Published in final edited form as:
Biometrics (2015), **71**(3): 782–791
DOI: 10.1111/biom.12302

Penalized Regression for Interval-Censored Times of Disease Progression: Selection of HLA Markers in Psoriatic Arthritis

YING WU

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: rjcook@uwaterloo.ca

Summary

Times of disease progression are interval-censored when progression status is only known at a series of assessment times. This situation arises routinely in clinical trials and cohort studies when events of interest are only detectable upon imaging, based on blood tests, or upon careful clinical examination. We consider the problem of selecting important prognostic biomarkers from a large set of candidates when disease progression status is only known at irregularly spaced and individual-specific assessment times. Penalized regression techniques (e.g. LASSO, adaptive LASSO and SCAD) are adapted to handle interval-censored time of disease progression. An expectation-maximization algorithm is described which is empirically shown to perform well. Application to the motivating study of the development of arthritis mutilans in patients with psoriatic arthritis is given and several important human leukocyte antigen (HLA) variables are identified for further investigation.

Keywords: EM algorithm; Interval-censoring; LASSO; Penalized regression; SCAD; Variable selection

1 INTRODUCTION

1.1 VARIABLE SELECTION AND PENALIZED REGRESSION

Breiman (1996) noted that the traditional method of best subset selection was unstable and that this instability could lead to poor predictive performance. Ridge regression (Hoerl and Kennard, 1970) imposes some shrinkage which leads to more stable models, but does not set any coefficients to zero and therefore does not "select" key variables. The LASSO (Tibshirani, 1996) attempts to maintain

the advantages of both subset selection and ridge regression by shrinking some coefficients and setting other coefficients to zero through use of a log-likelihood with an L_1 penalty function. Other penalty functions which have recently been proposed include the smoothly-clipped absolute deviation (SCAD) (Fan and Li, 2001, Zou and Li, 2008), the adaptive LASSO (Zou, 2006), the elastic net (Zou and Hastie, 2005), the grouped LASSO (Yuan and Lin, 2005), and the minimax concave penalty (MCP) (Zhang, 2010).

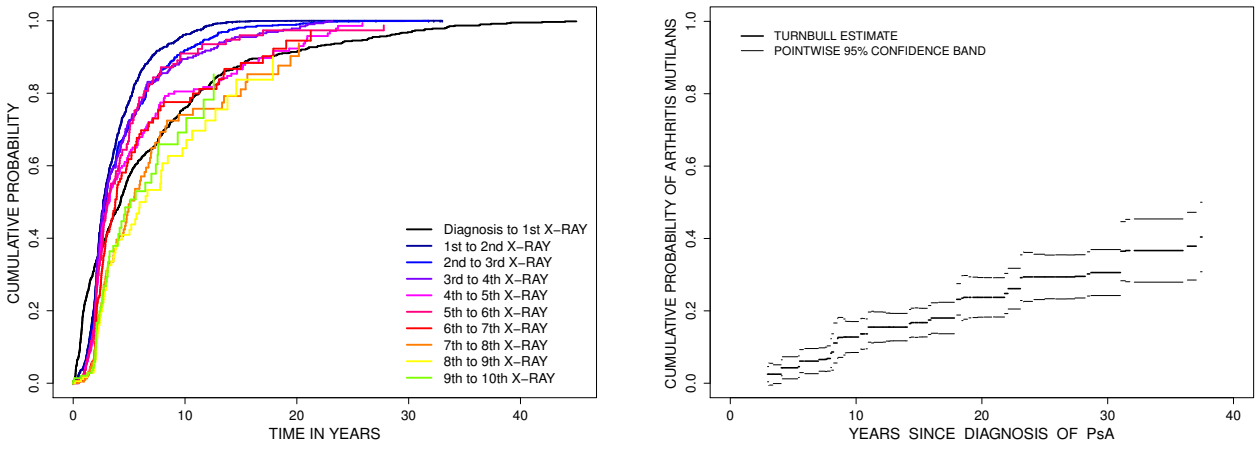
While much of the work on variable selection techniques was initially carried out in the context of continuous responses, advances have been made to deal with binary responses and time to event responses. For the latter, the penalty term is typically applied to the partial likelihood arising from a semiparametric Cox regression model (Cox, 1972) when data are right-censored.

Witten and Tibshirani (2009) give an excellent overview of the challenges arising with particularly high dimensional covariate data in settings with censored outcomes and provide an extensive discussion of the specific objectives one might have in particular scientific contexts; another useful account can be found in Li and Ma (2013). The inherent difficulty in obtaining robust and generalizable findings from samples with censored responses and high dimensional covariates is evident from the inconsistency of findings across seemingly similar patient populations (McShane et al., 2005a). The limitations due to inadequate sample size (Polley et al., 2013) and the inconsistency of findings across studies has led to an increased interest in synthesizing findings over multiple studies. Assimilating information from several sources can be helpful, but it is important to clearly understand the differences between the frameworks and goals of the studies contributing to this synthesis. Guidelines have been developed for reporting findings from biomarker studies with this in mind, which advocate clear statements of study objectives, study design, methods of processing samples, and the approach to statistical analysis (McShane et al., 2005b, Altman et al., 2012).

Many prospective studies, however, involve event times subject to interval censoring (Sun, 2006). In cancer clinical trials, for example, new metastatic lesions are often only detectable by imaging (Hortobagyi et al., 1996), so the time from randomization to the development of a new lesion is unknown. In patients infected with cytomegalovirus, the time from infection to viral shedding in the blood is only known to lie between the last negative and first positive serum sample (Betensky and Finkelstein, 1999). The occurrence of an asymptomatic fracture in osteoporosis patients is only detected by radiographic examination (Riggs et al., 1990).

We consider the problem of variable selection in the context of interval-censored time to event data. We adopt a flexible piecewise exponential model (Friedman, 1982) for the event of interest and penalize the complete data likelihood constructed by treating the interval-censored failure times as known. An expectation-maximization (EM) algorithm (Dempster et al., 1977) is then used for variable selection through optimization of the observed data likelihood incorporating the LASSO, adaptive LASSO or SCAD penalty function.

The remainder of the article is organized as follows. In Section 1.2 we describe the motivating study with the goal of identifying key human leukocyte antigens associated with the development of arthritis mutilans in a cohort of individuals with psoriatic arthritis. In Section 2 we describe a penalized EM algorithm based on a piecewise exponential response model, for which existing techniques for variable selection can be exploited to handle interval-censored event times. Simulation studies involving multivariate normal and correlated binary covariates reported on in Section 3 demonstrate superior performance of the proposed method over analyses based on mid-point imputation. Additional simulation studies are described in *Supplementary Material: Web Appendix A* and studies of different criteria for selection of tuning parameters (Bradic et al., 2011) are given in *Supplementary Material: Web Appendix B*. The data from the psoriatic arthritis clinic are analyzed in Section 4 using a variety of penalty functions, and concluding remarks are given in Section 5.



(a) (b)

Figure 1: Plots of the estimated cumulative distribution functions for the time from psoriatic arthritis diagnosis and clinic entry (Kaplan-Meier estimate) and the times between radiological assessments based on a semi-Markov model with a gamma frailty (panel (a)) and the Turnbull estimate with a pointwise 95% confidence band for the marginal cumulative distribution function of the time from disease onset to arthritis mutilans (panel (b))

1.2 Prognostic HLA Markers in Psoriatic Arthritis

The University of Toronto Psoriatic Arthritis Clinic is a tertiary referral center for individuals with psoriatic arthritis (PsA), an immunological condition which features both skin and joint involvement (Chandran et al., 2010). A registry was created in 1976, which has been recruiting and following patients continuously since its inception, making it one of the largest cohorts of patients with PsA in the world.

Patients undergo a detailed clinical and radiological examination upon entry to the clinic, and provide serum samples for genetic testing. Follow-up clinical and radiological assessments are scheduled annually and every two years respectively in order to track changes in joint damage. At each radiological assessment the degree of damage is recorded in sixty-four joints on a five point scale (Rahman et al., 1998). Arthritis mutilans is a particularly aggressive form of arthritis characterized here by five or more joints with the highest grade of damage. Identification of genetic features associated with this condition is important to help identify patients warranting prophylactic treatment with more effective but costly anti-TNF therapy (Kyle et al., 2005) and to help guide the selection of high risk patients for inclusion in clinical trials of experimental treatments. The aim of the current analysis is to identify key human leukocyte antigens which are associated with increased risk of arthritis mutilans in this cohort of patients.

To date, 1191 patients have been recruited to the University of Toronto Psoriatic Arthritis Clinic, and 604 of these have undergone genetic testing to determine their human leukocyte antigen profile. A total of 96 human leukocyte antigen covariates were available for study but 20 of these markers had a frequency in the sample of less than 1% and so were excluded from further consideration. Among the 604 patients the median time from clinic entry to last radiological assessment is 6.3 years and there is a median of 3 radiological assessments per patient. The estimated cumulative distribution functions of the times between the first 10 radiological assessments are displayed in Figure 1 (a), which were obtained by fitting a semi-Markov model with an individual-specific gamma distributed frailty term (Klein, 1992) and stratified on the cumulative number of radiological assessments. The median inter-assessment times range from 2.7 years for the first two or three assessments after clinic

entry, to over 6 years for later assessments. Also plotted is a marginal Kaplan-Meier estimate of the time from psoriatic arthritis diagnosis to clinic entry.

Five hundred and seven (83.9%) of the 604 individuals in this dataset were not observed to develop arthritis mutilans and hence provided right-censored times, whereas 97 (16.1%) individuals yielded interval-censored times. Figure 1 (b) contains a nonparametric estimate (Turnbull, 1976) and pointwise 95% confidence bands for the cumulative distribution function of the time from onset of psoriatic arthritis to arthritis mutilans. The estimate reflects a steadily increasing risk with roughly 23% of psoriatic arthritis patients developing the condition within 20 years of disease onset.

2 VARIABLE SELECTION WITH INTERVAL-CENSORED DATA

2.1 NOTATION AND THE PENALIZED COMPLETE DATA LIKELIHOOD

We let T_i denote the time from disease onset to the event of interest for individual i in a sample of m independent individuals, $i = 1, \ldots, m$. We assume individuals are examined at assessment times governed by a conditionally independent inspection process (Grüger et al., 1991) and let $\mathcal{C}_i = [L_i, R_i)$ denote the interval known to contain the event for subject i, $i = 1, \ldots, m$. For left-censored data $L_i = 0$, for right censored data $R_i = \infty$, and for interval censored data $0 < L_i < R_i < \infty$. We let $X_i = (X_{i1}, \ldots, X_{ip})'$ denote a $p \times 1$ covariate vector.

