BIOMETRICS 54, 1498–1507
December 1998

A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data Under the Cox Proportional Hazards Model

William B. Goggins,[1] Dianne M. Finkelstein,[1,2,*] David A. Schoenfeld,[2] and Alan M. Zaslavsky[3]

[1]Biostatistics Department, Harvard School of Public Health, Boston, Massachussetts 02115, U.S.A.
[2]Massachusetts General Hospital, Boston, Massachusetts 02114, U.S.A.
[3]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.

SUMMARY

This paper proposes a Monte Carlo EM (MCEM) algorithm for fitting the proportional hazards model for interval-censored failure-time data. The algorithm generates orderings of the failures from their probability distribution under the model. We maximize the average of the log-likelihoods from these completed data sets to obtain updated parameter estimates. As with the standard Cox model, this algorithm does not require the estimation of the baseline hazard function. The performance of the algorithm is evaluated using simulations, and the method is applied to data from AIDS and cancer studies. Our results indicate that our method produced more precise and unbiased estimates than methods of right and midpoint imputation.

1. Introduction

Failure-time data are said to be interval censored when the events of interest are not observed exactly but instead are only known to lie in intervals that may overlap and vary in length. Such data occur when the failure event can only be detected by a screening examination. For example, Kim, De Gruttola, and Lagakos (1993) describe data from an HIV screening study in which stored blood samples were tested retrospectively to determine seropositivity. For patients who had seroconverted, we only know that this occurred between the time of their last negative screening and the time of their first positive screening. Interval-censored data also appear in clinical trials, e.g., when the occurrence of a clinical or laboratory change is monitored at routine medical visits. Finkelstein (1986) describes data from a breast cancer study where patients were monitored for cosmetic changes following radiation therapy. In this study, changes were noted at clinic visits, which occurred at irregular intervals, resulting in a data set consisting of overlapping intervals of failure.

The Cox (1972) proportional hazards model is the most commonly applied methodology for assessing the effect of covariates on the hazard of failure. An appealing feature of the Cox model is that, for the usual case of right-censored and exact failure-time data, the method does not require specification or estimation of the baseline hazard. Although a functional form is assumed for the effect of covariates, estimation relies only on the rank ordering of the exact and censored failure times. For interval-censored data, it is not possible to identify the exact ranking of the failure times, so the methodology that has been developed for exact and right-censored data cannot be directly applied.

Often, interval-censored data are analyzed by first replacing each failure-time interval with a single value. For example, in one such method, midpoint imputation, the analysis is performed as

* *Corresponding author's email address:* finkel@hsph.harvard.edu

Key words: Cancer; HIV; Missing data; Rank-based inference; Survival.

though the midpoint of each subject's interval was the exact event time. This method, however, can lead to biased estimates if the intervals are wide and varied (Law and Brookmeyer, 1992). Furthermore, it underestimates the standard errors of the coefficients since it treats the survival times as known when, in fact, they are not. Another common approach is to use the right endpoint of the interval; this estimator has drawbacks similar to those of midpoint imputation.

Other, more complex methods have also been proposed for analyzing interval-censored failure-time data. Brookmeyer and Goedert (1989) propose a two-stage regression model for the analysis of interval-censored latency times, but this relies on parametric assumptions. Finkelstein (1986) develops a score test that can be used for two-group comparisons when the data are interval censored. This method is based on the full likelihood under the proportional hazards model and requires estimation of the underlying baseline hazard function. Since the number of parameters to be estimated may increase with the number of event times, this can be numerically unstable and computationally intensive for some data sets. Huang and Wellner (1995) obtain maximum likelihood estimates for the Cox model with interval-censored data using the convex minorant algorithm. They show that their parameter estimates are asymptotically normal and efficient under appropriate regularity conditions. Their method, however, requires estimation of the baseline hazard and can have numerical problems when the number of subjects is large.

As interval-censored data can be viewed as incomplete data, a natural approach to the regression problem is to derive the maximum marginal likelihood estimates from the proportional hazards model using an EM algorithm (Dempster, Laird, and Rubin, 1977). Because the required expected log-likelihood is hard to calculate in closed form, a Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) is useful. Sinha, Tanner, and Hall (1994) use MCEM to estimate Cox model parameters with grouped failure-time data. However, they sample from the distribution of the exact failure orderings by enumerating every possible ordering, and this calculation becomes excessively complex for many data sets with interval-censored survival times due to the large number of such orderings.

