l(t)P_{g|l}}. \quad (7)$$

For such data, Craiu and Duchesne (2004a) developed an EM algorithm (Dempster, Laird, and Rubin 1977) in which the δ_{ij}'s (for those masked items) are treated as missing data and that allows estimation of all of the parameters of the model. In the model selection procedures used in the following sections, essential ingredients are the observed likelihood function as well as the estimators for each parameter of interest. We briefly review here the algorithm used for estimation and refer the reader to the article by Craiu and Duchesne (2004a) for details on additional properties, such as convergence and variance estimation.

Using (1)–(6), we obtain the log-likelihood function under the complete data as

$$l_C(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\left\{\left[\delta_{ij}\ln\sum_{k=1}^{K}\lambda_{jk}\mathbb{1}_k(t_i) - \sum_{k=1}^{K}\lambda_{jk}\int_0^{t_i}\mathbb{1}_k(u)\,du\right] + \delta_{ij}\left[\left(1 - \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\right)\ln\left(1 - \sum_{g\in\mathcal{G}_j^*}P_{g|j}\right) + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln P_{g|j}\right]\right\}, \quad (8)$$

where in this case $\boldsymbol{\theta}$ is the vector of parameters λ_{jk} and $P_{g|j}$ for all $1 \le j \le J$, $1 \le k \le K$, and all masking groups g, and $\mathcal{G}_j^*$ is the number of proper masking groups that contain cause j, for all $1 \le j \le J$.

For right-censored observations, the term on the second line of (8) vanishes, and hence the γ_{ig}'s are not needed for right-censored observations. We emphasize again that for the stated purpose of this article (i.e., the choice of the intervals' limits), we consider that there are no right-censored observations. The EM algorithm consists of the following steps:

Initial step. Set $\hat{\lambda}_{jk}^{(0)} = \sum_{i=1}^{N}\mathbb{1}[\delta_{ij}$ observed and equal to $1]/e_k$, $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $\hat{P}_{g|j}^{(0)} = 1/\#\mathcal{G}_j$, $j = 1, \ldots, J$, $g = g_1, \ldots, g_{G+J}$, where # denotes cardinality and $\mathcal{G}_j$ is the set of all masking groups that contain cause j.

E-step. Using (7), compute $E_{\hat{\boldsymbol{\theta}}^{(l-1)}}[\delta_{ij}|OBS]$ as

$$E_{\boldsymbol{\theta}}[\delta_{ij}|OBS] = \begin{cases} 1, & \text{cause of failure of } i \text{ known to be } j \\ 0, & \text{cause of failure of } i \text{ known not to be } j \\ \hat{\pi}_{j|g_i}(t_i), & \text{cause of } i \text{ masked in } g_i \text{ and} \\ & \text{no stage 2 data for } i. \end{cases} \tag{9}$$

The $\hat{\pi}_{j|g_i}(t_i)$ is computed using (7).

M-step. Set

$$\hat{\lambda}_{jk}^{(l)} = \frac{\sum_{i=1}^{N} E_{\hat{\boldsymbol{\theta}}^{(l-1)}}[\delta_{ij}|OBS]\mathbb{1}_k(t_i)}{e_k} \quad \text{and}$$
$$\hat{P}_{g|j}^{(l)} = \frac{\sum_{i=1}^{N} E_{\hat{\boldsymbol{\theta}}^{(l-1)}}[\delta_{ij}|OBS]\gamma_{ig}}{\sum_{i=1}^{N} E_{\hat{\boldsymbol{\theta}}^{(l-1)}}[\delta_{ij}|OBS]}. \tag{10}$$

Note that the expression for $\hat{\lambda}_{jk}^{(l)}$ in (10) is the same as the one in (4) with the exception of those δ_{ij}'s that are unknown and must be replaced by their estimates computed in the E-step. Also note that the observed log-likelihood can also be computed in this situation, because

$$\begin{aligned} l_{\mathrm{OBS}}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}'}[l_C(\boldsymbol{\theta})|OBS] - E_{\boldsymbol{\theta}'}[l_{\mathcal{M}}(\boldsymbol{\theta})|OBS] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg\{\Bigg[E_{\boldsymbol{\theta}'}[\delta_{ij}|OBS]\ln\sum_{k=1}^{K}\lambda_{jk}\mathbb{1}_k(t_i) \\ &\qquad - \sum_{k=1}^{K}\lambda_{jk}\int_0^{t_i}\mathbb{1}_k(u)\,du\Bigg] \\ &\qquad + E_{\boldsymbol{\theta}'}[\delta_{ij}|OBS] \\ &\qquad \times\Bigg[\Bigg(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Bigg)\ln\Bigg(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}\Bigg) \\ &\qquad + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln P_{g|j}\Bigg]\Bigg\} \\ &\quad - \sum_{i\in\mathcal{M}}\sum_{j\in g_i}E_{\boldsymbol{\theta}'}[\delta_{ij}|OBS]\ln\pi_{j|g_i}(t_i). \end{aligned} \tag{11}$$

3. MODEL SELECTION METHODS

In this section we consider the problem of selecting a "best"-fitting model, that is, the problem of choosing a "best" number of intervals K and a "best" combination of the interval endpoints a_k. Recall that in our convention, $a_0 = 0$ and $a_K = t_{\max}$, and hence there are $K-1$ endpoints to be determined. For simplicity, we write $\mathbf{A}_K = (a_1, \ldots, a_{K-1})$. Note that once $\mathbf{A}_K$ is specified, unique maximum likelihood estimates for $\boldsymbol{\theta} = (\lambda_{11}, \ldots, \lambda_{JK})$ can be obtained using (4) or the EM algorithm described in Section 2.2.

We suggest using the following strategy to solve the problem of finding a "best" $\mathbf{A}_K$. First, a model selection principle is applied to define a "best"-fitting model. Then a fast-splitting algorithm is adopted to practically obtain such a defined "best"-fitting model. In the next two sections, we discuss the use of three different model selection principles for defining a "best"-fitting model.