Interest lies in the relation between the covariates and the time of interest based on a proportional hazards model with $h(t|X_i;\theta) = h_0(t;\alpha)\exp(X_i'\beta)$ where α parameterizes the baseline hazard, $\beta = (\beta_1, \ldots, \beta_p)'$, and $\theta = (\alpha', \beta')'$. We adopt a weakly parametric piecewise constant baseline hazard function which requires specification of the number and location of break-points, the times that the baseline hazard changes value. If $b_0 = 0$ and $0 < b_1 < \cdots < b_{K-1} < b_K = \infty$ denote K break-points, the baseline hazard function is $h_0(s;\alpha) = \rho_k = \exp(\alpha_k)$, for $s \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. The survivor function is then $\mathcal{F}(t|X_i;\theta) = \exp\{-H(t|X_i;\theta)\}$ where $H(t|X_i;\theta) = \int_0^t h(s|X_i;\theta)ds$. Given the covariate vector X_i and a conditionally independent inspection process, the observed (partial) likelihood is

$$\boldsymbol{L(\theta) \propto \prod_{i=1}^{m} \{\mathcal{F}(L_i|X_i;\theta) - \mathcal{F}(R_i|X_i;\theta)\}}$$

and the corresponding observed data log-likelihood is

$$\boldsymbol{\log L(\theta) \propto \sum_{i=1}^{m} \log\{\mathcal{F}(L_i|X_i;\theta) - \mathcal{F}(R_i|X_i;\theta)\}}. \tag{1}$$

When viewing this as a variable selection problem, we are specifically interested in identifying the covariates for which the regression coefficients are non-zero. Many common methods of variable selection are based on a penalized likelihood of the form

$$\boldsymbol{\log L_{\text{PEN}}(\theta) = \frac{1}{m}\log L(\theta) - p_{\gamma,\lambda}(\beta)}, \tag{2}$$

where the function $p_{\gamma,\lambda}(\beta)$ determines the extent of the penalty for each value of β, modulated by the tuning parameters (γ, λ). Ridge regression (Hoerl and Kennard, 1970) is implemented with the L_2 penalty $p_{\gamma,\lambda}(\beta) = \lambda\sum_{j=1}^{p}\beta_j^2$ and the LASSO (Tibshirani, 1996) uses the L_1 penalty $p_{\gamma,\lambda}(\beta) = \lambda\sum_{j=1}^{p}|\beta_j|$; there is no tuning parameter γ in these penalty functions. The value of the scalar λ is typically found by cross-validation (Shao, 1993) or generalized cross-validation (Golub et al., 1979). The adaptive LASSO uses adaptively weighted L_1 penalties of the form

$$\boldsymbol{p_{\gamma,\lambda}(\beta) = \sum_{j=1}^{p}\lambda_j|\beta_j|}, \tag{3}$$

with small penalties λ_j chosen for large coefficients to reduce their shrinkage, and large penalties for small coefficients to address the selection objective (Zou, 2006). One option is to set $\lambda_j = \lambda/|\widetilde{\beta}_j|$, where $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_p)'$ is the maximum likelihood estimate (Zou, 2006, Zhang and Lu, 2007). Alternatively, the penalties can be updated iteratively. In this case, at the $(\ell+1)$st implementation, λ_j is set to $\lambda_j^{(\ell)} = \lambda/|\widetilde{\beta}_j^{(\ell)}|$ where $\widetilde{\beta}^{(\ell)}$ is obtained on the ℓth iteration; when $\ell = 0$, we set $\lambda_j^{(0)} = \lambda/|\widetilde{\beta}_j|$ as in the first implementation (Fan and Lv, 2010). We investigate the iterative implementation of the adaptive LASSO in the next section.

The smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) is defined by

$$p'_{\gamma,\lambda}(\beta) = \lambda \sum_{j=1}^{p} \left\{ I(|\beta_j| \leq \lambda) + \frac{(\gamma\lambda - |\beta_j|)_+}{(\gamma-1)\lambda} I(|\beta_j| > \lambda) \right\} ,$$

where $\gamma > 2$ and $y_+ = I(y \geq 0) \times y$. This penalty function is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but singular at 0 with its derivatives zero outside the range $[-\gamma\lambda, \gamma\lambda]$. Therefore, the SCAD penalty results in "small" coefficients being set to zero, "moderate" coefficients being shrunk towards zero, and "large" coefficients retained as they are. In principle, the optimal pair (γ, λ) could be obtained using a two dimensional grid search by cross validation or generalized cross validation. Empirical studies by Fan and Li (2001) suggest $\gamma = 3.7$ is a reasonable choice for a variety of problems so we use this in what follows and select λ by (generalized) cross validation.

2.2 An Expectation-Maximization Algorithm

Let $D_k(u) = I(u \in \mathcal{B}_k)$ denote whether or not the time u is in the interval $\mathcal{B}_k$ and $W_k(u) = \int_0^u I_k(s)ds$ denote the duration at risk in interval k over $[0, u)$. If the event time t_i is known, then under the piecewise constant model of Section 2.1 and given a covariate vector X_i, the complete data log-likelihood $\log L_{\text{COMP}}(\theta)$ is

$$\sum_{i=1}^{m} \sum_{k=1}^{K} \{ D_k(t_i)(\alpha_k + X_i'\beta) - W_k(t_i) \exp(\alpha_k + X_i'\beta) \} . \tag{4}$$

Let $Z_{ik\ell} = I(k = \ell)$ indicate $k = \ell$, $\ell = 1, \ldots, K$ and $Z_{ik} = (Z_{ik1}, \ldots, Z_{ikK})'$ denote the corresponding vector of indicator functions, $k = 1, \ldots, K$; thus $Z_{i1} = (1, 0, \ldots, 0)'$, $Z_{i2} = (0, 1, \ldots, 0)'$, …, $Z_{ik} = (0, 0, \ldots, 1)'$. If $\alpha = (\alpha_1, \ldots, \alpha_K)'$ we can write

$$\boldsymbol{\log L_{\text{COMP}}(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \{ D_k(t_i) V_{ik}'\theta - W_k(t_i) \exp(V_{ik}'\theta) \} .} \tag{5}$$

where $V_{ik} = (Z_{ik}', X_i')'$ and $\theta = (\alpha', \beta')'$. Since the penalty in (2) is simply a function of the regression parameters, maximization of the penalized likelihood (2) can be achieved by applying the EM algorithm to the penalized complete data likelihood

$$\boldsymbol{\frac{1}{m} \log L_{\text{COMP}}(\theta) - p_{\gamma,\lambda}(\beta) .} \tag{6}$$

The E-step

We let $D_i = (L_i, R_i, X_i)$ represent the observed data from individual i and $D = \{D_i, i = 1, \ldots, m\}$ denote the observed data for the full sample. The conditional expectation of (6) at the $(r+1)$st iteration is evaluated as

$$\boldsymbol{Q_{\text{PEN}}(\theta; \theta^{(r)}) = E \left\{ \log L_{\text{COMP}}(\theta) | D; \theta^{(r)} \right\} - p_{\gamma,\lambda}(\beta) ,} \tag{7}$$

where $\theta^{(r)}$ is the estimate obtained from the rth iteration. The required conditional expectations are therefore $\widehat{\Delta}_{ik}^{(r)} = E[D_k(T_i)|D_i;\theta^{(r)}]$ and $\widehat{\omega}_{ik}^{(r)} = E[W_k(T_i)|D_i;\theta^{(r)}]$.

Let $\mathcal{C}_{ik} = \mathcal{C}_i \cap \mathcal{B}_k = [L_{ik}, R_{ik})$ denote the sub-interval of the censoring interval $\mathcal{C}_i$ contained within $\mathcal{B}_k$. When $\mathcal{C}_{ik} = \emptyset$, the required expectations are relatively easy to compute since, for instance, it is clear that $D_k(t_i) = 0$ and $\widehat{\Delta}_{ik}^{(r)} = 0$. Moreover, if $b_k < L_i$, then it is known that individual i was at risk for the entire interval $\mathcal{B}_k$ so $W_k(t_i) = \widehat{\omega}_{ik}^{(r)} = b_k - b_{k-1}$, and if $R_i < b_{k-1}$, then $W_k(t_i) = \widehat{\omega}_{ik}^{(r)} = 0$ since they are known to have failed prior to the start of interval $\mathcal{B}_k$. If $\mathcal{C}_{ik} \neq \emptyset$,

$$\widehat{\Delta}_{ik}^{(r)} = \frac{\mathcal{F}(L_{ik}|X_i;\theta^{(r)}) - \mathcal{F}(R_{ik}|X_i;\theta^{(r)})}{\mathcal{F}(L_i|X_i;\theta^{(r)}) - \mathcal{F}(R_i|X_i;\theta^{(r)})} \tag{8}$$

$$\widehat{\omega}_{ik}^{(r)} = \max(L_i - b_{k-1}, 0) + \int_{\max(L_i, b_{k-1})}^{\min(R_i, b_k)} \frac{\mathcal{F}(s|X_i;\theta^{(r)})}{\mathcal{F}(L_i|X_i;\theta^{(r)}) - \mathcal{F}(R_i|X_i;\theta^{(r)})} \mathrm{d}s \, . \tag{9}$$

Given these results, (7) can be written more explicitly as

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left\{ \widehat{\boldsymbol{\Delta}}_{\boldsymbol{ik}}^{(\boldsymbol{r})} \boldsymbol{V}_{\boldsymbol{ik}}' \boldsymbol{\theta} - \widehat{\boldsymbol{\omega}}_{\boldsymbol{ik}}^{(\boldsymbol{r})} \exp(\boldsymbol{V}_{\boldsymbol{ik}}' \boldsymbol{\theta}) \right\} - \boldsymbol{p}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{\beta}) \, . \tag{10}$$

THE M-STEP

The objective function (10) has the form of a penalized Poisson likelihood.The value $\theta^{(r+1)}$ that maximizes (10) can therefore be obtained using software for penalized Poisson regression by creating a dataset comprised of pseudo-individuals indexed by (i,k). If $R_i \geq b_{k-1}$, then at the $(r+1)$st iteration this dataset should include a contribution from pseudo-individual (i,k) with pseudo-count $\widehat{\Delta}_{ik}^{(r)}$ and offset $\log \widehat{\omega}_{ik}^{(r)}$; if $R_i < b_{k-1}$ then no such contribution is required. The function $Q_{\text{PEN}}(\theta;\theta^{(r)})$ is then maximized with respect to θ using standard software for penalized Poisson regression (e.g. the `glmnet(.)` function (R Core Team, 2013, Friedman et al., 2010) or `SIS(.)` (Fan et al., 2010)).