Others have proposed simulation methods that are not based on the EM algorithm. Self and Grossman (1986) developed a two-sample linear rank test for interval-censored data using an algorithm that shuffles under the null distribution. Satten (1996) estimates the Cox model for interval-censored data using a stochastic approximation based on the Robbins–Munro process and a Gibbs sampler to sample rankings consistent with the observed data.

This paper proposes a method for fitting the Cox model using Monte Carlo EM. We use a Markov chain Monte Carlo (MCMC) sampling method in the E step that makes enumeration of all rankings unnecessary. Our sampling algorithm draws from the set of complete rankings by repeatedly shuffling the exact ranks of failures, keeping them consistent with both the observed data and the proportional hazards model. The algorithm is intuitive and computationally efficient. In Section 2, we describe the estimation methodology; in Section 3, we describe variance estimation; and in Section 4, we discuss techniques for improving the performance of the algorithm. Section 5 provides results of a simulation study, Section 6 applies the method to several examples, and Section 7 gives conclusions.

2. Maximum Likelihood Estimates by Monte Carlo EM

Under the proportional hazards model, the hazard of the event at time t given a vector of covariates x is

$$\lambda(t \mid x) = \lambda_0(t)e^{\beta x},$$

where λ_0 is the unknown baseline hazard rate and β is a vector of regression coefficients. For untied exact and right-censored data, the N failure times are ranked and only the vector $\mathbf{r}$ of rankings is used in the analysis. Inference is then based on the partial likelihood

$$\mathcal{L}(\beta \mid \mathbf{r}) \propto P(\mathbf{r} \mid \beta) = \prod_{i=1}^{N} \frac{e^{\beta x_i}}{\sum_{j \in R(i)} e^{\beta x_j}}, \tag{1}$$

where $R(i)$ is the set of individuals at risk for failure at the time of the ith ranked failure and x_i is the covariate vector for that individual. Kalbfleisch and Prentice (1980) also derive (1) as the marginal likelihood for β based on the ranking of the failure times. It is assumed that the observations are censored into intervals noninformatively, i.e., that, conditional on x, the censoring mechanism is independent of the actual failure times.

With interval-censored data, we observe the vector $\mathbf{I} = ((L_1, R_1), (L_2, R_2), \ldots, (L_N, R_N))$ of ordered pairs defining the censoring intervals. The failure time for case i is known to fall in the

interval $(L_i, R_i]$, where L_i is the left endpoint of the interval for observation i and R_i is the right endpoint of its interval. Both right censoring, $R_i = \infty$, and left censoring, $L_i = -\infty$, are special cases of interval censoring.

If observations are censored into overlapping intervals, we do not know their relative rankings with certainty. Then (1) cannot be directly maximized since, instead of knowing the rankings of the subjects, we have a set of admissible rankings $\mathcal{A}(\mathbf{I})$ that are consistent with the interval-censored data. The likelihood based on the incomplete ranks is the marginal likelihood

$$\mathcal{L}(\beta \mid \mathbf{I}) \propto \sum_{\mathbf{r} \in \mathcal{A}(\mathbf{I})} P(\mathbf{r} \mid \beta)$$

(Kalbfleisch and Prentice, 1980, pp. 72–73). This likelihood is difficult to maximize because of the potentially large set of rankings $\mathcal{A}(\mathbf{I})$. Instead, we maximize the likelihood using an EM algorithm. For interval-censored data, the E step of the EM algorithm finds the expected value of the log-likelihood of the data under the probability distribution of the orderings,

$$\sum_{\mathbf{r} \in \mathcal{A}(\mathbf{I})} \log[\mathcal{L}(\beta \mid \mathbf{r})] P\left(\mathbf{r} \mid \hat{\beta}_c\right), \tag{3}$$

where $\hat{\beta}_c$ denotes the current estimate of β and $P(\mathbf{r} \mid \hat{\beta}_c)$ is the probability of complete ranking $\mathbf{r}$ given $\hat{\beta}_c$. The M step maximizes (3) to update the parameter estimate.