3.1 Akaike and Bayesian Information Criteria

With the Akaike information criterion (AIC), the best-fitting model is defined as the minimizer of an estimator of the Kullback–Leibler (KL) distance measure between a fitted model and the "true" model (see, e.g., Burnham and Anderson 2002). If r is the number of parameters that need to be estimated in a fitted model, then, under some mild regularity conditions, it can be shown that such a KL distance estimator is $-2\times$ "maximized log likelihood" $+\,2r$. Here, for a candidate model with K intervals, there are JK λ_{jk}'s to be estimated and $M = \sum_{h=1}^{G}\#g_h - J$ masking probabilities, where $\#g_h$ denotes the cardinality of the masking group g_h. The number of independent parameters is thus $r = JK + M$, and the AIC best-fitting model is defined as the one that minimizes

$$\mathrm{AIC}(\mathbf{A}_K) = -2l_{\mathrm{OBS}}(\boldsymbol{\theta}) + 2(JK + M). \tag{12}$$

It has been known that this criterion is biased when the sample size is small, and for many problems, biased-corrected versions of it have been proposed (e.g., Burnham and Anderson 2002; McQuarrie and Tsai 1998). Such small-sample version criteria are often termed AICC, and for the present problem it is given by

$$\mathrm{AICC}(\mathbf{A}_K) = -2l_{\mathrm{OBS}}(\boldsymbol{\theta}) + 2(JK+M) + \frac{2(JK+M)(JK+M+1)}{N-JK-M-1}. \tag{13}$$

Our simulation study suggests that AICC is uniformly better than AIC, and hence AIC is not included in our summary of findings.

The form of the BIC (Schwarz 1978) is very similar to AIC. Instead of a constant value 2, it replaces the penalty for each parameter with $\log N$. Thus the BIC best-fitting model is defined as the one that minimizes

$$\mathrm{BIC}(\mathbf{A}_K) = -2l_{\mathrm{OBS}}(\boldsymbol{\theta}) + (M + JK)\log N. \tag{14}$$

As stated by Hastie, Tibshirani, and Friedman (2002), choosing the model with the minimum BIC value is approximately equivalent to choosing the model with the largest posterior probability with respect to an uniform prior. Because the penalty term for BIC is larger than the one for AIC, it is expected that BIC tends to produce more parsimonious best-fitting models than AIC. However, this comparison is less clear with respect to the AICC, especially if the sample size N is not very large. An astute reader will have noticed that in the case of masked data, the term involving the number of masking probabilities, M, can be omitted in (12) and (14), because this term remains the same no matter how many intervals we use. However, this is not the case for AICC, as can be seen from (13), so the number M must be taken into consideration in that case.

3.2 Minimum Description Length Principle

The MDL principle uses ideas from the information theory and signal processing literature and was adapted by Rissanen (1989) as a model selection tool for statisticians. It defines the best-fitting model as the one that produces the shortest code length of the data. Loosely speaking, the code length of an object can be treated as the amount of memory space required to

store the object (for details, see Rissanen 1989). (Also see, e.g., Hansen and Yu 2001 and Lee 2001 for introductory tutorials to the MDL principle.)

One common approach to applying the MDL principle is to split the code length for a set of data into two components: a fitted model plus the data "conditioned on" the fitted model, that is, the part in the data not explained by the fitted model. For the present problem, a fitted model can be specified by $\mathbf{A}_K$ and the maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{\lambda}_{11}, \ldots, \hat{\lambda}_{JK})$ for $\boldsymbol{\theta}$. We choose to omit the number of masking parameters, M, from the criterion because this number remains the same across models with different intervals. If $CL(z)$ denotes the code length of the object z, then we have the following decomposition:

$$\begin{aligned} CL(\text{“data”}) &= CL(\mathbf{A}_K, \hat{\boldsymbol{\theta}}) + CL(\text{“data”}|\mathbf{A}_K, \hat{\boldsymbol{\theta}}) \\ &= CL(\mathbf{A}_K) + CL(\hat{\boldsymbol{\theta}}) + CL(\text{“data”}|\mathbf{A}_K, \hat{\boldsymbol{\theta}}). \end{aligned}$$

Now the task is to find an expression for CL(“data”) so that the MDL best-fitting model can be defined and obtained as its minimizer. We show in Appendix B that in the case of competing-risks data without masking CL(“data”) can be well approximated by

$$\text{MDL}(\mathbf{A}_K) = \sum_{k=1}^{K} \log n_k + \frac{J}{2} \sum_{k=1}^{K} \log(n_k + n_{k+1} + \cdots + n_K) - l_{\text{OBS}}(\boldsymbol{\theta}), \quad (15)$$

where n_k is the number of observations inside the interval $(a_{k-1}, a_k]$. We propose to select the minimizer of $\text{MDL}(\mathbf{A}_K)$ as our MDL-based estimate. Note that, unlike in AICC or BIC, in MDL the penalty for each interval is not the same. First, the penalty for the kth interval is a function of its width n_k. Second, from the double summation in the second term of $\text{MDL}(\mathbf{A}_K)$, one can see that those “late” intervals (i.e., large k) are penalized more than those “early” intervals (i.e., small k). This agrees with the intuition that stronger penalties (or, loosely, more prior information) are required for those “late” intervals, because as time passes, more and more items die, and hence a smaller amount of information is available for those intervals.

Ideally, one would like to adapt the MDL principle to the situation of masked data. However, due to a technical difficulty given in Appendix B, we decide to use the same criterion for unmasked data. Simulations show that this choice performs better, on average, than AICC and BIC.

3.3 A Fast-Splitting Algorithm

Minimizing any one of the foregoing selection criteria with respect to A_K is not a trivial task, because the search space is enormous. Here we describe a simple, fast, and yet effective search algorithm for approximating the minimizers of the criteria.

The algorithm starts with fitting a model with only a $K = 1$ interval (i.e., no breakpoints) and calculates the corresponding value of the selection criterion used (i.e., MDL, BIC, or AICC). Denote this value by S_1. Then the algorithm adds one breakpoint to the model or, equivalently, splits the entire domain into two intervals. The location of this first breakpoint is chosen in the following manner. Among all possible breakpoint locations, if the whole domain is split at this particular breakpoint, then it will produce the largest increase (or the smallest decrease) of the likelihood value. To locate such a breakpoint, one could conduct a grid search on $[0, t_{\max}]$ or, as in our implementation, limit the set of all possible breakpoints to be the midpoints between any two adjacent observations. To further speed up the algorithm, one could consider, say, every other midpoint. That is, if the set of all midpoints is $\{x_1, x_2, \ldots, x_{N-1}\}$, then one can consider $\{x_1, x_3, x_5, \ldots, x_{N-1}\}$ instead of all of the x_i's. Once such a breakpoint is located, the algorithm computes the value of the selection criterion being used. This selection criterion value is denoted by S_2.

