Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

On lasso for censored data

Brent A. Johnson

Department of Biostatistics
Rollins School of Public Health
Emory University
1518 Clifton Rd., NE
Atlanta, GA 30322
U. S. A.
e-mail: bajohn3@emory.edu

Abstract: In this paper, we propose a new lasso-type estimator for censored data after one-step EM update. While several penalized likelihood estimators have been proposed for censored data variable selection through hazards regression, many such estimators require parametric or proportional hazards assumptions. The proposed estimator, on the other hand, is based on the linear model and least-squares principles. Penalized Buckley-James estimators are also popular in this setting but have been shown to be unstable and unreliable. Unlike path-based learning based on least-squares approximation, our method requires no covariance assumption and the method is valid for even modest sample sizes. Our calibration estimator is equivalent to the minimizer of a well-defined convex loss function and, thus, yields an exact regularized solution path. Thus, the numerical algorithms are fast, reliable, and readily available because they build on existing software for complete, uncensored data. We examine the large and small sample properties of our estimator and illustrate the method through simulation studies and application to two real data sets.

Keywords and phrases: Accelerated failure time model, Buckley-James estimator, Least angle regression, Survival analysis, Synthetic data.

1. Introduction

Variable and model selection are very important topics in modern statistical inference. Partly due to advances in supervised learning, computer and statistical scientists have made significant strides in model selection over the last decade. Many such ideas are first conceived and initially developed in a least squares framework because no error distribution is necessary; a likelihood development soon ensues. Extending computational methods and model selectors to censored data applications presents a tremendous new challenge, particularly if the goals of model selection from the original, uncensored case remain unchanged and censoring is simply viewed as a nuisance. In this paper, we use a one-step EM update (Wang and Robins, 1999; Tsiatis, 2006; Jin et al., 2006) along with least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) to offer a principled new estimator for the censored data problem. The ideas described

*The research of the author was supported, in part, by a grant from the National Institutes of Allergies and Infectious Diseases (R03 AI068484) and Emory's Center for AIDS Research (P30 AI050409).

here persist whenever the complete data are unobserved; in this sense, the ideas are rather general and apply to many missing data problems.

Let y_i be the natural logarithm of the failure time variable for the i-th subject and consider the linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \ (i = 1, \ldots, n), \tag{1}$$

where $\mathbf{x}_i$ is a d-vector of fixed predictors for the ith subject, $\boldsymbol{\beta}$ is a d-vector of regression coefficients, and $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed errors with distribution function F. The lasso estimator estimates coefficients in model (1) while setting some coefficients exactly equal to zero. Specifically, the lasso estimator is given by the quadratic programming problem

$$\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \quad \text{subject to } \sum_{j=1}^{d} |\beta_j| \leq \tau, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, $\|\cdot\|$ is Euclidean norm and $\tau \geq 0$ is a user-specified regularization parameter. Now, define the observable random pair (U_i, δ_i) where $U_i = \min(y_i, C_i)$ and $\delta_i = I(y_i \leq C_i)$ for $i = 1, \ldots, n$, where C_i is a random censoring variable for the i-th subject and $I(\cdot)$ denotes the indicator function. The method proposed in this paper efficiently estimates the lasso coefficients consistent with model (1) through the original quadratic programming problem (2) using the observed data $\{(U_i, \delta_i, \mathbf{x}_i), \ i = 1, \ldots, n\}$.

When the response refers to the (natural) logarithm of a failure time, the semiparametric linear regression model is often referred to as the accelerated life or accelerated failure time (AFT) model (Cox and Oakes, 1984; Kalbfleisch and Prentice, 2002). Alternatively, statisticians (e.g. Tibshirani, 1997; Zhang and Lu, 2007) have extended lasso to censored data applications through the proportional hazards (PH) model (Cox, 1972), where a subject's hazard (i.e. instantaneous probability of failing) is modeled

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta}), \tag{3}$$

for an unspecified baseline hazard function $h_0(t)$. By many accounts, the PH model is most popular while the AFT model plays second fiddle. However, because the AFT model is based on the linear model, this has lead many prominent statisticians, most notably Sir D. R. Cox himself, to observe that the accelerated failure time model and the estimated regression coefficients to have a rather 'direct physical interpretation,' (Reid, 1994, p. 450). It is well-known that the linear model and the proportional hazards model cannot hold simultaneously except in the case of extreme value error distribution. For this and other reasons, developing methods for the AFT model is sufficiently interesting and significant.