If there are many overlapping intervals, it is difficult to maximize (3) because of the large number of possible orderings. Therefore, we propose using a Monte Carlo EM (MCEM) algorithm with Monte Carlo integration in the E step. In this step, we draw a sample of M complete orderings, $\mathbf{r}_1, \ldots, \mathbf{r}_M$, with replacement, from the distribution of all possible orderings, given the incomplete data $\mathbf{I}$ and the current parameter estimate $\hat{\beta}_c$. In the M-step, we maximize the average of the individual log-likelihoods

$$\frac{1}{M} \sum_{m=1}^{M} \log \mathcal{L}(\beta \mid \mathbf{r}_m) \tag{4}$$

to update $\hat{\beta}_c$. We repeat these steps, increasing M if necessary, until the parameter estimates converge. Note that we are using the MCEM to maximize the likelihood for an uncensored proportional hazards model. This is slightly different from the usual procedure of maximizing the censored data likelihood because we are imputing rankings for the failure times of all subjects, including right-censored observations. The marginal likelihood (2) is identical to the standard partial likelihood for right-censored data. Nonetheless, our method uses a complete ranking because the ranks of some of the right-censored observations may be required to determine the risk sets of some of the interval-censored observations.

Each $\mathbf{r}_m$ must be drawn from $\mathcal{A}(\mathbf{I})$ with probabilities

$$P(\mathbf{r} \mid \hat{\beta}_c, \mathbf{I}) = \frac{P(\mathbf{r} \mid \hat{\beta}_c)}{\sum_{\mathbf{r}' \in \mathcal{A}(\mathbf{I})} P(\mathbf{r}' \mid \hat{\beta}_c)}. \tag{5}$$

Direct calculation of these probabilities can be difficult with interval-censored data because it requires the enumeration of all possible orderings. Instead, we draw from $\mathcal{A}(\mathbf{I})$ in a way that makes the probability of selecting each ranking approximately equal to (5), using an MCMC algorithm (Hastings, 1970). The algorithm repeatedly shuffles the exact ranks of the failures, while keeping them consistent with the interval-censored data, in such a way that the distribution of the exact ranking converges to (5). In each shuffle, we take one pass through the data from the first-ranked pair of observations to the last. We randomly switch or leave as is the first two failures under the current ranking, then the second and third, and so on, up to the last two. The probabilities of each switch are chosen so that repetition of the overall shuffle produces draws from the desired distribution.

We now describe the details of a single possible switch. Assume a current parameter estimate $\hat{\beta}_c$ and a current ranking $\mathbf{r} \in \mathcal{A}(\mathbf{I})$. Consider any two observations whose failure times are adjacent in $\mathbf{r}$. If the ranks of their failure times are switched, we obtain a new ranking $\mathbf{r}^*$. If the censoring intervals of the two observations do not overlap, then $\mathbf{r}^* \notin \mathcal{A}(\mathbf{I})$ and we move on to consider another pair of observations. Conversely, if the intervals do overlap, then $\mathbf{r}^* \in \mathcal{A}(\mathbf{I})$. Then we can compare the probabilities that $\mathbf{r}$ or $\mathbf{r}^*$ is the correct ranking, given $\hat{\beta}_c$ and the rankings of all failure times

other than the two that were switched (which restricts the possible rankings to $\mathbf{r}$ and $\mathbf{r}^*$). The ratio of these probabilities is easy to compute because the failures and risk sets at every failure except the two that were switched are the same for $\mathbf{r}$ and $\mathbf{r}^*$; the ratio is

$$\frac{P(\mathbf{r}^* \mid \hat{\beta}_c)}{P(\mathbf{r} \mid \hat{\beta}_c)} = \frac{\left(\sum_{i \in R} e^{\hat{\beta}_c x_i}\right) - e^{\hat{\beta}_c x_k}}{\left(\sum_{i \in R} e^{\hat{\beta}_c x_i}\right) - e^{\hat{\beta}_c x_j}}, \tag{6}$$

where observation j fails just before observation k under ranking $\mathbf{r}$ and just after k under ranking $\mathbf{r}^*$ and R is the risk set just prior to the failures of observations j and k.

Suppose that the probability of each switch is

$$\alpha \min\left(1, \frac{P(\mathbf{r}^* \mid \hat{\beta}_c)}{P(\mathbf{r} \mid \hat{\beta}_c)}\right). \tag{7}$$

Choosing $\alpha = 1$ results in the well-known Metropolis algorithm (Hastings, 1970). This produces the highest acceptance rates possible with an algorithm of this form. On the other hand, with $\alpha = 1$, it is possible that the chain of draws will be periodic or nearly so. In particular, with $\alpha = 1$ and $\hat{\beta}_c = 0$, every potential switch is accepted, so the algorithm alternates between odd and even permutations of the starting ordering. With α slightly less than one, we stay put with probability $1 - \alpha$ and perform a Metropolis step otherwise. The proof that this algorithm converges to the desired distribution (5) is outlined in the Appendix; the choice of α and the choice of other parameters of the algorithm are discussed in Section 4.