The next step of the algorithm is to add one additional breakpoint to the existing two-interval model; that is, to produce a model with $K = 3$ intervals. This second breakpoint is chosen in a similar manner as before; among all possible splitting locations, it produces the largest increase of the likelihood value after the splitting. After this breakpoint is chosen, the algorithm computes the selection criterion value, S_3. If this computed selection criterion value (S_3) is larger than the value (S_2) obtained with $K = 2$ intervals, then the algorithm stops and the fitted model that has the smallest selection criterion value among all of the fitted models examined so far is taken as the final fitted model. Otherwise, the algorithm continues to add breakpoints to the model, to recompute, and to compare the selection criterion values $S_4, S_5, \ldots$ in a similar fashion as before. The process stops when the selection criterion value S_i increases, and the fitted model with the smallest criterion value S_i is taken as the final fitted model. Some timing figures on the computational speed of this algorithm are reported in the next section.

It should be noted that for the algorithm defined in Section 2.2, there are certain restrictions on the width of the intervals. More precisely, we need to have for each interval $I_k = (a_{k-1}, a_k]$ and for each cause j at least one item that has failed during I_k and with a failure cause masked in a group that contains j. Such restrictions can be easily incorporated within the fast-splitting algorithm.

4. EXAMPLES

4.1 Simulation Study

We conducted a simulation study to empirically evaluate the performances of the foregoing model selection methods. We used two sets of λ_{jk}'s as our test functions. These two functions have the same number of failure causes, $J = 3$, but different numbers of intervals, 3 and 7. The locations of the interval endpoints and the corresponding values for the λ_{jk}'s are listed in Tables 1 and 2.

Altogether, three sample sizes, $\mathbf{N} = (100, 200, 800)$, and three values for the probability p that a masked item is sent to second-stage analysis, $p \in \{.3, .6, 1.0\}$, were used (e.g., $p = 1$

Table 1. True λ_{jk} Values for the Test Function With Three Intervals

j	λ_{j1}	λ_{j2}	λ_{j3}
1	.0030	.0200	.0120
2	.0045	.0100	.0300
3	.0045	.0100	.0300

NOTE: The two interval endpoints are 30 and 50.

Table 2. True λ_{jk} Values for the Test Function With Seven Intervals

j	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	λ_{j5}	λ_{j6}	λ_{j7}
1	.0013	.0052	.0151	.0001	.0200	.0050	.0500
2	.0015	.0081	.0151	.0021	.0300	.0050	.0600
3	.0012	.0071	.0161	.0017	.0180	.0060	.0500

NOTE: The six interval endpoints are 53, 65, 75, 89, 101, and 173.

means that there are no missing data in the sample). Thus the total number of different experimental configurations was $2 \times 3 \times 3 = 18$.

For each experimental configuration, 400 simulated datasets were generated, and the following methods were applied to each dataset to obtain a fitted model:

- *mdl*: the MDL criterion (15) minimized by the splitting algorithm described in Section 3.3,
- *aicc*: similar to *mdl* but for the AICC criterion (13)
- *bic*: similar to *mdl* but for the BIC criterion (14)
- $f5$: a model with five equilength intervals in $[0, t_{\max}]$. This approach of fixing five intervals is a generic approach for situations in which the researcher does not have additional experience with the type of failure data under study and is included in this simulation as a baseline for comparison.

A discrete approximation of the mean squared error (MSE) was used to measure the quality of the fitted models,

$$\mathrm{MSE} = \sum_{j=1}^{J} \int_0^{t_{\max}} \{\lambda_j(t) - \hat{\lambda}_j(t)\}^2 \, dt,$$

where the true and known λ_j is $\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} \mathbb{1}_k(t)$ and the estimate $\hat{\lambda}_j$ is $\hat{\lambda}_j(t) = \sum_{k=1}^{\hat{K}} \hat{\lambda}_{jk} \mathbb{1}_k(t)$. Boxplots of the logs of these MSE values are given in Figures 1 and 2. Paired Wilcoxon tests were also applied to test whether the difference between the median MSE values of any two methods was significant. The significance level used was 1.25%. If the median MSE value of a particular method was significantly less than the median MSE values of the other three methods, then this method was assigned rank 1. If its median MSE value was significantly less than two but greater than one of the other three, then the method was assigned rank 2, and so on for ranks 3 and 4. Methods with insignificantly different MSE values share the same averaged rank. These paired Wilcoxon rankings are also shown in Figures 1 and 2.

Note from Figures 1 and 2 that none of the three criteria is uniformly optimal. Table 3 and also our numerical experience suggest that *aicc* has a tendency to overestimate the number of intervals. When the number of intervals is smaller (i.e., three) *bic* performs the best, with *mdl* lagging not far behind. But