Two general estimation strategies in the AFT model (1) include extensions of least-squares estimators through missing data techniques (Buckley and James,

1979; Miller and Halpern, 1982; Koul et al., 1981; Ritov, 1990; Lai and Ying, 1991) and rank-based methods (Prentice, 1978; Tsiatis, 1990; Lai and Ying, 1991). The foci of this paper are numerical and theoretical properties for extensions of (2) to censored data while comments regarding rank-based estimators are relegated to Section 5. Several authors have extended penalized Buckley-James statistics for variable selection (Datta et al., 2007; Wang et al., 2008; Johnson, 2008) and explored their small sample properties. An exemplary small sample finding is given by Wang et al. (2008), who found their penalized Buckley-James statistic converged in less than 50% of simulation studies. Similarly, Huang et al. (2006) developed a coordinate-descent algorithm (Friedman and Popescu, 2004) for a regularized Buckley-James statistic but found their algorithm to be rather unstable. Large sample properties of penalized Buckley-James estimators are also not encouraging. Recently, Johnson et al. (2008) proved that, under suitable regularity conditions, a penalized Buckley-James estimator yields only an *approximate*, root-n consistent solution, which is in stark contrast to the complete data case (Tibshirani, 1996; Knight and Fu, 2000; Zou, 2006). Using ideas initially put forth elsewhere (Ritov, 1990; Lai and Ying, 1991; Jin, Lin, and Ying, 2006), the method here is proposed to address the aforementioned numerical and theoretical problems with penalized Buckley-James statistics.

In a recent paper, Wang and Leng (2007) addressed similar concerns with penalized Buckley-James statistics through least-squares approximation. In simulation studies, the authors showed their procedure worked effectively. We note that the method of Wang and Leng (2007) requires a technical covariance assumption and, in practice, requires an estimate of the asymptotic covariance of the original unregularized regression coefficients and sufficiently large sample size to ensure the least-squares approximation is valid. In semi-parametric Buckley-James estimator, one estimates the asymptotic covariance of regression coefficients through resampling algorithms. First, our algorithm for estimating coefficients requires no resampling. Second, we achieve similar theoretical conclusions as in Wang and Leng (2007) but with no covariance assumption. Third, our method works for even modest sample sizes. The remainder of the paper is organized as follows: the method is described in Section 2 and its operating characteristics in Section 3. The utility of the proposed method is demonstrated through real and simulated examples in Section 4.

2. Methods

For mean-zero predictors, one version of the penalized Buckley-James estimating function (Datta et al., 2007; Wang et al., 2008; Johnson, 2008) is defined:

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbb{S}(\boldsymbol{\beta}) - n\lambda \cdot s(\boldsymbol{\beta}), \quad \mathbb{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i(\widetilde{Y}_i(\boldsymbol{\beta}) - \mathbf{x}_i'\boldsymbol{\beta}), \tag{4}$$

where $s(\boldsymbol{\beta}) = (\text{sgn}(\beta_1), \ldots, \text{sgn}(\beta_d))'$, λ is a user-defined regularization parameter,

$$\widetilde{Y}_i(\boldsymbol{\beta}) = \delta_i y_i + (1-\delta_i)\left[\mathbf{x}_i'\boldsymbol{\beta} + \frac{\int_{e_i(\boldsymbol{\beta})}^{\infty}\{1-\widehat{F}(s,\boldsymbol{\beta})\}\,ds}{1-\widehat{F}\{e_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}}\right],$$