After a predetermined number of shuffles, H, the current permutation is saved as a draw $\mathbf{r}_m$. This process is repeated to generate M draws of completed rankings $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M$, completing an E step.

We obtain an initial $\mathbf{r} \in \mathcal{A}(\mathbf{I})$ by ranking observations by the midpoints of their intervals. An initial estimate of β is the MLE for this completed data set.

3. Estimation of Variance

There are three sources of variation of our estimate of β. The first is the ordinary complete data variance. The second is the extra variance resulting from the fact that we are estimating the parameters in the presence of incomplete data. The third source of variance is the randomness of the Monte Carlo E step. Using a result from Louis (1982), the information matrix for β is given by

$$\mathcal{I}(\beta \mid \mathbf{I}) = -E\left(\frac{\partial^2 \log \mathcal{L}(\beta \mid \mathbf{r})}{\partial^2 \beta}\right) - \operatorname{var}\left(\frac{\partial \log \mathcal{L}(\beta \mid \mathbf{I})}{\partial \beta}\right). \tag{8}$$

The first term of (8) is the ordinary complete data information and is estimated by

$$-\frac{1}{M}\sum_{1}^{M} \frac{\partial^2 \log \mathcal{L}(\beta \mid \mathbf{r}_m)}{\partial^2 \beta} \tag{9}$$

evaluated at $\beta = \hat{\beta}$.

Defining $S_m = \partial \log \mathcal{L}(\beta \mid \mathbf{r}_m)/\partial \beta$, we estimate the second term in (8) with

$$\frac{1}{M}\sum_{1}^{M} S_m^2 \tag{10}$$

evaluated at $\beta = \hat{\beta}$. (Note that, because at the MLE $\Sigma_1^M S_m = 0$, the variance can be estimated by a sum of squares.) If β is vector valued, interpret the square as outer product. This term represents the loss of information due to the uncertainty about the exact ranking of the failure times.

The third source of variance results from use of a Monte Carlo sampling scheme to approximate the expected value of the score statistic. This contribution is independent of the other two and is a consequence of our algorithm and not the actual data. To account for the additional Monte Carlo variance, also subtracted from the observed information is

$$\frac{1}{M(M-1)}\sum_{1}^{M} S_m^2. \tag{11}$$

This quantity is just $1/(M-1)$ times (10) and, by making M large enough, we can make this component of variance negligible. To estimate the variance of our parameter estimates, we subtract (10) and (11) from (9) and invert the resulting matrix.

4. Setting Parameters of the Algorithm

To use this algorithm, we must choose the number of shuffles between draws, H, the number M of completed data sets to be created, and the shuffling parameter α. Since positive autocorrelation of the sequence $S_1, \ldots, S_m$ increases the Monte Carlo variance, increased autocorrelation also increases the number of draws M required to obtain desired precision. By retaining the ranking on only one out of every H shuffles, we are discarding some information. This is justified since the shuffling requires few additional computations once we have calculated $e^{\hat{\beta}_c x_i}$ for each i, whereas each draw saved also requires calculation of a score and Hessian for the M step. Therefore, increasing M is more expensive computationally than increasing H. Furthermore, we would like to have H large enough so that the $\mathbf{r}_1, \ldots, \mathbf{r}_M$ are nearly independent, as assumed by the variance estimator in Section 3. Because the autocorrelation of the S_m goes to zero as $H \to \infty$, we can increase H until the autocorrelation is suitably small.

In determining M, we need to look at the extra Monte Carlo variance that results from the fact that we are sampling the possible complete data vectors rather than enumerating all of them. We can determine an acceptable value of M by directly calculating an estimate of the variance from the formulas given above and by monitoring the convergence of the successive estimates of β. Wei and Tanner (1990) recommend using a small value of M for the first few iterations to get close to the maximizer and then increasing M for later iterations to get a more precise estimate. Once the additional variance that results from using a moderate value of M, as opposed to a very large value of M, is negligible compared to the overall variance of the estimate, there is no benefit in increasing M further.

Larger values of α mean higher acceptance rates for new permutations and faster movement through the possible orderings. Consequently, fewer shuffles are required between draws to achieve desired levels of autocorrelation. Because the amount of shuffling required increases rapidly as α moves away from one, a value of $\alpha = 0.99$ is recommended. For very small data sets, a slightly smaller value for α is preferable in order to guarantee that periodicity of the orderings does not occur.

The number of EM iterations required for the parameter estimate to converge ranged between 5 and 10 in our examples and simulations. Because of the Monte Carlo error in our estimate of the expected score, our estimate of β will not converge completely but rather will move in the direction of the correct value and then fluctuate around it. Data sets with a larger percentage of missing information require more iterations to achieve convergence.