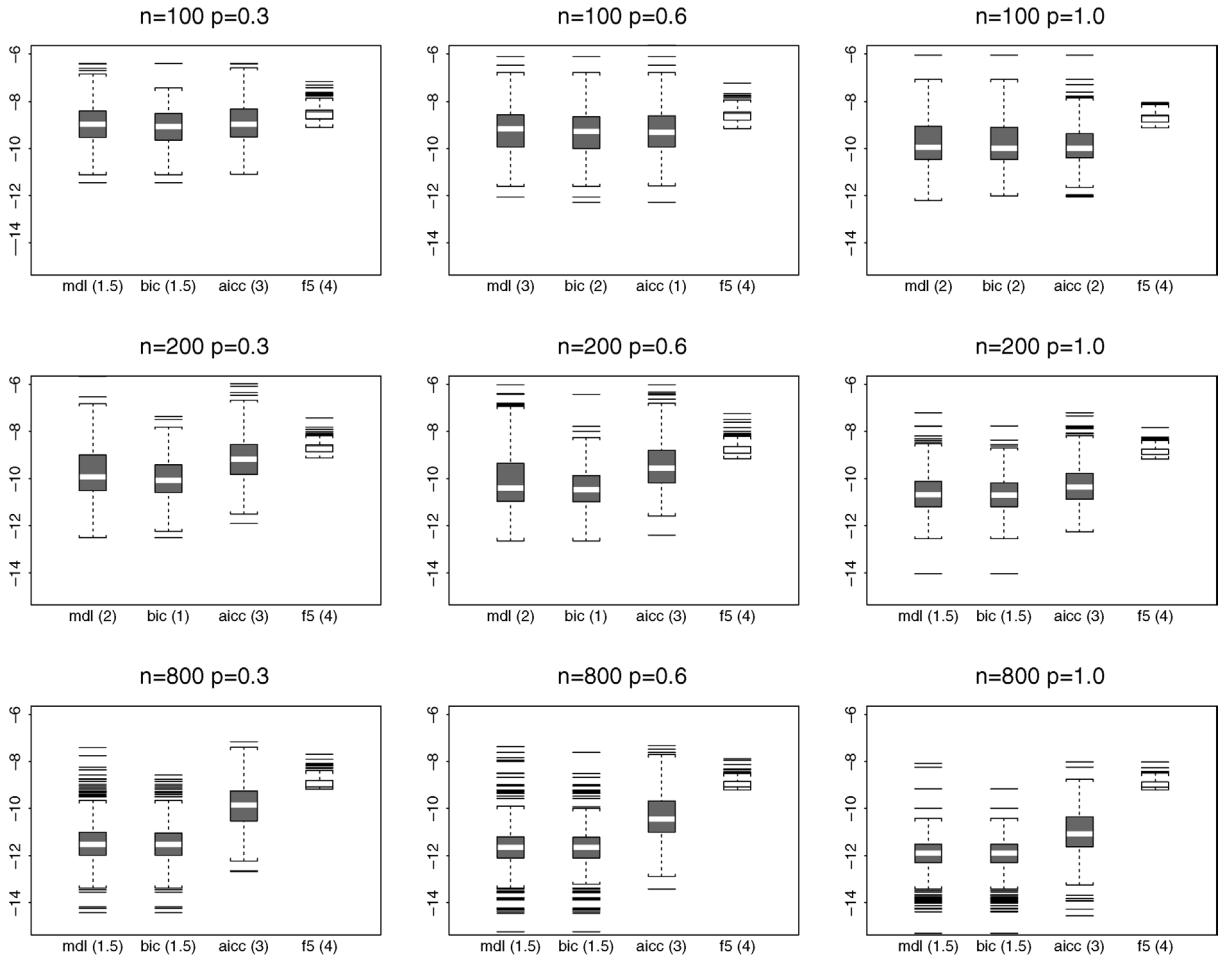

Figure 1. Boxplots of the Log of the MSE Values for the Test Function With Three Intervals. The paired Wilcoxon rankings are listed inside the parentheses.

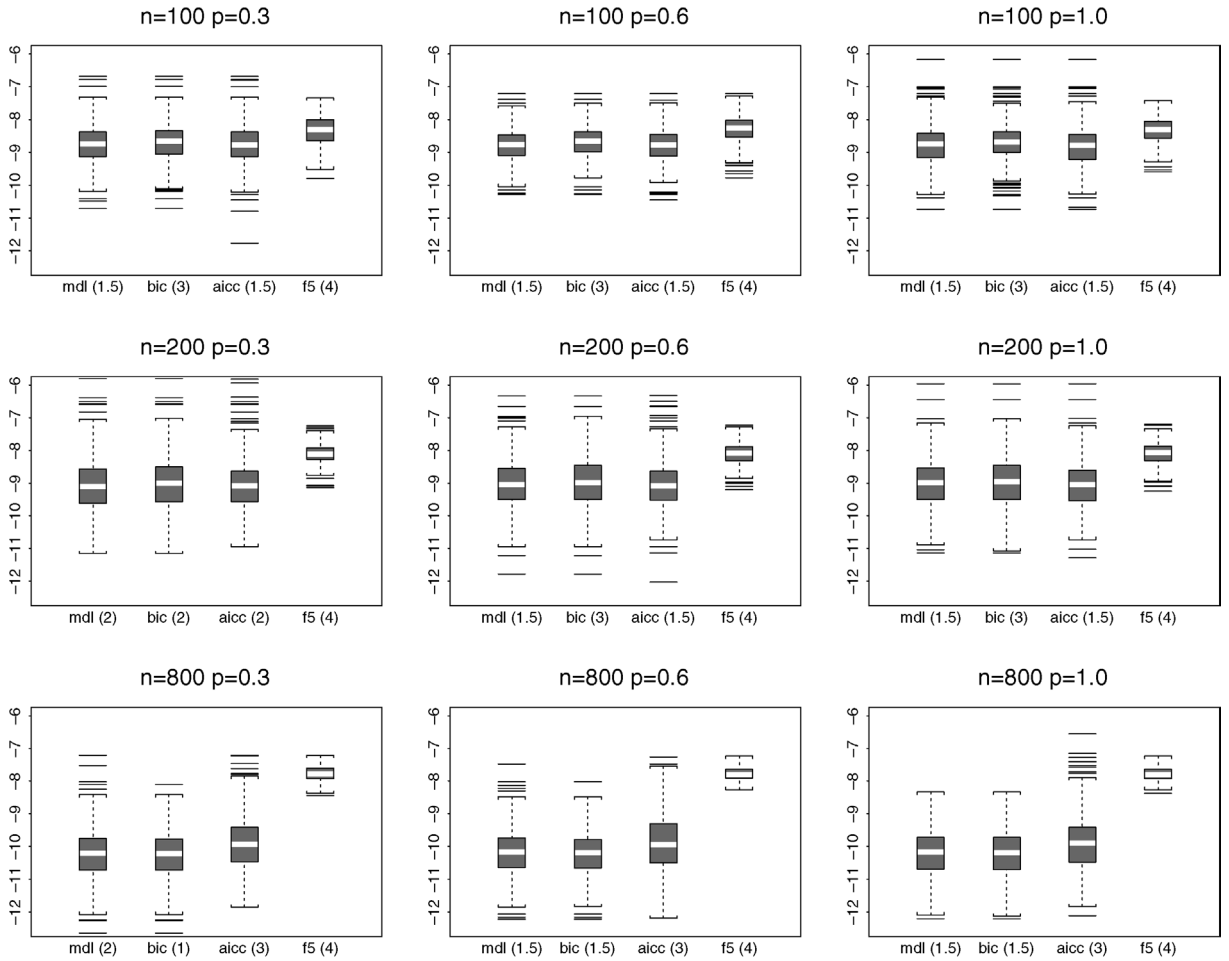

Figure 2. Boxplots of the Log of the MSE Values for the Test Function With Seven Intervals. The paired Wilcoxon rankings are listed inside the parentheses.

when the number of intervals is larger (i.e., seven), both *mdl* and *aicc* do better than *bic*, with *mdl* the best.