This optimization procedure is repeated iteratively with updated values of (8) and (9) in (10) until the difference between successive estimates becomes small enough to satisfy the convergence criterion. In our implementation the iterations were terminated when $\max_j(|\theta_j^{(r+1)} - \theta_j^{(r)}|/|\theta_j^{(r)}|) < \epsilon$, where $\epsilon = 10^{-6}$.

SELECTION OF THE OPTIMAL TUNING PARAMETER λ_{OPT}

The criterion for selecting the optimal λ is similar to traditional cross validation. Here we use G-fold cross validation and so partition the dataset into G subsamples $\mathcal{S}_1, \ldots, \mathcal{S}_G$; we refer to $\mathcal{S}_g$ and $\mathcal{S} - \mathcal{S}_g$ as the gth test and training sets, $g = 1, \ldots, G$. For the SCAD penalty we fixed $\gamma = 3.7$. For a given λ, the *cross-validation statistic* is

$$\widehat{CV}(\lambda) = \sum_{g=1}^{G} \left\{ \log L(\widehat{\theta}_{-g}(\lambda)) - \log L_{-g}(\widehat{\theta}_{-g}(\lambda)) \right\} \, . \tag{11}$$

where L_{-g} is the observed data likelihood (1) for the gth training dataset and $\widehat{\theta}_{-g}(\lambda)$ is the estimate for the gth training data, obtained through the penalized EM algorithm. The optimal λ maximizes $\widehat{CV}(\lambda)$.

Simulation studies reported in *Supplementary Material: Web Appendix B* assess the relative performance of cross-validation, use of the Bayesian information criterion, and the sparse generalized cross-validation (Bradic et al., 2011). While it is difficult to make general statements, the different penalty functions yielded good performance under cross-validation (i.e. good sensitivity for picking up important factors) and small mean squared error (MSE) of the β parameter estimates, with a

slightly higher tendency to claim association when there is none. Since there is often strong interest in identifying important variables for further study, it is reasonable to place high importance on the sensitivity and MSE criteria and so we adopt the standard cross-validation approach to selection of the tuning parameter in the following empirical studies; this statistic is also used in the R package `glmnet`.

3 Design and Interpretation of Simulation Studies

In this section, we report on the results of simulation studies designed to assess the performance of the penalized EM algorithm for variable selection with interval-censored data. We consider a sample size of $m = 500$ to correspond roughly to the size of the sample in the psoriatic arthritis study. In the first setting, $p = 100$ and $X_i \sim \text{MVN}_p(0, \Sigma)$ are i.i.d. where the (j, k) element of Σ is $\Sigma_{jk} = \rho^{|j-k|}$, with $\rho = 0.5$ to represent a strong autoregressive dependence, $i = 1, 2, \ldots, m$. The conditional hazard for T_i is based on a Weibull regression model where $h(t|X_i; \theta) = \kappa\eta(\eta t)^{\kappa-1}\exp(X_i'\beta)$. We set $\beta_j = 0.5$ for $j = 1, \ldots, 5$ and $j = 96, \ldots, 100$, so that high values of $X_{i,1}, \ldots, X_{i,5}, X_{i,96}, \ldots, X_{i,100}$ are associated with shorter times to the event, and $\beta_j = 0$, $j = 6, \ldots, 95$ so that $T_i \perp (X_{i,6}, \ldots, X_{i,95})|X_{i,1}, \ldots, X_{i,5}, X_{i,96}, \ldots, X_{i,100}$. The elements of X_i with non-zero coefficients were chosen to give both weak and strong dependence within the set of important covariates.

We consider a study with follow-up planned over $[0, 1]$, where for each of $\kappa = 1.0$ and 1.25, we solve for η so that $P(T_i < 1|X_i = 0; \theta) = 0.95$. We let N_i denote the number of assessments for individual i, generated according to a Poisson distribution with mean μ, truncated to ensure at least one follow-up assessment, given by

$$P(N_i = n|N_i \geq 1; \mu) = \frac{\mu^n \exp(-\mu)}{n!\{1 - \exp(-\mu)\}}, n = 1, \ldots .$$

The n_i inspection times $0 < a_{i1} < \cdots < a_{in_i} < 1$ were then generated uniformly over $[0, 1]$. The left and right endpoints of the censoring interval for individual i are then $L_i = \max(a_{ij} \cdot I(a_{ij} < t_i))$ and $R_i = \min(a_{ij} \cdot I(a_{ij} > t_i))$ respectively. One hundred datasets were then simulated ($nsim = 100$) for $\mu = 10$ and 20 respectively.

For each dataset, variable selection was carried out based on the penalized EM (P-EM) algorithm of Section 2.2 with the LASSO, adaptive LASSO (ALASSO) and SCAD penalty ($\gamma = 3.7$). For each analysis, 5-fold cross validation was carried out to select the unknown tuning parameter. Analyses were conducted based on proportional hazards models with a piecewise constant baseline hazards; hazard functions with four pieces (PWC-4) where the break-points were located at the quartiles of the baseline survival function. For comparison with a simple alternative approach, datasets were created by an *ad hoc* mid-point imputation approach (Lindsey and Ryan, 1998) in which event times for individuals with $R_i < \infty$ were take to be $t_i^* = (L_i + R_i)/2$. The resulting datasets were analysed based on the proportional hazards assumption with piecewise constant baseline hazards with the same break-points as used in the P-EM analyses; the corresponding results are labeled MID. The more traditional methods of variable selection based on forward selection and backward elimination were also considered under the true parametric Weibull regression model where we used $p = 0.10$ for inclusion or removal of terms; the R function `survreg` (R Core Team, 2013, Therneau, 2013) was used in this case as it handles parametric modeling with interval-censored data.

The number of variables selected was recorded. Among those that were truly associated with the response, the average number selected across all simulated datasets is reported as the mean number of true positive (TP) selections; the correct number of non-zero coefficients is given in parentheses in the column headings as TP(10). Among the covariates having no (conditional) association with the event time, the number selected for each dataset was averaged and reported as the mean number of

Table 1: Empirical results for interval-censored data with normally distributed covariates ($p = 100$, $E(X_{ij}) = 0$, $\mathrm{var}(X_{ij}) = 1$ and $\mathrm{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, where $\rho = 0.5$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation.

		$\mu = 10$			$\mu = 20$		
Method		TP(10)	FP(90)	MSE (SD)	TP(10)	FP(90)	MSE (SD)
		Shape parameter: $\kappa = 1$					
LASSO	P-EM	10.00	14.80	0.312 (0.126)	10.00	14.83	0.261 (0.105)
	MID	10.00	13.05	1.346 (0.286)	10.00	12.05	0.912 (0.251)
ALASSO	P-EM	10.00	0.12	0.057 (0.047)	10.00	0.07	0.047 (0.040)
	MID	9.69	0.30	0.953 (0.328)	10.00	1.57	0.499 (0.201)
SCAD	P-EM	9.98	0.36	0.059 (0.073)	9.99	0.24	0.050 (0.048)
	MID	9.39	0.96	0.946 (0.354)	9.91	1.01	0.521 (0.213)
FORWARD		10.00	9.17	0.218 (0.088)	10.00	9.50	0.201 (0.082)
BACKWARD		10.00	15.35	0.322 (0.130)	10.00	14.80	0.289 (0.099)
		Shape parameter: $\kappa = 1.25$					
LASSO	P-EM	10.00	14.88	0.291 (0.118)	10.00	14.13	0.245 (0.109)
	MID	10.00	15.28	1.037 (0.271)	10.00	12.94	0.685 (0.216)
ALASSO	P-EM	9.99	0.23	0.055 (0.050)	10.00	0.08	0.045 (0.031)
	MID	9.75	0.29	0.724 (0.327)	10.00	1.25	0.314 (0.160)
SCAD	P-EM	9.98	0.29	0.055 (0.052)	9.99	0.13	0.044 (0.036)
	MID	9.53	0.76	0.741 (0.336)	9.97	0.91	0.317 (0.167)
FORWARD		10.00	8.66	0.324 (0.089)	10.00	8.81	0.313 (0.089)
BACKWARD		10.00	14.35	0.383 (0.092)	10.00	14.17	0.363 (0.092)

false positive (FP) selections; the number of truly independent covariates is given in parentheses as FP(90). These statistics, along with the mean squared error (MSE = $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$), and the empirical standard errors of the mean square error, are reported in Table 1 based on 100 simulations.

All three penalty functions generally led to selection of the ten covariates associated with the response for the P-EM and mid-point implementations, with slightly worse performance of the ALASSO and SCAD penalty functions following mid-point imputation. The ALASSO and SCAD penalty functions had the lowest FP values which were lower in the P-EM implementation than following mid-point imputation. For any particular penalty function the MSE and the respective standard deviation were always lower when the penalized EM algorithm was used rather than mid-point imputation. These findings point to the advantages of the proposed method which include slightly lower FP values and substantially lower MSE. The forward and backward selection algorithms also featured high FP values. There were little differences between the findings with the exponential ($\kappa = 1$) and Weibull ($\kappa = 1.25$) regression models.

In a second simulation study, we considered correlated binary covariates with $p = 100$ to more closely represent the dimension of the HLA variables in the psoriatic arthritis study. We set $P(X_{ij} = 1) = 0.20$, $j = 1, \ldots, 100$. For the dependence structure we considered the covariates as arising in ten independent blocks such that the correlation between covariates X_{ij} and X_{ik} within the same block is $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$ with $\rho = 0.2$. Ten covariates were specified to have coefficients equal to one such that the pairwise dependencies among them ranged from weak to strong; all others covariate effects were set to zero. The results displayed in Table 2 again demonstrate that all methods tend to select the covariates with the non-zero coefficients on average, although the methods based on the adaptive LASSO and SCAD penalties have negligibly lower TP values. As in the previous simulations, the false positive selection rate is lower with the adaptive LASSO and SCAD penalty functions compared to the LASSO as well as the forward and backward selection algorithms. The respective mean squared errors are always substantially lower in the penalized EM algorithms compared to the respective implementation following mid-point imputation.