A modification of this MCEM algorithm can be useful when the fraction of missing information is large and hence both the shuffling step and convergence of the MCEM are slow. Instead of running the shuffling algorithm to draw a new sample at every E step, the sample from the previous E step can be reweighted for replacement of the previous estimate $\hat{\beta}_c$ by the new estimate $\hat{\beta}_c^*$. The weighting factor for $\mathbf{r}_m$ is proportional to $P(\mathbf{r}_m \mid \hat{\beta}_c^*)/P(\mathbf{r}_m \mid \hat{\beta}_c)$, which can be calculated directly. The average likelihood in (4) is then replace by a weighted average. This procedure obviates the need for reshuffling during a series of small steps in β. Furthermore, no additional Monte Carlo randomness is introduced from step to step, so convergence acceleration methods (Louis, 1982) are applicable. (Further details, including variance estimation with weighted draws, appear in Goggins (1997).)

5. Simulation Studies

5.1 *Simulation Design*

In order to test the performance of our algorithm we did simulation studies to assess the level, power, and possible bias of the algorithm. All simulations involved generating exact, exponentially distributed data for two samples with a specified true hazard ratio and then randomly censoring the data into irregular and overlapping intervals. We did this as follows:

(1) For group 1, we generate exact failure times from an exponential distribution with hazard rate λ_0.
(2) For group 2, we generate exact failure times from an exponential distribution with hazard rate $\lambda_1 = e^{\beta}\lambda_0$, where e^{β} is the desired true hazard ratio for this simulation.
(3) We randomly censor the data into intervals as follows:

(a) Every integer time is a potential screening time.
(b) For each subject, we draw a vector of independent Bernoulli variables with fixed probability of success; failures represent missed check-ups and successes represent made check-ups.
(c) Each outcome is interval censored to the (integral) times of the nearest success before and after.
(d) We use our algorithm (and a traditional alternative) to estimate $\hat{\beta}$ from the censored data and (for comparison) fit the Cox model to the uncensored data.

This simulates the kind of interval censoring that might occur if individuals were screened at monthly visits and a certain percentage of these visits were missed independently of covariates or event time. In order to test whether hypothesis tests based on our method have the correct size (rejection rate), data were simulated for two groups of cases with $\lambda_0 = \lambda_1 = 0.1$, and $P(\text{missed visit}) = 0.8$. Once we estimated $\hat{\beta}$ and $\widehat{\text{s.e.}}(\hat{\beta})$ for these data sets, we then compared the observed rejection rate, where rejection means $|z| = |\hat{\beta}/\widehat{\text{s.e.}}(\hat{\beta})| > 1.96$, with the nominal rate of 5% and the observed rate for the uncensored tests.

To assess the power of our method, we again simulated exponential data from two groups, this time with $\lambda_0 = 0.1$ and $\lambda_1 = 0.1487$, the hazard ratio necessary in order to have 80% power to detect the difference with the Cox model given a sample size of 200 (100 in each group) and uncensored data. We estimated the power of each method by calculating the observed proportion of times that the null was rejected. Our method should have less power than the analysis of uncensored data because we lose some information through interval censoring.

In order to assess bias, we compared the average of $\hat{\beta}$ from each set of simulations to the true β from which the data was generated and to the $\hat{\beta}$ estimated using Cox regression on the corresponding exact data. We also compared our results to those obtained using midpoint imputation on the interval-censored data.

To assess whether the variance estimates for the coefficient estimates were correct, we compared the sample variance of the $\hat{\beta}$ obtained from the simulations to the estimated variances derived in Section 3. We also separately tested our method of estimating the additional Monte Carlo variance. We did this by running our algorithm for 200 MCEM iterations on an interval-censored data set and then calculating the sample variance of the $\hat{\beta}$'s from each iteration (excluding the estimates before convergence). We compared this variance to the estimated additional variance resulting from subtracting (11) from the information before inverting it. This was done for 20 simulated data sets.

5.2 *Simulation Results*

The simulation program was written using the Splus statistical programming language, with calls to C subroutines for the more computing-intensive parts of the algorithm. The average CPU time used per data set was 2.7 minutes on a Sparc 2000. We simulated 1000 data sets each for the size test and the power test. A simple summary of the severity of interval censoring is the fraction of all pairs of intervals that overlap. Setting the proportion of missed visits at 80% resulted in mean proportion of overlap of 56%, with a range over the data sets of 50–61%.