Surprisingly, $f5$ does not seem to take advantage of additional data. One can see that as the amount of complete data increases from top to bottom and from left to right, *mdl*, *bic*, and *aicc* become sensibly more accurate.

The averaged Wilcoxon rankings for *mdl*, *bic*, *aicc*, and $f5$ are 1.72, 1.92, 2.37, and 4. Judging from this measure, it seems that *mdl* should be the preferred method for a researcher without any prior knowledge on the expected number of intervals required by the particular application.

Table 3. Further Results for the Test Function With Three Intervals and $N = 200$

		Number of intervals				Endpoint 1	Endpoint 2
p	*Method*	*2*	***3***	*4*	*5+*	*(at $t = 30$)*	*(at $t = 50$)*
.3	*mdl*	54	**285**	54	7	30.06 (.017)	50.90 (.296)
	bic	39	**316**	34	11	30.06 (.016)	50.40 (.238)
	aicc	0	**24**	66	310	30.00 (.050)	50.04 (.432)
.6	*mdl*	31	**323**	37	6	30.03 (.016)	50.19 (.241)
	bic	20	**346**	28	3	30.04 (.015)	50.09 (.219)
	aicc	0	**25**	81	291	30.08 (.040)	50.24 (.442)
1.0	*mdl*	12	**376**	12	0	30.02 (.015)	50.18 (.155)
	bic	7	**386**	7	0	30.03 (.015)	50.14 (.157)
	aicc	0	**117**	166	117	30.05 (.027)	50.62 (.271)

NOTE: The left half of the table lists the number of times out of 400 repetitions that the number of intervals a particular method selected. The right half of the table provides the averaged locations of the interval endpoints for those repetitions that the correct number of intervals were identified. Numbers in the parentheses are estimated standard errors. Numbers in bold represent the number of correct choices for each method.

To assess the performance of the methods in terms of selecting the correct number of intervals and the correct locations of the endpoints, we recorded, for those experimental settings associated with $N = 200$ and the test function with three intervals, the number and the endpoints of the intervals that *mdl*, *bic*, and *aicc* selected. Results are summarized in Table 3. The methods *mdl* and *bic* seem to be preferable in this case.

To visually evaluate the quality of the estimates, Figure 3 plots the true λ_{jk}'s for the three-interval test function, together with one representative estimate sampled from the 400 simulations. The number of observations was $N = 400$, and the masking probability was $p = .6$. From this figure, one can see that both *mdl* and *bic* selected the correct number of intervals, whereas *aicc* overfitted the data. For $f5$, one can see that it inflated the variance of the estimates confirming the result of Lemma 1.

We also applied the foregoing methods to data generated from hazards following a Weibull distribution. In particular, we generated datasets of size $N = 400$ with a probability of a second-stage analysis of $p = .6$. Figure 4 displays the true hazards and the estimate obtained from a typical dataset. Not surprisingly, the performances of the three procedures depend on the gradient of the true hazards. However, it seems fair to say