Figure 2 displays box plots of the errors in estimates (i.e. $\widehat{\beta}_k - \beta_k$) for four of the hundred coefficients in the setting with binary covariates, $\kappa = 1.25$, and $\mu = 20$; β_5 and β_{95} (both zero) and β_{22} and β_{96} (both 1.0). For each penalty function the estimates for the P-EM and mid-point imputation methods are displayed, along with estimates from an analysis using the true failure time subject only to administrative right censoring (RC) at $C = 1$; the latter analysis is only possible in a simulation study, but is presented for comparison purposes since it provides a natural benchmark for assessing the performance of the proposed algorithm for interval-censored data. It is important to note that different datasets are used for the P-EM, mid-point imputation and RC analyses, with only the former corresponding to the observed data.

In *Supplementary Material: Web Appendix A*, we present the results of further simulation studies with multivariate normal and correlated binary covariates when $p = 10$. Here we consider analyses with an exponential (time homogeneous) regression model and a piecewise constant baseline hazard (4 pieces) model. The former is included to examine the effect of having a more elaborate (four piece) baseline hazard when a single piece is sufficient as is the case when $\kappa = 1.0$, as well as the effect of gross misspecification of the baseline hazard when $\kappa = 1.25$. When $\kappa = 1.0$ and the P-EM algorithm is used, the PWC-4 model yields a very slightly higher MSE than was seen for the exponential model, but the results suggest there is little price to pay when the piecewise constant model is used unnecessarily.

When $\kappa = 1.25$, the piecewise constant model (PWC-4) had a slightly lower rate of false positive selections and a lower MSE than the exponential model. A similar study was conducted with binary covariates ($p = 10$) with findings that suggest that the adaptive LASSO and SCAD penalties again are again preferable to the LASSO since they generally lead to smaller MSE; among these two methods the relative performance tends to depend on the criteria used (TP, FP or MSE) but they appear broadly

Table 2: Empirical results for interval-censored data with correlated binary covariates ($p = 100$, $E(X_{ij}) = 0.2$ and $\text{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$ if X_{ij}, X_{ik} are in the same block as discussed in Section 3 and $\rho = 0.2$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-dataset obtained by mid-point imputation; the tuning parameter is selected by five-fold cross-validation.

		$\mu = 10$			$\mu = 20$		
Method		TP(10)	FP(90)	MSE (SD)	TP(10)	FP(90)	MSE (SD)
		Shape parameter: $\kappa = 1$					
LASSO	P-EM	10.00	12.49	0.304 (0.068)	10.00	15.30	0.201 (0.052)
	MID	10.00	17.64	0.690 (0.117)	10.00	19.01	0.436 (0.086)
ALASSO	P-EM	9.88	0.82	0.071 (0.067)	9.98	0.26	0.039 (0.033)
	MID	9.18	0.78	0.491 (0.149)	9.83	0.49	0.255 (0.097)
SCAD	P-EM	9.94	0.54	0.063 (0.063)	10.00	0.10	0.038 (0.031)
	MID	9.02	0.96	0.505 (0.166)	9.79	0.40	0.254 (0.102)
FORWARD		10.00	11.14	0.244 (0.078)	10.00	11.09	0.183 (0.057)
BACKWARD		10.00	15.18	0.299 (0.083)	10.00	14.64	0.231 (0.064)
		Shape parameter: $\kappa = 1.25$					
LASSO	P-EM	10.00	12.04	0.277 (0.064)	10.00	15.65	0.186 (0.053)
	MID	9.99	18.15	0.609 (0.100)	10.00	17.91	0.374 (0.074)
ALASSO	P-EM	9.98	0.59	0.051 (0.042)	10.00	0.22	0.034 (0.023)
	MID	9.59	0.60	0.404 (0.116)	9.97	0.26	0.186 (0.064)
SCAD	P-EM	10.00	0.48	0.053 (0.038)	10.00	0.16	0.033 (0.021)
	MID	9.54	0.93	0.414 (0.118)	9.95	0.42	0.186 (0.064)
FORWARD		10.00	10.86	0.198 (0.060)	10.00	10.81	0.180 (0.045)
BACKWARD		10.00	14.49	0.233 (0.064)	10.00	13.76	0.195 (0.052)

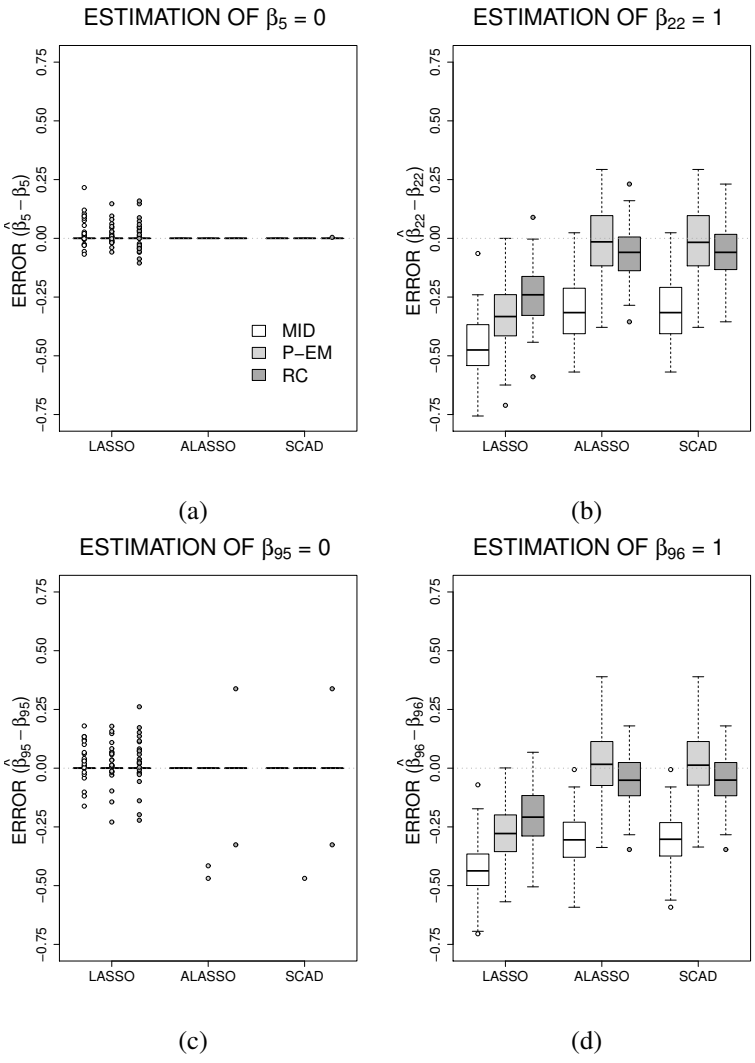


Figure 2: Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 5, 22, 95, 96$, for each penalty function for datasets with correlated binary covariates ($p = 100$) with $\kappa = 1.25$, $\mu = 20$.

comparable overall.

4 HLA Markers and Risk of Arthritis Mutilans

Interest lies in identifying which among the 76 human leukocyte antigen markers are associated with increased risk of developing arthritis mutilans from the time of diagnosis with psoriatic arthritis. The first, second and third quartiles for the length of the closed censoring intervals for the 97 individuals known to have developed arthritis mutilans were 2.50, 8.06 and 15.00 years respectively. These quantiles are much wider than one might expect from a protocol in which radiological assessments are to be scheduled every two years because of the variation between individuals in the propensity to attend the clinic, as well as the potentially long delay from the onset of psoriatic arthritis to clinic entry; see Figure 1 (a). We also remark that the proportion of individuals generating interval-censored times to arthritis mutilans is smaller than that represented in the simulation study, and that the variability in the width of the censoring intervals is considerable; the P-EM algorithm can accommodate this setting.

All models considered control for gender, age at onset of PsA, family history of psoriasis (yes/no), and family history of psoriatic arthritis (yes/no). We report here on the results of applying the penalized EM algorithm using the LASSO, adaptive LASSO and SCAD penalty functions. For comparison purposes, results are also reported for a right-censored dataset obtained by using mid-point imputation (MID) as examined in the simulation studies. Given the findings from the simulation studies, however, we restrict our attention primarily to the results form the penalized EM procedure. The standard errors of the estimates are calculated using the bootstrap (Efron and Tibshirani, 1994); details are given in *Supplementary Material: Web Appendix C*.

The break-points for the piecewise constant hazard functions were chosen based on the nonparametric estimate of the marginal cumulative probability distribution function for the time from disease onset to arthritis mutilans; see Figure 1 (b). The cumulative probability is about 35% over 28 years so the break-points chosen were 6.5, 10.5, 18, and 22 years corresponding to the cumulative probabilities of 7%, 14%, 21% and 28%.

The union of all HLA variables selected by any method are listed in Table 3, where it can be seen that the SCAD penalty function with the P-EM procedure selected the fewest HLA markers including HLA-A11, HLA-A29, HLA-B27 and HLA-DQB1-02; HLA-B27 and HLA-DQB1-02 are two factors well known to incur increased risk of joint damage and we found that the presence of HLA-A11 and HLA-A29 has a protective effect. Under the P-EM algorithm the LASSO penalty function also selected HLA-C04, and the corresponding implementation of the ALASSO further selected HLA-A25, HLA-A30 and HLA-DRB1-10. With the ALASSO penalty the same variables were selected whether the P-EM or mid-point imputation was used. For the other penalty functions more variables were selected under mid-point imputation than with the P-EM procedure, as found in the empirical investigations. The findings are in broad agreement with those from recent analyses (Chandran et al., 2012) and a validation exercise is currently underway involving three independent cohorts from Spain, Ireland and Newfoundland, Canada. The empirical correlations among the union set of all variables selected by any method range from -0.105 to 0.198.

The top row of Figure 3 contain plots of the cross-validation statistic to reveal how the optimal values of the tuning parameters are found for the LASSO, adaptive ALASSO and SCAD functions; The plots in the bottom row of Figure 3 give the profile plots of the coefficients, showing the degree of shrinkage and selection of covariates as a function of the tuning parameter. The stage at which each variable is selected conveys the relative importance of the covariates; the optimal value of the tuning parameter is designated by the vertical dotted lines.

Table 3: HLA markers selected following variable selection with LASSO, ALASSO or SCAD penalty in analysis of interval-censored progression data in psoriatic arthritis.