The results of simulations done to test size with $\beta = 0$ are shown in Table 1. The observed rejection rate for our method was 4.5%, close to the nominal 5%. The mean of the observed $\hat{\beta}$'s estimated using our method was very close to both the true β of zero and the observed mean of the $\hat{\beta}$'s estimated from the uncensored data. The observed standard error of the $\hat{\beta}$'s from our method was slightly less than the mean of standard error estimates for the individual data sets, 0.149. A comparison of our results with those obtained using midpoint imputation shows that the two

Table 1
Results of simulations to test size. True hazard ratio = 1.0, true $\beta = 0.00$, number of data sets = 1000, $N = 200$.

	Cox on exact data	Midpoint imputation	Proposed method
Mean of $\hat{\beta}$	−0.0078	−0.0074	−0.0090
95% C.I. for E[$\hat{\beta}$]	(−0.017, 0.001)	(−0.016, 0.002)	(−0.018, 0.001)
Mean $\widehat{\text{s.e.}}$ of $\hat{\beta}$	0.141	0.140	0.147
Observed size	4.3%	4.2%	4.5%

Table 2
Results of simulations to test power. True hazard ratio = 1.487, *true* $\beta = 0.397$, *number of data sets* $= 1000$, $N = 200$.

	Cox on exact data	Midpoint imputation	Proposed method
Mean of $\hat{\beta}$	0.396	0.366	0.405
95% C.I. for E[$\hat{\beta}$]	(0.387, 0.405)	(0.357, 0.375)	(0.395, 0.415)
Mean $\widehat{\text{s.e.}}$ of $\hat{\beta}$	0.145	0.144	0.156
Observed power	77.8%	72.4%	75.5%

methods performed equally well in this situation, as would be expected because, with $\beta = 0$, the failure model provides no information for imputation of censored failure times.

The observed power when using standard Cox inference procedures on the exact data was close to that expected by design. As anticipated, both our method and the midpoint imputation method showed slightly lower power when applied to the corresponding interval-censored data. Our method rejected more often than midpoint imputation and, by McNemar's test, the difference was significant ($\chi^2 = 17.6, p < 0.001$). The mean and 95% confidence intervals for β for each method are shown in Table 2. The mean of the coefficient estimates for our method agrees very closely with both the true β and the observed mean of the estimates from the exact data. The estimates from midpoint imputation had a substantial and significant negative bias, about 20% of the estimated standard error. As expected, the variance of the coefficient estimates from our method was somewhat higher than the variance of the estimates from the exact data, reflecting the additional uncertainty due to interval censoring. The observed standard error of our coefficient estimates was nearly identical to mean of the estimated standard errors for the individual data sets. On the whole, our method performed better than midpoint imputation under these circumstances, showing lower overall bias and coming closer to the answer obtained using the exact data for 58% of the data sets ($p < 0.001$).

The results of the simulations done to test our method of calculating the extra Monte Carlo variance showed close agreement of the sample between-iteration variance of the $\hat{\beta}$'s with the additional variance predicted by using our method. This estimated additional variance was quite small, adding less than 1% to our standard error estimates in all cases. We were able to obtain this close agreement between the actual Monte Carlo variance of the estimates and the variance estimated by (11) by increasing H until the estimated lag 1 autocorrelation was consistently less than 0.1.

6. Applications

6.1 *Breast Cosmesis Data*

For this data set, described in Finkelstein (1986), breast cancer patients were monitored to determine the long-term effects of therapy on the occurrence of cosmetic deterioration. In this analysis, the outcome was time until the onset of breast retraction and the covariate was a binary indicator of whether or not a patient received adjuvant chemotherapy following initial radiation treatment. A total of 94 patients were followed and were seen, on average, every 4–6 months. The frequency of visits decreased with increasing time from the completion of treatment. Forty percent of the patients had not experienced retraction by the time of their final visit and therefore were right censored. Because the time between visits was irregular and many patients missed visits, the data must be treated as interval censored rather than grouped. The proportion of overlap (fraction of all intervals that overlap) was 46%.

The parameter estimate obtained using our method, $\hat{\beta} = 1.45$ (s.e. $= 0.371$), was much higher than that obtained using midpoint imputation, 0.66 (s.e. $= 0.224$), and resulted in a higher z-statistic, 3.91, than was obtained using Finkelstein's (1986) method, 2.86.