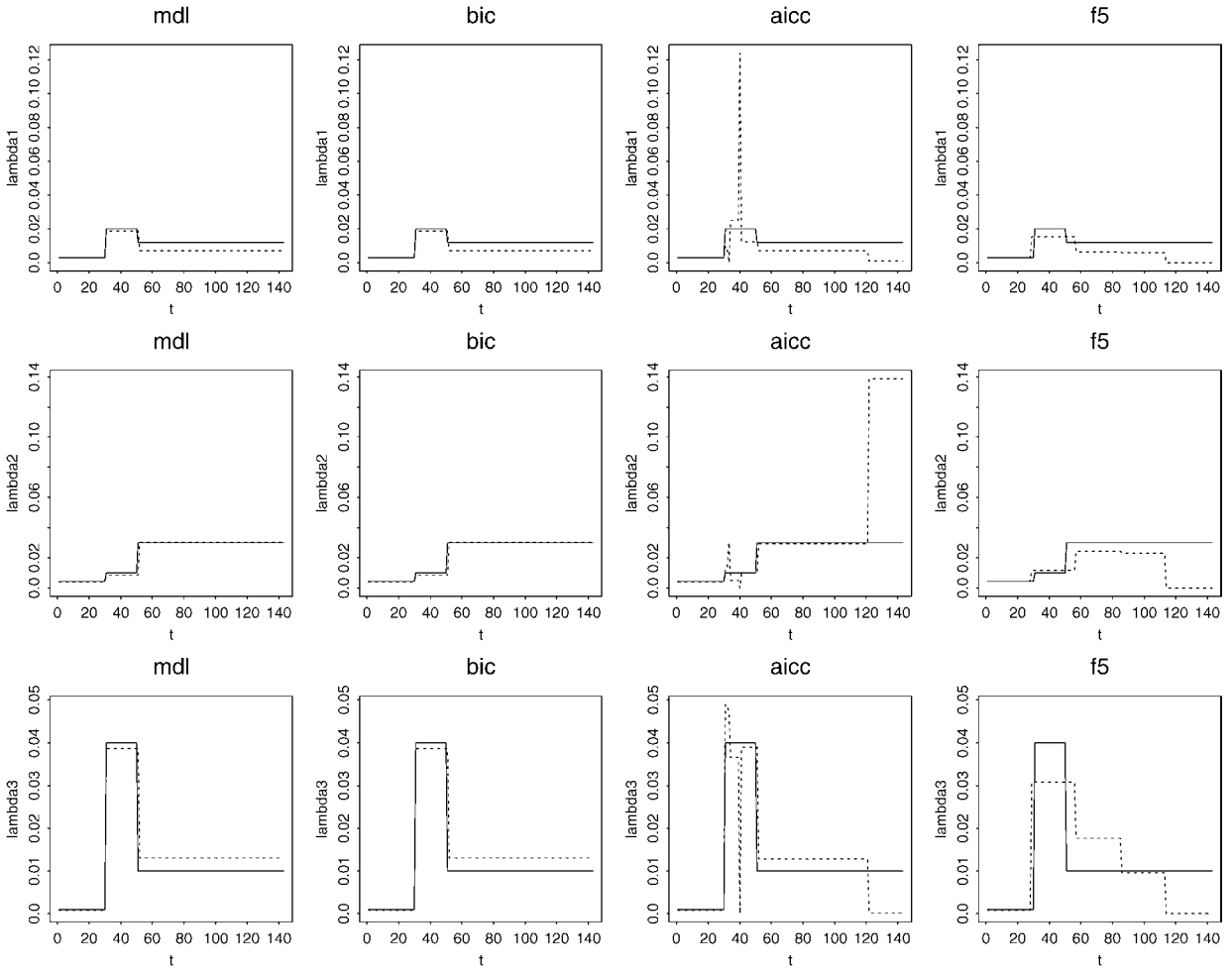

Figure 3. Plots of True (solid lines) and Estimated (dotted lines) λ_{jk}'s for the Three-Interval Test Functions.

that the *mdl* and *bic* perform well for the three hazards shown. The middle row has a poor fit on the last piece, because most of the items die early and few data points are available later in the study. This obviously affects *aicc* and $f5$ even more, because many of the intervals artificially added in the model will contain very few data points.

Finally, we report some timing figures. If the true number of intervals is 7 and $N = 200$, then our implementation took an average of 5 seconds on a Sun Ultra 60 Unix workstation for any of the three model selection methods to finish. If $N = 800$, then on average it took 17 seconds on the same machine.

4.2 Hard Drive Data

We consider here a real dataset analyzed by Flehinger et al. (2002) using a model with Weibull cause-specific hazards. Craiu and Duchesne (2004a) analyzed the same data using piecewise-constant hazards, and their conclusions were very close to those obtained by Flehinger et al. (2002). However, selection of the intervals was based on a subjective choice, a situation that we want to remedy here.

We are interested in the failure causes of a certain subassembly of hard disk drives. Some of these causes are related to particular components (e.g., defective head), but others, such as particle contamination, are not. The analysis does not discriminate among these and simply treats them as causes of failure.

The data consist of 10,000 hard drives, of which 172 have failed during an observation period of 4 years. There are three possible failure causes, and for many of the failed items the true cause of death is group-masked. There are two masking groups, $\{1, 3\}$ and $\{1, 2, 3\}$. The results obtained using *mdl* and *bic* are similar to those obtained by Craiu and Duchesne (2004a), where the endpoints of the intervals were chosen subjectively, as can be seen from Table 4. The *aic* suggests eight intervals, whereas the *aicc* suggests six intervals. Both models result in inflated variances for the maximum likelihood estimators. The asymptotic standard errors, as measured using the SEM algorithm (Meng and Rubin 1991), are smaller with the new cutpoints $(0, .81, 1.58, 3.77, 4)$ compared with the ones obtained previously using the cutpoints $(0, 1, 2, 3, 4)$.

5. CONCLUSIONS AND FURTHER WORK

Choosing the endpoints of the piecewise cause-specific hazard intervals can play an important part in solving a competing-risks problem with or without group masking. Here we discuss three possible approaches using the MDL, AICC, and BIC criteria. The MDL and BIC seem to be more robust with respect to the number of intervals required for a good approximation. We recommend using MDL in situations in which little is known about the number of intervals required.

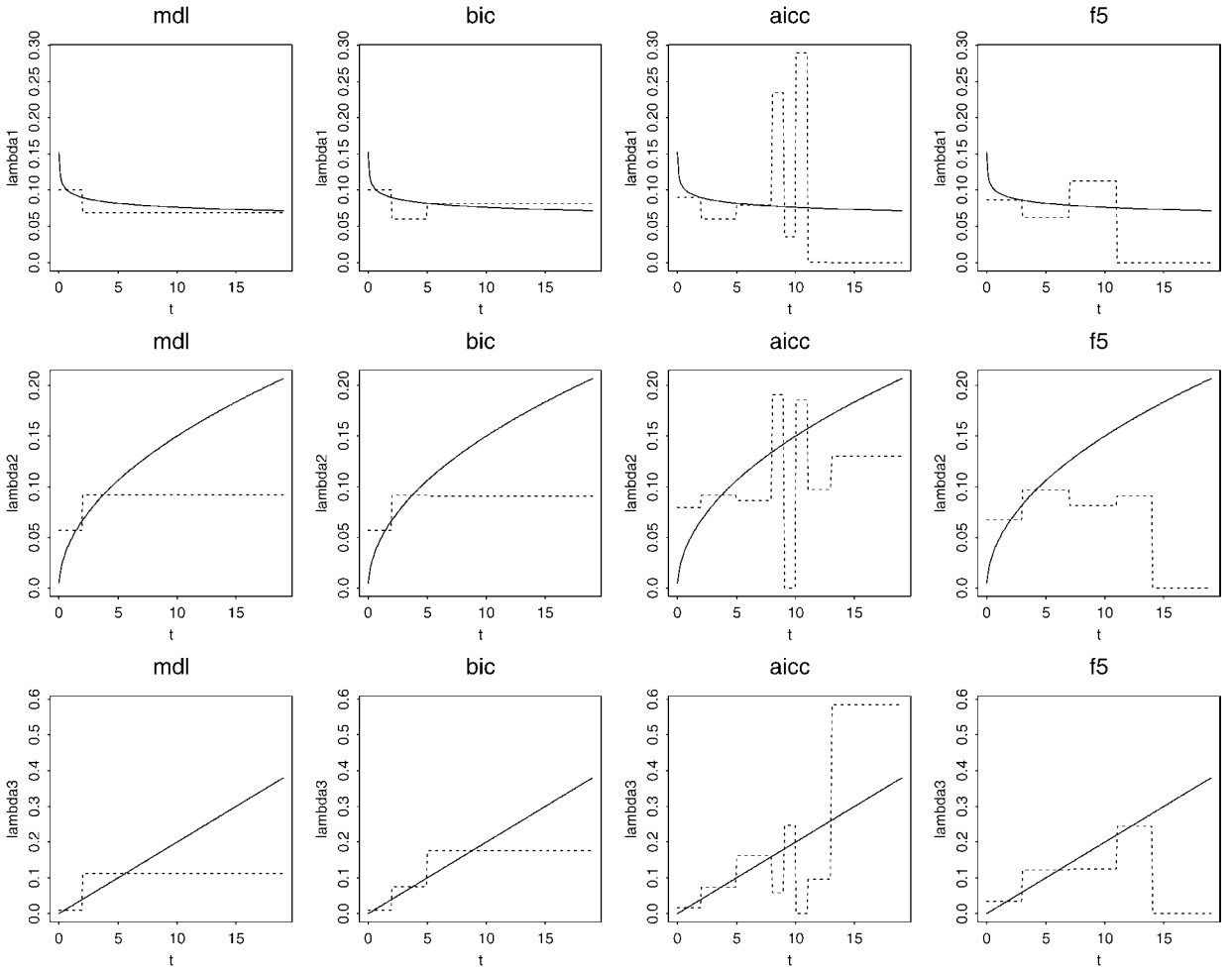

Figure 4. Plots of True (solid lines) and Estimated (dotted lines) λ_{jk}'s for Weibull Hazards.