	LASSO				ALASSO				SCAD			
	P-EM		MID		P-EM		MID		P-EM		MID	
HLA Marker	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)	β	s.e.(β)
HLA-A11	-0.135	0.199	-0.280	0.263	-0.516	0.629	-0.556	0.836	-1.021	0.746	-0.922	0.947
HLA-A25			-0.232	0.288	-3.265	0.707	-3.229	1.529				
HLA-A29	-0.216	0.254	-0.502	0.353	-1.388	1.284	-1.385	1.440	-1.605	2.376	-1.658	2.482
HLA-A30			0.101	0.260	0.494	0.417	0.494	0.525				
HLA-B27	0.249	0.232	0.397	0.272	0.588	0.356	0.595	0.547	0.763	0.312	0.725	0.425
HLA-C04	-0.012	0.134	-0.170	0.233	-0.578	0.492	-0.569	1.086			-0.637	0.611
HLA-DQB1-02	0.134	0.164	0.270	0.205	0.514	0.307	0.503	0.540	0.609	0.276	0.623	0.415
HLA-DRB1-10					-2.713	1.007	-2.714	1.725				

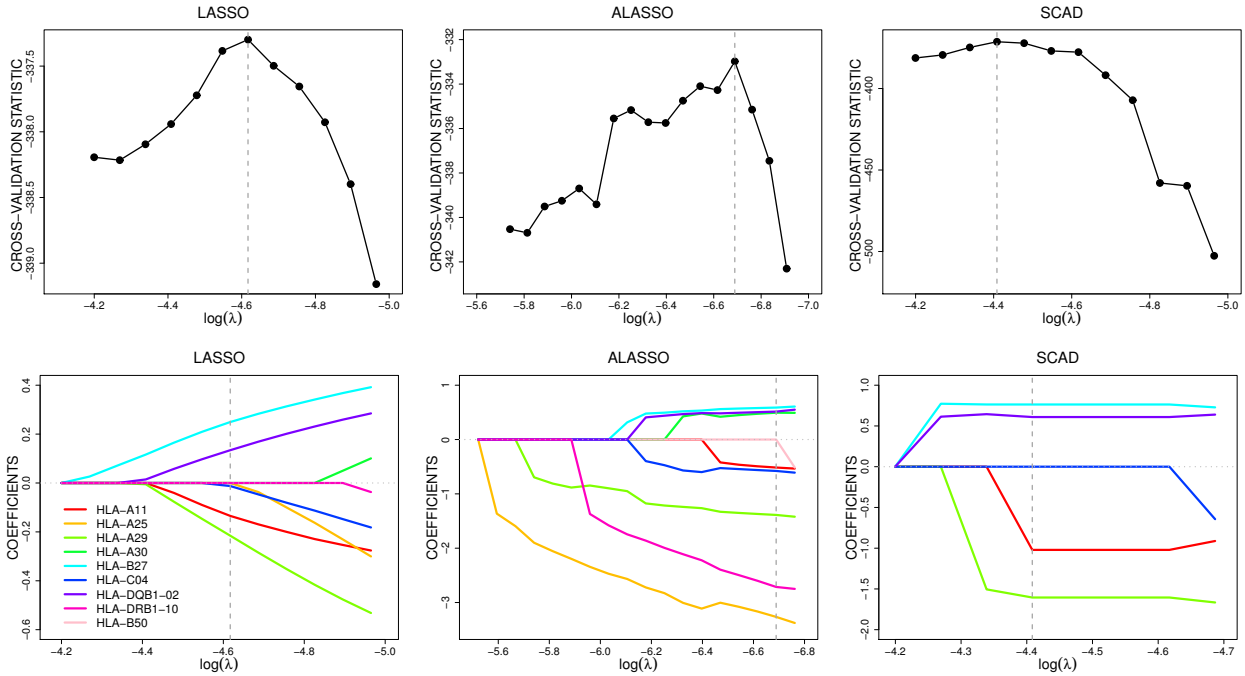


Figure 3: Plots of the cross-validation statistics (top row) and shrinkage estimates of coefficients (bottom row) from penalized regression of the PsA dataset based on a piecewise constant hazard model (PWC-5) fitted via an EM algorithm with the LASSO, ALASSO or SCAD penalty.

5 DISCUSSION

In this paper we have proposed a simple adaptation of existing algorithms for variable selection to deal with interval-censored failure time data. A complete data log-likelihood formed based on a proportional hazards model with a piecewise constant baseline hazard is augmented by including one of several possible penalty terms. The simulation studies showed that the proposed algorithm led to better performance for each penalty function compared to ad hoc methods using mid-point imputation. We experienced no convergence problems with the penalized expectation-maximization algorithm; Wu (1983) should help assess whether this can be relied upon generally. The adaptive LASSO, as implemented here with iteratively updated weights, had the best performance. The relative performance of the different penalty functions depended heavily on the method for selecting the optimal tuning parameter in the penalty functions. It can be seen in Table B.2 of the *Supplementary Material: Web Appendix B*, for example, that the performance of the LASSO in terms of FP was much better when tuning parameter λ was chosen by BIC or SGCV. The purpose of this article is not to carry out an exhaustive study of variable selection techniques based on the different penalty functions, but further study of the various options for choosing the tuning parameters seems worthwhile.

An application to the PsA data was conducted, and the results of these analyses agree quite well with previous analysis. Lockhart et al. (2014) point out the properties of coefficients obtained following variable selection are not well understood. In *Supplementary Material: Web Appendix C* we explore techniques for variance estimation following variable selection, but we rely on bootstrap standard errors in the application.

The piecewise exponential model is a simple, flexible and weakly parametric approach to dealing with interval-censored data. We set $K = 4$, following the observation of Lawless and Zhan (1998) that a modest number of pieces is usually sufficient, particularly when inferences about covariate effects are of greatest interest. More flexible semiparametric methods could be considered in this setting, including methods based on local likelihood (Betensky et al. 2002) or penalized splines (Cai and Betensky, 2003). These, and other semiparametric methods, may offer a more suitable framework for studying the limiting behaviour of these algorithms and the resultant estimators.

6 SUPPLEMENTARY MATERIAL

Web Appendices referenced in Sections 2 (Web Appendix B), 3 (Web Appendix A) and 4 (Web Appendix C) are available with this paper at the Biometrics website on Wiley Online Library. The R code implementing the proposed penalized expectation-maximization (P-EM) algorithm and an example dataset discussed in Section 3 are available with this paper at the Biometrics website on Wiley Online Library.

ACKNOWLEDGEMENTS

The authors thank Dr. Dafna Gladman, Dr. Vinod Chandran and Dr. Lihi Eder for stimulating collaboration and helpful discussions involving the psoriatic arthritis research program. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

REFERENCES

Altman, D. G., McShane, L. M., Sauerbrei, W., and Taube, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Medicine*,

10(1):51.

Betensky, R. A. and Finkelstein, D. M. (1999). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18(22):3089–3100.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21:263–275.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Coxs proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39(6):3092–3120.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized splines. *Biometrics*, 59(3):570–579.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., and Gladman, D. D. (2010). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7):1399–1405.

Chandran, V., Cook, R. J., Thavaneswaran, A., Lee, K.-A., Pellett, F., and Gladman, D. (2012). Parametric survival analysis as well as multi-state analysis confirms the association between human leukocyte antigen alleles and the development of arthritis mutilans in patients with psoriatic arthritis. *Journal of Rheumatology*, 39(8):1723–1723.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):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 (Methodological)*, 39(1):1–38.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*, volume 57. CRC press.

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). *SIS: Sure Independence Screening*. R package version 0.6.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *Annals of Statistics*, 10:101–113.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47:595–605.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hortobagyi, G. N., Theriault, R. L., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J., and Knight, R. D. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335(24):1785–1792.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics*, 48(3):795–806.

Kyle, S., Chandler, D., Griffiths, C. E. M., Helliwell, P., Lewis, J., McInnes, I., Oliver, S., Symmons, D., and McHugh, N. (2005). Guideline for anti-TNF-α therapy in psoriatic arthritis. *Rheumatology*, 44(3):390–397.

Lawless, J. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565.

Li, J. and Ma, S. (2013). *Survival Analysis in Medicine and Genetics*. CRC Press.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine*, 17:219–238.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the LASSO. *Annals of Statistics*, 42(2):413–468.

McShane, L. M., Altman, D. G., and Sauerbrei, W. (2005a). Identification of clinically useful cancer prognostic factors: what are we missing? *Journal of the National Cancer Institute*, 97(14):1023–1025.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M., et al. (2005b). Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute*, 97(16):1180–1184.

Polley, M.-Y. C., Freidlin, B., Korn, E. L., Conley, B. A., Abrams, J. S., and McShane, L. M. (2013). Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*, 105(22):1677–1683.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, P., Gladman, D. D., Cook, R. J., Zhou, Y., Young, G., and Salonen, D. (1998). Radiological assessment in psoriatic arthritis. *Rheumatology*, 37(7):760–765.

Riggs, B. L., Hodgson, S. F., O'Fallon, W. M., Chao, E. Y., Wahner, H. W., Muhs, J. M., Cedel, S. L., and Melon, L. J. (1990). Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *New England Journal of Medicine*, 322(12):802–809.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.

Therneau, T. M. (2013). A package for survival analysis in S. R package version 2.37-4.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Witten, D. M. and Tibshirani, R. (2009). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19:29–51.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *Annals of statistics*, 11(1):95–103.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533.

Published in final edited form as:
Biometrics (2015), **71**(3): 782–791
DOI: 10.1111/biom.12302

Web-based Supplementary Materials for *Penalized Regression for Interval-Censored Times of Disease Progression: Selection of HLA Markers in Psoriatic Arthritis*

YING WU

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: rjcook@uwaterloo.ca

WEB APPENDIX A: SUPPLEMENTARY SIMULATION STUDIES

Here we conduct simulation studies with a relatively small number of covariates ($p = 10$). The generating procedure for the normally distributed covariates is the same as described in Section 3 with $p = 100$ multivariate normal covariates. We set $\beta_j = 0.5$, $j = 1, 2, 9, 10$ and $\beta_j = 0$, $j = 3, \ldots, 6$. In Table A.1, we report the results of applying the penalized expectation-maximization algorithm (P-EM) to proportional hazards models with exponential (EXP) and piecewise constant baseline hazards with four pieces (PWC-4). We also report corresponding results following mid-point imputation when the resulting data are treated as right-censored (MID). Traditional methods of variable selection based on forward selection and backward elimination are also considered based on the correct parametric Weibull regression model.

For the case of correlated binary covariates, the data are generated using a series of conditional binary probability mass functions as described by Preisser et al. [3]. We set the marginal probabilities such that $E(X_{ij}) = 0.05$, $j = 1, \ldots, 5$ and $E(X_{ij}) = 0.20$, $j = 6, \ldots, 10$, using a 10×10 correlation matrix with entry $\mathrm{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, where $\rho = 0.3$ or 0.6. The coefficients in the proportional hazards model are set to $\beta_j = 1$ for $j = 1, 2, 9, 10$ and $\beta_j = 0$, $j = 3, \ldots, 6$. The analyses are the same as those used for the multivariate normal covariates; and Table A.2 shows the results which is analogous to Table A.1.