There was a considerable loss of information due to interval censoring. The complete data information, from (9), was 14.4, and the amount subtracted, (10) and (11), was 7.2, representing a 50% loss. This was a considerably greater loss than that observed for the simulated data sets, even though the simulations produced data with a higher proportion of overlap. This illustrates that proportion of overlap does not necessarily capture the effect of the interval censoring on standard error.

6.2 *AZT Resistance Data*

This data set, which is described in Richman, Grimes, and Lagakos (1990), concerns the development of resistance to AZT in AIDS patients. The data set describes 31 patients from four AIDS Clinical Trials Group (ACTG) studies who had at least one drug resistance assay. The data are sparse because assays are expensive. All patients were assumed to be sensitive at baseline so, if a patient's first assay showed resistance (13 of the patients), the observation would be considered interval censored, with a left endpoint of zero and a right endpoint equal to the time of the first assay. There were also 13 patients who did not show resistance by their last assay (right censored). The covariate analyzed here is dose of AZT, which is coded as low = 0 and high = 1. The sparseness of the data is reflected in the fact that the average proportion of overlap is 83%.

We found $\hat{\beta} = 1.01$ with a standard error of 0.75. This gave a z-statistic of 1.34 and a p-value of 0.18. Not surprisingly, given the sparseness of the data, there was a considerable loss of information due to the interval censoring. The complete data information was 5.04, and the amount subtracted was 3.26, representing a 65% loss. Had we calculated the standard error using just the complete data information, it would have been 0.45, yielding a z-statistic of 2.3, a misleadingly significant result. This illustrates the importance of accounting for the additional uncertainty due to censoring. For comparison, midpoint imputation yields a parameter estimate of 0.70, a standard error of 0.58, and a z-statistic of 1.21. Finkelstein's score test gives a z-statistic of 1.43.

6.3 *Hemophilia Data*

This data set, from Kim et al. (1993), describes a population of 257 hemophiliacs who were treated at two French hospitals beginning in 1978. By the time this data set was compiled, 188 of these patients were found to be HIV positive. Because individual HIV infection status was determined by retrospectively testing stored blood sera, the infection time is only known to fall between the times of the last seronegative sample and the first seropositive one. While Kim et al. (1993) analyze the latency, here we analyze the infection times themselves. The covariate to be analyzed is treatment level, which is coded as heavily treated ($x = 1$) if the patient received at least 1000 mg/kg for 1 or more years between 1982 and 1985 and lightly treated ($x = 0$) otherwise. We assumed that patients who were heavily treated during this period had been heavily treated since 1978.

Time to infection is measured in 6-month intervals starting from January 1, 1978. While the width of the censoring intervals varies, the data are far less sparse than the AZT resistance data, with an average percentage of overlap of 38%. Our analysis estimates $\hat{\beta} = 0.89$, with a standard error of 0.15 and a z-statistic of 5.86. This agrees closely with the z-statistic of 5.99 obtained from the Finkelstein (1986) method. The use of midpoint imputation gives a lower parameter estimate, 0.60, and z-statistic, 4.51. This reflects bias of this method when the true hazard ratio is far from one. The loss of information associated with the interval censoring was fairly small. The complete data information was 52.1, and the amount subtracted was 8.96, representing a 17% loss. The use of all three methods lead to the same conclusion—that heavily treated patients tended to be infected earlier.

For all three examples, the required CPU time for the analysis was approximately 1.5 minutes, the number of shuffles H required to achieve small lag 1 autocorrelation was 20–30, and the number of imputations at each iteration required to make the Monte Carlo variation negligible was $M = 500$.

7. Conclusions

We have presented here an approach, using an adaptation of the Cox proportional hazards model, that can be used to analyze the effect of covariates on failure-time random variables when the failure times are interval censored. This approach can be utilized when the censoring is arbitrary, e.g., irregular intervals, or when the data is grouped. Although parametric methods may be more powerful when their assumptions are in fact correct, our method has the appeal of the Cox model for uncensored data in that it makes no parametric assumptions about the underlying distribution of survival times and does not require estimation of the baseline hazard function. Our method also yields standard error estimates that take into account the extra variance associated with interval censoring. Simulation studies have shown good overall performance, i.e., correct size, power close to that obtained with exact data, and little evidence of bias, on data that was fairly heavily censored. Although the method is computationally intensive, even large data sets with heavy censoring can be analyzed with reasonable expenditure of CPU time. The program is available from the authors.

Acknowledgement

Support for this research came from ACTG under NIAID contract N0-AI 95030 and NIH grant CA 74302.