One can adapt the present work to the case where not all failure causes share the same number of intervals and interval endpoint locations. It is straightforward to modify the AICC and BIC criteria for this generalization, because the penalty terms in these two criteria are proportional to the number of parameters in the model being fitted. It is also straightforward to derive a corresponding MDL criterion. The necessary modification is to derive new expressions for $CL(\mathbf{A}_K)$ and $CL(\hat{\boldsymbol{\theta}})$, and the material in Appendix B can be applied to derive such expressions. The fast-splitting algorithm discussed in Section 3.3 can also be modified to minimize any of these new criteria. The main idea is that at each time step, instead of adding a new same breakpoint to all failure causes, the new algorithm adds one breakpoint to only one failure cause. However, this would be a lengthy procedure, because at each time step many comparisons are needed to determine a best breakpoint.

In addition, we would like to expand the present study to fitting splines instead of constant functions on each interval. However, this significantly increases the amount of data required and the complexity of the computation, and further research is necessary.

ACKNOWLEDGMENT

The authors thank the editor, an associate editor, and two anonymous referees for constructive remarks and the suggestion to use AICC. The first author gratefully acknowledges the support of a Connaught grant from the University of Toronto and an individual research grant from the Natural Sciences and

Table 4. Masking Probability Estimates (Flehinger et al. 2002 hard drive data)

	Estimates of the P_{glj}'s								
	Flehinger et al.			*Craiu and Duchesne*			*Our estimates*		
Masking group	*j = 1*	*j =2*	*j = 3*	*j = 1*	*j = 2*	*j = 3*	*j = 1*	*j = 2*	*j = 3*
$g = \{1, 3\}$	.412	0	.446	.410 (.0789)	0	.445 (.0563)	.410 (.0680)	0	.443 (.0353)
$g = \{1, 2, 3\}$	.310	.469	.436	.308 (.0766)	.457 (.1190)	.439 (.0565)	.305 (.0658)	.448 (.0988)	.442 (.0356)

NOTE: Numbers in parentheses are asymptotic standard errors computed with SEM.

Engineering Research Council of Canada. The second author gratefully acknowledges the partial support of National Science Foundation grant DMS-02-03901.

APPENDIX A: PROOF OF LEMMA 1

To show part a, write $\hat{\lambda}_{11} = \frac{n_{11}/N}{e_1/N}$ and note that because of the law of large numbers, the sequences $x_n = n_{11}/N$ and $y_n = e_1/N$ converge almost surely to $\Pr_{M_0}(T \in (0, a_1], C = 1)$ and $E_{M_0}[\min(T, a_1)]$. (The index M_0 signifies that the probability and expectation are computed using the distribution under that model.)

From (1), we get

$$\Pr_{M_0}\big(T \in (0, a_1], C = 1\big) = \frac{\lambda_1}{\lambda_1 + \lambda_2}\big(1 - e^{-(\lambda_1+\lambda_2)a_1}\big) \quad \text{(A.1)}$$

and

$$\begin{aligned} &E_{M_0}[\min(T, a_1)] \\ &\quad = E\big[T\mathbb{1}_{\{T \le a_1\}} + a_1\mathbb{1}_{\{T > a_1\}}\big] \\ &\quad = -a_1 e^{(\lambda_1+\lambda_2)a_1} + \int_0^{a_1} e^{-(\lambda_1+\lambda_2)t}\,dt + a_1 e^{-(\lambda_1+\lambda_2)a_1} \\ &\quad = \frac{1 - e^{-(\lambda_1+\lambda_2)a_1}}{\lambda_1 + \lambda_2}. \end{aligned} \quad \text{(A.2)}$$

Under fairly general regularity conditions, (A.1) and (A.2) imply that $\hat{\lambda}_{11}$ converges almost surely to λ_1. Similar calculations can be done to show that the same holds for $\hat{\lambda}_1$.

Taking second derivatives of the log-likelihood obtained from (3), we can deduce that the asymptotic variance of $\hat{\lambda}_{11}$, obtained using the observed Fisher information, is n_{11}/e_1^2 whereas the asymptotic variance of $\hat{\lambda}_1$, is $(n_{11} + n_{12})/(e_1 + e_2)^2$. For N large, using part a, we have $n_{11}/e_1 \approx \lambda_1 \approx (n_{11} + n_{12})/(e_1 + e_2)$, so that the desired result part b follows.

APPENDIX B: DERIVATION OF THE MINIMUM DESCRIPTION LENGTH CRITERION (15)