When comparing the results between the midpoint imputed and interval-censored datasets with the PWC-4 model in Table A.1 and Table A.2, there is generally a comparable ability to detecting important covariates (TP) and number of false positive (FP) selections, but the proposed P-EM algorithm leads to lower MSE. The results from traditional variable selection methods also feature high mean squared errors and slightly higher FP values.

Figure A.1 contains box plots of the empirical estimation errors ($\widehat{\beta}_j - \beta_j$) for four of the ten coefficients (β_1 and β_2 (both equal to 1) and β_3 and β_5 (both equal to zero)) when data are simulated with $\kappa = 1.25$, $\mu = 10$ and $\rho = 0.3$. We report on results for an exponential and piecewise constant baseline hazard, for datasets featuring by mid-point imputation (MID), interval-censoring (P-EM), and for the case where the actual event time is used, subject only to right-censoring (RC). The performance of the piecewise constant model is generally better than the exponential model since $\kappa \neq 1$, and for

this hazard function, the P-EM algorithm leads to performance which is more like the analysis using the right-censored (RC) failure time; the latter analysis is only possible in a simulation study such as this where the interval-censored time is actually known.

			$\rho = 0.3$						$\rho = 0.6$					
			$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Model	Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
Shape parameter: $\kappa = 1$														
EXP	LASSO	P-EM	4.00	3.08	0.021 (0.012)	4.00	2.51	0.019 (0.013)	4.00	2.15	0.020 (0.014)	4.00	2.48	0.020 (0.012)
		MID	4.00	2.73	0.077 (0.029)	4.00	2.38	0.033 (0.018)	4.00	2.09	0.117 (0.037)	4.00	2.59	0.044 (0.029)
	ALASSO	P-EM	4.00	0.52	0.012 (0.013)	4.00	0.30	0.012 (0.011)	4.00	0.45	0.014 (0.012)	4.00	0.68	0.012 (0.013)
		MID	4.00	0.45	0.048 (0.023)	4.00	0.31	0.018 (0.013)	4.00	0.53	0.084 (0.029)	4.00	0.66	0.029 (0.018)
	SCAD	P-EM	4.00	0.51	0.010 (0.013)	4.00	0.40	0.012 (0.012)	4.00	0.33	0.013 (0.011)	4.00	0.61	0.011 (0.012)
		MID	4.00	0.38	0.048 (0.023)	4.00	0.35	0.018 (0.012)	4.00	0.44	0.082 (0.029)	4.00	0.55	0.028 (0.019)
PWC-4	LASSO	P-EM	4.00	3.05	0.026 (0.017)	4.00	2.54	0.020 (0.015)	4.00	2.10	0.027 (0.018)	4.00	2.38	0.022 (0.015)
		MID	4.00	2.84	0.057 (0.030)	4.00	2.50	0.029 (0.019)	4.00	2.19	0.085 (0.038)	4.00	2.55	0.037 (0.024)
	ALASSO	P-EM	4.00	0.45	0.012 (0.012)	4.00	0.21	0.013 (0.014)	4.00	0.45	0.015 (0.013)	4.00	0.59	0.013 (0.014)
		MID	4.00	0.29	0.029 (0.021)	4.00	0.34	0.015 (0.013)	4.00	0.38	0.052 (0.029)	4.00	0.56	0.019 (0.017)
	SCAD	P-EM	4.00	0.45	0.012 (0.013)	4.00	0.31	0.014 (0.014)	4.00	0.34	0.015 (0.013)	4.00	0.56	0.012 (0.013)
		MID	4.00	0.34	0.029 (0.021)	4.00	0.46	0.015 (0.012)	4.00	0.38	0.052 (0.029)	4.00	0.59	0.020 (0.017)
FORWARD			4.00	0.57	0.014 (0.012)	4.00	0.46	0.017 (0.011)	4.00	0.45	0.017 (0.012)	4.00	0.47	0.014 (0.011)
BACKWARD			4.00	0.65	0.014 (0.012)	4.00	0.48	0.017 (0.011)	4.00	0.60	0.018 (0.012)	4.00	0.65	0.016 (0.011)
Shape parameter: $\kappa = 1.25$														
EXP	LASSO	P-EM	4.00	2.80	0.057 (0.022)	4.00	2.53	0.057 (0.023)	4.00	2.26	0.063 (0.026)	4.00	2.50	0.064 (0.022)
		MID	4.00	2.68	0.113 (0.034)	4.00	2.47	0.076 (0.025)	4.00	2.10	0.157 (0.039)	4.00	2.39	0.094 (0.029)
	ALASSO	P-EM	4.00	0.47	0.030 (0.017)	4.00	0.33	0.032 (0.017)	4.00	0.48	0.038 (0.021)	4.00	0.72	0.041 (0.018)
		MID	4.00	0.38	0.082 (0.028)	4.00	0.35	0.047 (0.020)	4.00	0.57	0.123 (0.034)	4.00	0.55	0.068 (0.023)
	SCAD	P-EM	4.00	0.59	0.030 (0.018)	4.00	0.32	0.032 (0.017)	4.00	0.35	0.039 (0.021)	4.00	0.52	0.040 (0.018)
		MID	4.00	0.60	0.082 (0.028)	4.00	0.42	0.049 (0.020)	4.00	0.47	0.123 (0.033)	4.00	0.53	0.067 (0.023)
PWC-4	LASSO	P-EM	4.00	2.94	0.025 (0.015)	4.00	2.52	0.021 (0.016)	4.00	2.26	0.023 (0.017)	4.00	2.45	0.022 (0.014)
		MID	4.00	3.04	0.043 (0.027)	4.00	2.78	0.028 (0.017)	4.00	2.27	0.066 (0.033)	4.00	2.54	0.031 (0.022)
	ALASSO	P-EM	4.00	0.42	0.010 (0.012)	4.00	0.27	0.012 (0.012)	4.00	0.44	0.015 (0.014)	4.00	0.58	0.011 (0.013)
		MID	4.00	0.46	0.022 (0.020)	4.00	0.30	0.014 (0.011)	4.00	0.29	0.038 (0.023)	4.00	0.53	0.017 (0.014)
	SCAD	P-EM	4.00	0.48	0.010 (0.012)	4.00	0.28	0.013 (0.012)	4.00	0.41	0.016 (0.013)	4.00	0.55	0.011 (0.013)
		MID	4.00	0.41	0.022 (0.020)	4.00	0.41	0.015 (0.011)	4.00	0.32	0.038 (0.023)	4.00	0.52	0.017 (0.015)
FORWARD			4.00	0.53	0.060 (0.022)	4.00	0.51	0.060 (0.022)	4.00	0.47	0.076 (0.028)	4.00	0.61	0.073 (0.024)
BACKWARD			4.00	0.69	0.061 (0.023)	4.00	0.51	0.060 (0.022)	4.00	0.62	0.078 (0.028)	4.00	0.72	0.072 (0.024)

Table A.1: Empirical results for interval-censored data with normally distributed covariates ($p = 10$, $E(X_{ij}) = 0$, $Var(X_{ij}) = 1$ and $\mathrm{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation.

			$\rho = 0.3$						$\rho = 0.6$					
			$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Model	Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
Shape parameter: $\kappa = 1$														
EXP	LASSO	P-EM	4.00	2.48	0.293 (0.230)	3.99	2.77	0.263 (0.225)	3.99	2.43	0.292 (0.227)	4.00	2.12	0.256 (0.176)
		MID	3.99	2.53	0.826 (0.353)	3.99	2.56	0.451 (0.217)	3.96	2.35	1.102 (0.481)	4.00	2.02	0.506 (0.276)
	ALASSO	P-EM	3.82	0.77	0.250 (0.486)	3.93	0.90	0.208 (0.313)	3.84	1.44	0.287 (0.300)	3.93	0.79	0.162 (0.409)
		MID	3.79	0.73	0.611 (0.451)	3.92	0.88	0.263 (0.300)	3.64	0.83	0.968 (0.573)	3.88	0.63	0.317 (0.276)
	SCAD	P-EM	3.84	0.62	0.200 (0.465)	3.91	0.79	0.217 (0.402)	3.71	1.16	0.337 (0.339)	3.89	0.58	0.217 (0.300)
		MID	3.78	0.45	0.573 (0.459)	3.94	0.77	0.261 (0.280)	3.61	0.79	0.961 (0.578)	3.84	0.58	0.331 (0.283)
PWC-4	LASSO	P-EM	4.00	2.34	0.314 (0.239)	4.00	2.47	0.262 (0.217)	3.99	2.21	0.305 (0.238)	4.00	1.91	0.265 (0.193)
		MID	4.00	2.43	0.602 (0.320)	3.99	2.48	0.363 (0.208)	3.97	2.35	0.844 (0.440)	4.00	1.97	0.381 (0.254)
	ALASSO	P-EM	3.79	0.65	0.276 (0.567)	3.92	0.82	0.223 (0.328)	3.80	1.13	0.309 (0.428)	3.88	0.73	0.215 (0.426)
		MID	3.72	0.48	0.343 (0.635)	3.92	0.61	0.218 (0.297)	3.60	0.84	0.729 (0.735)	3.88	0.49	0.220 (0.397)
	SCAD	P-EM	3.85	0.69	0.210 (0.481)	3.95	0.89	0.213 (0.301)	3.75	1.18	0.314 (0.444)	3.89	0.69	0.236 (0.310)
		MID	3.74	0.52	0.351 (0.599)	3.92	0.73	0.229 (0.295)	3.63	0.87	0.720 (0.638)	3.88	0.43	0.221 (0.275)
FORWARD			3.99	0.63	0.216 (0.318)	3.97	0.67	0.227 (0.272)	3.79	0.59	0.269 (0.335)	3.91	0.56	0.202 (0.339)
BACKWARD			3.99	0.64	0.216 (0.315)	3.97	0.69	0.227 (0.275)	3.79	0.78	0.275 (0.332)	3.91	0.81	0.223 (0.334)
Shape parameter: $\kappa = 1.25$														
EXP	LASSO	P-EM	4.00	2.36	0.544 (0.254)	4.00	2.70	0.443 (0.222)	3.98	2.34	0.536 (0.275)	3.99	2.04	0.508 (0.217)
		MID	4.00	2.32	0.994 (0.303)	4.00	2.56	0.613 (0.221)	3.96	2.17	1.303 (0.412)	3.99	2.19	0.771 (0.244)
	ALASSO	P-EM	3.89	0.77	0.296 (0.468)	3.93	0.77	0.252 (0.275)	3.82	1.19	0.383 (0.281)	3.90	0.74	0.312 (0.219)
		MID	3.91	0.73	0.721 (0.343)	3.97	0.93	0.423 (0.238)	3.76	0.82	1.036 (0.443)	3.92	0.72	0.553 (0.238)
	SCAD	P-EM	3.88	0.54	0.270 (0.426)	3.94	0.64	0.247 (0.265)	3.75	0.97	0.428 (0.299)	3.88	0.70	0.314 (0.244)
		MID	3.87	0.47	0.718 (0.317)	3.96	0.75	0.425 (0.243)	3.68	0.68	1.049 (0.385)	3.83	0.39	0.550 (0.291)
PWC-4	LASSO	P-EM	3.99	2.15	0.284 (0.240)	3.99	2.32	0.249 (0.197)	3.98	2.18	0.308 (0.222)	3.99	1.86	0.245 (0.184)
		MID	4.00	2.51	0.489 (0.251)	4.00	2.70	0.304 (0.181)	3.95	2.28	0.617 (0.383)	3.99	2.00	0.332 (0.217)
	ALASSO	P-EM	3.83	0.56	0.173 (0.561)	3.94	0.62	0.153 (0.307)	3.83	1.17	0.271 (0.293)	3.91	0.76	0.182 (0.267)
		MID	3.88	0.57	0.279 (0.317)	3.95	0.64	0.159 (0.267)	3.69	0.70	0.480 (0.568)	3.88	0.72	0.210 (0.247)
	SCAD	P-EM	3.84	0.55	0.165 (0.514)	3.95	0.74	0.156 (0.299)	3.81	1.19	0.282 (0.292)	3.88	0.69	0.148 (0.280)
		MID	3.85	0.62	0.288 (0.336)	3.94	0.57	0.159 (0.271)	3.70	0.66	0.480 (0.437)	3.88	0.63	0.189 (0.255)
FORWARD			3.96	0.61	0.326 (0.234)	3.98	0.70	0.303 (0.197)	3.80	0.61	0.372 (0.297)	3.92	0.59	0.345 (0.213)
BACKWARD			3.96	0.65	0.326 (0.233)	3.98	0.73	0.303 (0.198)	3.80	0.76	0.395 (0.289)	3.91	0.82	0.383 (0.233)