Résumé

Cet article propose un algorithme de Monte Carlo EM (MCEM) pour ajuster le modèle à risques proportionnels pour des données censurées par intervalle. L'agorithme génère l'ordre des échecs à partir de leur distribution de probabilité sous le modèle. Nous maximisons la moyenne des log-vraisemblances à partir de ces jeux de données complétés pour obtenir des estimations de paramètres mis à jour. Comme dans le modèle de Cox (1972) standard, cet algorithme ne nécessite pas l'estimation de la fonction de risque instantané de base. Les performances de cet algorithme sont évaluées à l'aide de simulation et la méthode est appliquée à des données provenant d'études sur le SIDA et le cancer. Nos résultats indiquent que cette méthode fournit des estimations non biaisées et plus précises que les méthodes imputant les valeurs de droite ou du milieu de l'intervalle.

References

Besag, J., Green, P., Higdon, D., and Mengerson, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10,** 3–66.

Brookmeyer, R. and Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* **45,** 325–335.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

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

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Goggins, W. B. (1997). Monte Carlo EM methods for analyzing survival data in the presence of interval censoring. Ph.D. thesis, Harvard School of Public Health, Boston.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109.

Huang, J. and Wellner, J. A. (1995). *Efficient estimation for the proportional hazards model with "case 2" interval censoring.* Technical Report 290, Department of Statistics, University of Washington, Seattle.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Kim, M. Y., De Gruttola, V., and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49,** 13–22.

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* **11,** 1569–1578.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Parzen, E. (1962). *Stochastic Processes.* San Francisco: Holden-Dey.

Richman, D. D., Grimes, J. M., and Lagakos, S. W. (1990). Effect of stage of disease and drug dose on zidovudine susceptibilities of isolates of human immunodeficiency virus. *Journal of AIDS* **3,** 1139–1145.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83,** 355–370.

Self, S. G. and Grossman, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics* **42,** 521–530.

Sinha, D., Tanner, M. A., and Hall, W. J. (1994). Maximizing the marginal likelihood from grouped survival data. *Biometrika* **81,** 53–60.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85,** 699–704.

Received February 1997; *revised November* 1997; *accepted December* 1997.

APPENDIX

We outline here the proof that the distribution of permutations produced by the shuffling algorithm converges to the probability distribution, under the proportional hazards model, of the possible orderings of the data. (Details of the proof appear in Goggins (1997).) The successive rankings constitute a Markov chain because each ranking has a probability distribution for which dependence on the previous ranking is only through the ranking immediately preceding it. To show that the chain converges to the desired distribution, we must show that it is homogeneous, irreducible, aperiodic, and positive recurrent (Parzen, 1962, p. 256) and that the derived distribution is a fixed point under a complete shuffle. Then the ergodic theorem guarantees (Besag et al., 1995) that, for any starting point,

$$\frac{1}{M}\sum_{m=1}^{M}\log\mathcal{L}(\beta \mid \mathbf{r}_m)$$

converges almost surely to $\mathrm{E}[\log\mathcal{L}(\beta \mid \mathbf{I})]$ as $M \to \infty$.

The Markov chain consisting of the exact rankings generated at the completion of each shuffle is homogeneous because its transition matrix is the composition of a fixed sequence of transition matrices corresponding to the schedule of possible switches and the matrix for each switch depends only on $\mathbf{I}$, which is fixed, and $\hat{\beta}_c$, which is fixed through each E step. The Markov chain is aperiodic because it can leave the ranking unchanged with positive probability, at each potential switch, from (7) with $\alpha < 1$. Therefore, the probability that a complete shuffle also leaves the ranking unchanged is also positive.

In order to show that the chain is irreducible, we must show that it is possible to get from any $\mathbf{r} \in \mathcal{A}$ to any other $\mathbf{r}^* \in \mathcal{A}$ by a sequence of shuffles. We first note that it is possible to go from $\mathbf{r}$ to the base ranking, where the observations are ordered by the left endpoints of the intervals into which their event times were censored, with ties broken by the ordering of the right endpoints if they differ and otherwise arbitrarily. This state is reached, with nonzero probability, by a sequence of switches in which the failure falling to the left in the base ranking is always switched to the left. A similar sequence of switches in reversed order leads from the base ranking to $\mathbf{r}^*$.

The chain must be positive recurrent because it is irreducible with finite state space (Parzen, 1962, p. 258).

Finally, to show that the derived distribution is a fixed point under a complete shuffle, it suffices to show that it is a fixed point for each possible switch. This follows from the usual arguments for the Metropolis algorithm because switching probabilities of the form (7) satisfy the conditions of Hastings (1970).