This appendix outlines the derivation of $\mathrm{MDL}(\mathbf{A}_K)$. Recall that the goal is to find an expression for

$$CL(\text{“data”}) = CL(\mathbf{A}_K) + CL(\hat{\boldsymbol{\theta}}) + CL(\text{“data”}|\mathbf{A}_K, \hat{\boldsymbol{\theta}}),$$

and we begin with $CL(\mathbf{A}_K)$. Because we limit the breakpoints between different intervals to be a subset of the midpoints of any pair of adjacent observations, the width of each interval can be specified with n_k, the number of observations falls within that interval. Thus $\mathbf{A}_K$ is completely specified by all n_k's. Using the fact that the code length for an integer I is $\log_2 I$, we have $CL(\mathbf{A}_K) = \sum_k CL(n_k) = \sum_k \log_2 n_k$. To calculate $CL(\hat{\boldsymbol{\theta}}) = CL(\hat{\lambda}_{11}) + \cdots + CL(\hat{\lambda}_{JK})$, we apply the following result of Rissanen (1989). If a maximum likelihood estimate is calculated from m data points, then its code length is $\frac{1}{2}\log_2 m$. It can be seen that when there is no masking, for all j, $\hat{\lambda}_{jk}$ is computed from $n_k + \cdots + n_K$ data points, that is, those items that are still alive. Thus $CL(\hat{\lambda}_{jk}) = \frac{1}{2}\log_2(n_k + \cdots + n_K)$, and hence $CL(\hat{\boldsymbol{\theta}}) = \frac{J}{2}\sum_k \log_2(n_k + \cdots + n_K)$. Finally, based on Shannon's classical results on information theory, Rissanen (1989) showed that the code length for “data given a fitted model” amounts to the negative of the conditional log (base 2) likelihood of the data given the fitted model. That is, for our problem, $CL(\text{“data”}|\mathbf{A}_K, \hat{\boldsymbol{\theta}}) = -l_{\mathrm{OBS}}(\boldsymbol{\theta})$. Now, combining these expressions and changing $\log_2$ to log, we obtain $\mathrm{MDL}(\mathbf{A}_K)$.

To derive an MDL criterion for masked data, one would need to recalculate $CL(\hat{\lambda}_{jk})$ for all j and k. This calculation requires knowledge of the number of items used in the computation of $CL(\hat{\lambda}_{jk})$. However, the EM algorithm makes it difficult to track the number of items used to estimate each of the λ_{jk}'s. In addition, not all items will have equal weight in (10), because their importance will depend on (9) via (7). We therefore decide to use $\mathrm{MDL}(\mathbf{A}_K)$ for unmasked data.

[Received June 2004. Revised December 2004.]

REFERENCES

Aalen, O. O. (1978), “Nonparametric Inference for a Family of Counting Processes,” *The Annals of Statistics*, 6, 701–726.

Barrett, A. J., Horowitz, M. M., Gale, R. P., Biggs, J. C., Camitta, B. M., Dicke, K. A., Gluckman, E., Good, R. A., Herzig, R. H., and Lee, M. B. (1989), “Marrow Transplantation for Acute Lymphoblastic Leukemia at Memorial Sloan-Kettering Cancer Center,” *Blood*, 74, 862–871.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.

Craiu, R. V., and Duchesne, T. (2004a), “Inference Based on the EM Algorithm for the Competing Risk Model With Masked Causes of Failure,” *Biometrika*, 91, 543–558.

——— (2004b), Using EM and Data Augmentation for the Competing Risks Model, in *Applied Bayesian Modeling and Causal Inference From an Incomplete-Data Perspective*, eds. A. Gelman and X. L. Meng, New York: Wiley, Chap. 22.

Crowder, M. (2001), *Classical Competing Risks*, New York: Chapman & Hall.

Dempster, A., Laird, N., and Rubin, D. (1977), “Maximum Likelihood From Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society*, Ser. B, 39, 1–22.

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.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1998), “Survival With Competing Risks and Masked Causes of Failures,” *Biometrika*, 85, 151–164.

——— (2002), “Parametric Modeling for Survival With Competing Risks and Masked Failure Causes,” *Lifetime Data Analysis*, 8, 177–203.

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D., and Brennan, M. F. (1993), “On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data,” *Journal of the American Statistical Association*, 88, 400–409.

Goetghebeur, E., and Ryan, L. (1990), “A Modified Log-Rank Test for Competing Risks With Missing Failure Types,” *Biometrika*, 77, 151–164.

——— (1995), “Analysis of Competing Risks Survival Data When Some Failure Types Are Missing,” *Biometrika*, 82, 821–833.

Hansen, M. H., and Yu, B. (2001), “Model Selection and the Principle of Minimum Description Length,” *Journal of the American Statistical Association*, 96, 746–774.

Hastie, T., Tibshirani, R., and Friedman, J. (2002), *The Elements of Statistical Learning*, New Yor: Springer-Verlag.

He, W., and Lawless, J. F. (2003), “Flexible Maximum Likelihood Methods for Bivariate Proportional Hazards Models,” *Biometrics*, 59, 837–848.

Hoel, D. G. (1972), “A Representation of Mortality Data by Competing Risks,” *Biometrics*, 28, 475–488.

Holt, J. D. (1978), “Competing Risk Analysis With Special Reference to Matched Pair Experiments,” *Biometrika*, 65, 159–166.

Kalbfleisch, J. D., and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* (2nd ed.), New York: Wiley.

Kodell, R. L., and Chen, J. J. (1987), “Handling Cause of Death in Equivocal Cases Using the EM Algorithm” (with discussion), *Communications in Statistics, Part A—Theory and Methods*, 16, 2565–2585.

Lagakos, S. W. (1977), “A Covariate Model for Patially Censored Data Subject to Competing Causes of Failure,” *Journal of the Royal Statistical Society*, Ser. B, 27, 235–241.

Lapidus, G., Braddock, M., Schwartz, R., Banco, L., and Jacobs, L. (1994), "Accuracy of Fatal Motorcycle Injury Reporting on Death Certificates," *Accident Analysis and Prevention*, 26, 535–542.

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

Lee, T. C. M. (2001), "An Introduction to Coding Theory and the Two-Part Minimum Description Length Principle," *International Statistical Review*, 69, 169–183.

McQuarrie, A. D. R., and Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Moeschberger, M. L., and David, H. A. (1971), "Life Tests Under Competing Causes of Failure and the Theory of Competing Risks," *Biometrics*, 27, 909–933.

Nelson, W. B. (1969), "Hazard Plotting for Incomplete Failure Data," *Journal of Quality Technology*, 1, 27–52.

Prentice, R. L., Kalbfleisch, J. D., Peterson, Jr., A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978), "The Analysis of Failure Times in the Presence of Competing Risks," *Biometrics*, 34, 541–554.

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.

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Sun, Y., and Tiwari, R. C. (1997), "Comparing Cumulative Incidence Functions of a Competing-Risks Model," *IEEE Transactions on Reliability*, 46, 247–253.

Taylor, H. M. (1994), "The Poisson–Weibull Flaw Model for Brittle Fiber Strength," in *Extreme Value Theory and Applications*, eds. J. Galambos, J. Lechner, and E. Simin, Amsterdam: Kluwer, pp. 43–59.