Table A.2: Empirical results for interval-censored data with correlated binary covariates ($p = 10$, $E(X_{ij}) = 0.2$ and $\mathrm{corr}(X_{ij}, X_{ik}) = \rho^{|j-k|}$) summarizing the number of correctly (TP) and incorrectly (FP) selected variables along with the median and the standard deviation (SD) of the mean squared error (MSE); P-EM denotes the analyses based on the proposed penalized EM method and MID denotes an analysis based on a pseudo-data set obtained by mid-point imputation; the tuning parameter is selected by five-fold cross validation.

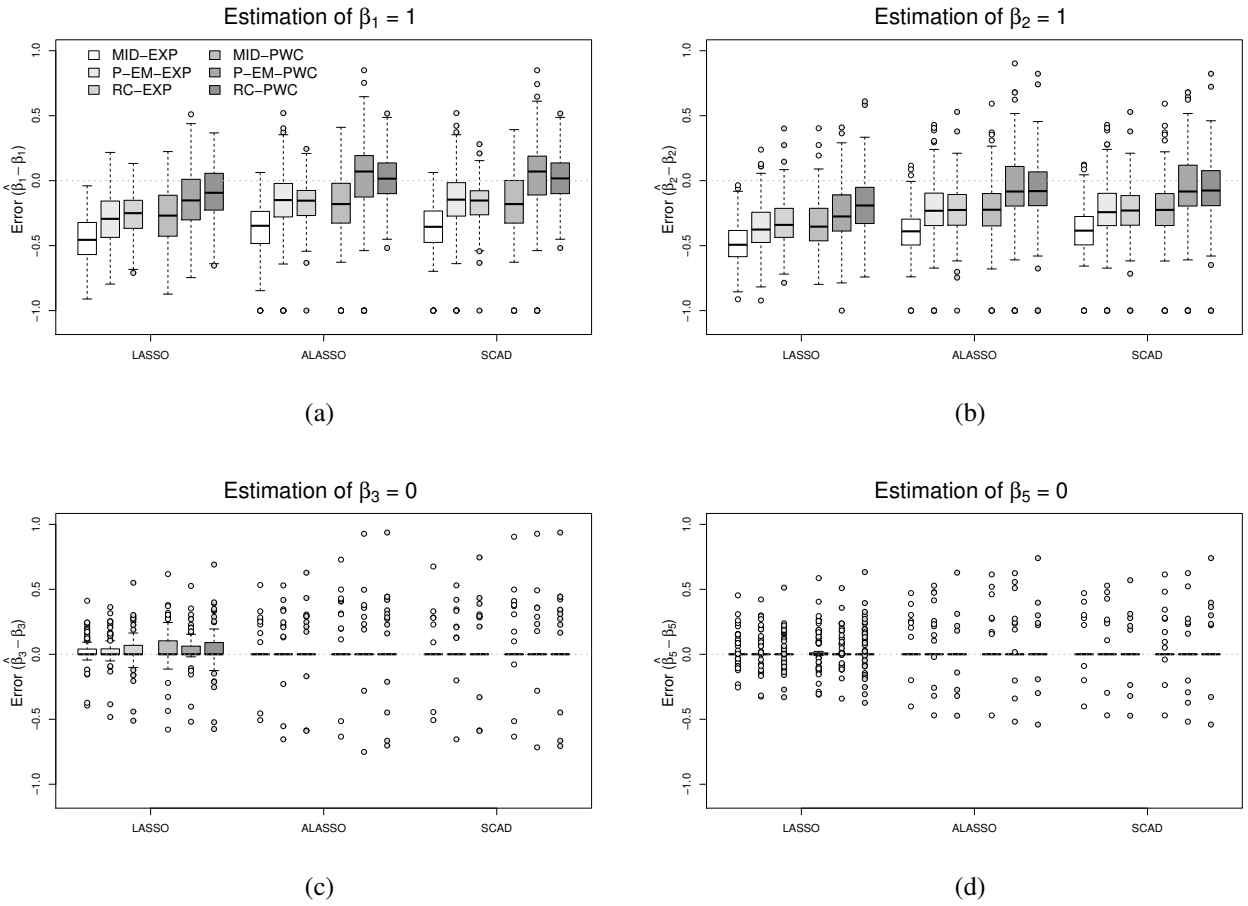


Figure A.1: Box plots of the error for the estimated regression coefficients $\widehat{\beta}_k - \beta_k$, $k = 1, 2, 3, 5$, for each penalty function for datasets with correlated binary covariates ($p = 10$) with $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$.

WEB APPENDIX B: COMPARISON OF METHODS FOR CHOOSING THE OPTIMAL TUNING PARAMETER

The selection of the tuning parameter λ is an important step in analyses based on penalized likelihood; when $\lambda = \infty$, none of the variables will be selected and when $\lambda = 0$, all of the variables will be selected in the usual fashion. Classical model selection methods are often based on the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) and more recent strategies have been based on cross-validation (CV) and generalized cross-validation (GCV). The traditional G-fold CV statistic is defined as

$$\widehat{CV}(\lambda) = \sum_{g=1}^{G} \left[\log L(\widehat{\boldsymbol{\theta}}_{-g}(\lambda)) - \log L_{-g}(\widehat{\boldsymbol{\theta}}_{-g}(\lambda)) \right]$$

where L_{-g} is the likelihood for the gth training dataset and $\widehat{\boldsymbol{\theta}}_{-g}(\lambda)$ is the estimate for the gth training data, obtained through the EM algorithm; the optimal λ maximizes $\widehat{CV}(\lambda)$.

Bradic et al. [1] mentioned that the measure of information contained in the full Cox partial likelihood is biased with respect to the number of nonzero elements and proper normalization is required. They proposed a sparse approximation to the generalized cross-validation statistic (SGCV) as

$$\widehat{SGCV}(\lambda) = \sum_{g=1}^{G} \left[\frac{\log L(\widehat{\boldsymbol{\theta}}_{-g}(\lambda))}{m\left(1 - \widehat{s}_{-g}(\lambda)/m\right)^2} - \frac{\log L_{-g}(\widehat{\boldsymbol{\theta}}_{-g}(\lambda))}{m_{-g}\left(1 - \widehat{s}_{-g}(\lambda)/m_{-g}\right)^2} \right]$$

where m_{-g} is the sample size of the gth training dataset and $\widehat{s}_{-g}(\lambda)$ is the number of non-zero coefficients. The optimal λ minimizes $\widehat{SGCV}(\lambda)$.

Here we compare three methods of selecting tuning parameters: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Table B.1 shows the results of comparisons of three methods for proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) for datasets with correlated binary covariates of dimension $p = 10$ and Table B.2 shows the corresponding results for datasets with multivariate normal covariates of dimension $p = 100$.

From these two tables, we see that for the LASSO penalty, SGCV shows some improvements in terms of a smaller number of incorrectly selected variables (FP), however, it also results in a smaller number of correctly selected variables (TP) and a larger mean squared error (MSE).

Compared with SGCV, both BIC and CV show good performance in terms of selecting tuning parameters for the ALASSO and SCAD penalties; BIC shows a smaller number of incorrectly selected variables (FP) than CV for LASSO penalty. Since, the R package `glmnet` uses cross-validation, we report the corresponding implementation of our algorithm using cross-validation to select tuning parameter.

		$\rho = 0.3$						$\rho = 0.6$					
		$\mu = 10$			$\mu = 20$			$\mu = 10$			$\mu = 20$		
Penalty	Method	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)	TP(4)	FP(6)	MSE (SD)
							Shape parameter: $\kappa = 1$						
LASSO	CV	4.00	2.34	0.314 (0.239)	4.00	2.47	0.262 (0.217)	3.99	2.21	0.305 (0.238)	4.00	1.91	0.265 (0.193)
	BIC	3.99	0.86	0.447 (0.305)	3.98	0.99	0.329 (0.251)	3.94	0.96	0.406 (0.315)	3.99	0.92	0.287 (0.231)
	SGCV	2.90	0.49	3.025 (1.411)	3.09	0.93	1.125 (1.463)	1.23	0.29	6.543 (2.290)	2.81	0.56	3.709 (1.982)
ALASSO	CV	3.79	0.65	0.276 (0.567)	3.92	0.82	0.223 (0.328)	3.80	1.13	0.309 (0.428)	3.88	0.73	0.215 (0.426)
	BIC	3.75	0.08	0.188 (0.502)	3.85	0.08	0.150 (0.373)	3.53	0.07	0.576 (0.355)	3.70	0.07	0.207 (0.377)
	SGCV	1.74	0.03	2.659 (1.564)	1.39	0.06	3.511 (1.591)	1.86	0.06	3.272 (2.079)	1.96	0.10	3.223 (2.006)
SCAD	CV	3.85	0.69	0.210 (0.481)	3.95	0.89	0.213 (0.301)	3.75	1.18	0.314 (0.444)	3.89	0.69	0.236 (0.310)
	BIC	3.73	0.06	0.176 (0.505)	3.85	0.07	0.148 (0.373)	3.48	0.04	0.670 (0.358)	3.65	0.05	0.237 (0.366)
	SGCV	1.50	0.00	3.516 (1.121)	1.60	0.02	3.511 (1.252)	1.44	0.02	3.914 (1.610)	1.38	0.03	3.908 (1.773)
							Shape parameter: $\kappa = 1.25$						
LASSO	CV	3.99	2.15	0.284 (0.240)	3.99	2.32	0.249 (0.197)	3.98	2.18	0.308 (0.222)	3.99	1.86	0.245 (0.184)
	BIC	3.99	0.82	0.364 (0.320)	3.98	0.91	0.303 (0.257)	3.94	0.68	0.371 (0.308)	3.98	0.83	0.269 (0.243)
	SGCV	2.95	0.30	2.235 (1.409)	3.09	0.85	1.157 (1.475)	2.97	0.75	2.111 (1.905)	2.97	0.71	1.756 (1.952)
ALASSO	CV	3.83	0.56	0.173 (0.561)	3.94	0.62	0.153 (0.307)	3.83	1.17	0.271 (0.293)	3.91	0.76	0.182 (0.267)
	BIC	3.84	0.05	0.127 (0.361)	3.87	0.14	0.141 (0.341)	3.49	0.03	0.653 (0.349)	3.67	0.02	0.166 (0.306)
	SGCV	1.83	0.02	2.624 (1.565)	1.44	0.02	3.511 (1.466)	1.33	0.08	3.846 (2.126)	2.29	0.07	3.203 (2.001)
SCAD	CV	3.84	0.55	0.165 (0.514)	3.95	0.74	0.156 (0.299)	3.81	1.19	0.282 (0.292)	3.88	0.69	0.148 (0.280)
	BIC	3.81	0.06	0.130 (0.380)	3.86	0.14	0.141 (0.356)	3.48	0.02	0.656 (0.402)	3.66	0.01	0.182 (0.308)
	SGCV	1.53	0.02	3.518 (1.314)	1.58	0.02	3.511 (1.326)	1.41	0.00	3.914 (1.564)	1.42	0.02	3.903 (1.557)

Table B.1: Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with correlated binary covariates ($p = 10$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE).

		$\mu = 10$			$\mu = 20$		
Penalty	Method	TP (10)	FP (90)	MSE (SD)	TP (10)	FP (90)	MSE (SD)
				Shape parameter: $\kappa = 1$			
LASSO	CV	10.00	14.80	0.312 (0.126)	10.00	14.83	0.261 (0.105)
	BIC	10.00	3.34	0.624 (0.199)	10.00	3.94	0.512 (0.184)
	SGCV	9.72	5.36	1.405 (0.823)	9.79	5.12	1.246 (0.692)
ALASSO	CV	10.00	0.12	0.057 (0.047)	10.00	0.07	0.047 (0.040)
	BIC	10.00	0.72	0.084 (0.072)	10.00	0.84	0.076 (0.057)
	SGCV	8.25	43.21	1.178 (1.329)	8.55	46.99	0.992 (1.011)
SCAD	CV	9.98	0.36	0.059 (0.073)	9.99	0.24	0.050 (0.048)
	BIC	10.00	0.84	0.082 (0.081)	10.00	0.79	0.068 (0.064)
	SGCV	9.55	58.93	1.275 (0.690)	9.51	53.23	0.940 (0.784)
				Shape parameter: $\kappa = 1.25$			
LASSO	CV	10.00	14.88	0.291 (0.118)	10.00	14.13	0.245 (0.109)
	BIC	10.00	3.37	0.604 (0.184)	10.00	3.78	0.501 (0.164)
	SGCV	9.81	0.96	1.277 (0.707)	9.64	2.70	1.227 (0.877)
ALASSO	CV	9.99	0.23	0.055 (0.050)	10.00	0.08	0.045 (0.031)
	BIC	10.00	0.59	0.068 (0.075)	10.00	0.90	0.071 (0.047)
	SGCV	9.54	62.70	1.024 (0.766)	7.14	29.37	0.983 (1.501)
SCAD	CV	9.98	0.29	0.055 (0.052)	9.99	0.13	0.044 (0.036)
	BIC	10.00	0.62	0.070 (0.085)	10.00	0.90	0.069 (0.058)
	SGCV	7.45	49.44	1.207 (1.987)	8.93	35.06	0.716 (0.735)

Table B.2: Comparison of three methods of choosing tuning parameter: cross-validation (CV), Bayesian information criterion (BIC) and sparse generalized cross-validation (SGCV). Analyses were based on interval-censored responses with multivariate normal covariates ($p = 100$) by using proportional hazards models with a piecewise constant baseline hazards with four pieces (PWC-4) and results are summarized in terms of the number of correctly (TP) and incorrectly (FP) selected variables and the median and standard deviation of the mean squared error (MSE).

WEB APPENDIX C: VARIANCE ESTIMATION

It is difficult to obtain an accurate estimate of the standard errors of the penalized estimator since the estimate is a non-linear and non-differentiable function of the responses, even for a fixed tuning parameter. One can, however, estimate the variance by using approximations or the bootstrap.

For the LASSO penalty, Tibshirani [4, 5] suggested estimating standard errors using either the bootstrap with either a fixed or an unfixed tuning parameter, or using an approximate form derived from ridge regression. For the SCAD penalty, Fan and Li [2] suggested that for moderate sample sizes, a sandwich-type variance formula derived from a local quadratic approximation (LQA) could be used for the covariance matrix, with modifications for large sample sizes. For the adaptive LASSO penalty, Zou [6] also used a LQA sandwich formula to approximate the variance of the estimators from penalized likelihood.

In the main paper, we propose an approach to variable selection for interval-censored failure times via a piecewise exponential model; it is not easy to derive an approximate approach to estimate standard errors. Therefore, we have employed a bootstrap approach to calculate standard errors of the penalized estimators. We draw a random sample $\boldsymbol{D}^*$ of size $m = 500$ with replacement from the original dataset $\boldsymbol{D}$ and we can obtain the penalized estimates $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ from $\boldsymbol{D}^*$ by using the proposed method with tuning parameter fixed at the optimal value that was determined from the original dataset $\boldsymbol{D}$. We repeat this process 500 times and get 500 bootstrap penalized estimates $\boldsymbol{\beta}^{*(1)}, \ldots, \boldsymbol{\beta}^{*(500)}$, so the bootstrap standard errors of the penalized estimators will be given by $\text{SE}(\beta_1^{*(1)}, \ldots, \beta_1^{*(500)}), \ldots, \text{SE}(\beta_p^{*(1)}, \ldots, \beta_p^{*(500)})$. Table C.1 shows the empirical biases, the average of the bootstrap standard errors, the empirical standard errors for the simulated datasets with $p = 10$, $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$ for both multivariate normal covariates and multivariate binary covariates. We can see that for the non-zero coefficients ($\beta_1, \beta_2, \beta_9, \beta_{10}$), the ASE and ESE agree well; for the zero coefficients, the ASE tends to be bigger than the ESE. We note that although we can calculate the standard errors based on the bootstrap or approximate approaches, it remains challenging to conceive how one would construct a confidence interval or compute a $p-$value based on a standard Wald-based pivotal or test statistic.

Penalty		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
						Multivariate Normal Covariate					
LASSO	EBIAS	-0.042	-0.043	0.008	-0.004	-0.002	0.000	0.005	-0.004	-0.052	-0.048
	ASE	0.057	0.057	0.035	0.036	0.036	0.035	0.036	0.036	0.058	0.057
	ESE	0.058	0.050	0.028	0.031	0.035	0.029	0.037	0.034	0.053	0.048
ALASSO	EBIAS	0.004	0.003	0.003	-0.001	0.001	0.000	0.002	-0.003	-0.006	-0.003
	ASE	0.061	0.062	0.043	0.045	0.046	0.044	0.045	0.045	0.062	0.060
	ESE	0.059	0.052	0.018	0.023	0.027	0.024	0.034	0.028	0.052	0.047
SCAD	EBIAS	0.004	0.002	0.005	-0.001	0.001	0.001	0.001	-0.003	-0.006	-0.003
	ASE	0.061	0.062	0.047	0.047	0.048	0.047	0.047	0.048	0.063	0.061
	ESE	0.059	0.052	0.020	0.021	0.029	0.020	0.036	0.030	0.052	0.048
						Multivariate Binary Covariate					
LASSO	EBIAS	-0.155	-0.231	0.033	0.032	0.020	0.008	0.003	0.005	-0.097	-0.100
	ASE	0.240	0.249	0.142	0.135	0.139	0.069	0.073	0.075	0.132	0.133
	ESE	0.258	0.258	0.122	0.118	0.100	0.059	0.057	0.075	0.132	0.139
ALASSO	EBIAS	-0.011	-0.071	0.018	0.036	0.016	0.005	0.001	-0.001	0.012	0.010
	ASE	0.259	0.273	0.182	0.170	0.175	0.094	0.096	0.095	0.143	0.141
	ESE	0.392	0.372	0.148	0.146	0.127	0.050	0.061	0.084	0.140	0.141
SCAD	EBIAS	-0.002	-0.071	0.018	0.031	0.007	0.005	-0.002	-0.003	0.013	0.009
	ASE	0.259	0.272	0.182	0.179	0.181	0.094	0.099	0.099	0.142	0.141
	ESE	0.381	0.366	0.145	0.139	0.119	0.042	0.053	0.083	0.142	0.141

Table C.1: Variance estimation by bootstrap for the simulated dataset with multivariate normal covariates and multivariate binary covariates for $\kappa = 1.25$, $\mu = 10$, $\rho = 0.3$.

REFERENCES

[1] Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics* **39**, 3092-3120.

[2] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74-99.

[3] Preisser, J. S. and Lohman, K. K. and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine* **21**, 3035-3054.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288.

[5] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385-395.

[6] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.