$e_i(\boldsymbol{\beta}) = U_i - \mathbf{x}_i'\boldsymbol{\beta}$, and $\widehat{F}(t,\boldsymbol{\beta})$ is the left-continuous version of the Kaplan-Meier estimator of $F(t)$ based on $\{(e_i(\boldsymbol{\beta}),\delta_i),\ i = 1,\ldots,n\}$. It is well-known that $\mathbb{S}(\boldsymbol{\beta})$ is not monotone in $\boldsymbol{\beta}$ and may contain multiple roots (Ritov, 1990; Lai and Ying, 1991). Unfortunately, the poor behavior of the estimating function $\mathbb{S}(\boldsymbol{\beta})$ carries over to the penalized estimating function $\boldsymbol{\Psi}(\boldsymbol{\beta})$. Johnson et al. (2008, Corollary 2) have shown that for a class of penalty functions, there exists an approximate root-n consistent zero-crossing of a penalized Buckley-James estimating function but not an exact zero-crossing. In practice, this leads to non-convergent estimation of the semi-parametric EM algorithm (Datta et al., 2007; Wang et al., 2008; Johnson, 2008). It does not apply to the Buckley-James boosting method by Schmid and Hothorn (2008) because they assume a parametric error distribution F.

As noted by Jin et al. (2006) as well as earlier authors (Ritov, 1990; Lai and Ying, 1991), numerical problems typically associated with Buckley-James algorithms could be avoided if we started with a root-n consistent initial estimator for $\boldsymbol{\beta}_0$. Hence, the proposed estimator solves the penalized estimating function (Fu, 2003; Johnson et al., 2008)

$$\boldsymbol{\Psi}^C(\boldsymbol{\beta}) = \mathbb{S}^C(\boldsymbol{\beta}) - n\lambda\cdot s(\boldsymbol{\beta}), \quad \mathbb{S}^C(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{x}_i(\widetilde{Y}_i(\boldsymbol{\beta}_0) - \mathbf{x}_i'\boldsymbol{\beta}), \tag{5}$$

where $\boldsymbol{\beta}_0$ is the true value for $\boldsymbol{\beta}$ in the linear model (1). For now, we assume the errors have mean zero, i.e. $E(\varepsilon_1) = 0$, and consider the minimizer of the following convex loss function

$$\frac{1}{2}\|\widetilde{\mathbf{Y}}(\boldsymbol{\beta}_0) - \mathbf{X}\boldsymbol{\beta}\|^2 + n\lambda\sum_{j=1}^{d}|\beta_j|, \tag{6}$$

where $\widetilde{\mathbf{Y}}(\boldsymbol{\beta}_0) = (\widetilde{Y}_1(\boldsymbol{\beta}_0),\ldots,\widetilde{Y}_n(\boldsymbol{\beta}_0))'$. It is not difficult to argue that a minimizer of (6) is a solution to the estimating equations in (5). Thus, if the true value $\boldsymbol{\beta}_0$ were known, we would define our regularized estimator as the solution to following equivalent quadratic programming problem:

$$\widehat{\boldsymbol{\beta}}_C = \min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\widetilde{\mathbf{Y}}(\boldsymbol{\beta}_0) - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{7}$$
$$\text{subject to } \sum_{j=1}^{d}|\beta_j| \le \tau,$$

which is exactly the complete data lasso (2) on synthetic data $\widetilde{\mathbf{Y}}(\boldsymbol{\beta}_0)$. Like the lasso for uncensored data, the solution to the quadratic programming program

(7) is exact for every τ. Since $\boldsymbol{\beta}_0$ is unknown in practice, we replace $\boldsymbol{\beta}_0$ with an initial consistent estimator $\widehat{\boldsymbol{\beta}}_I$. As in Jin et al. (2006), we adopt the Gehan (1965) estimator,

$$\widehat{\boldsymbol{\beta}}_I = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \left\{ e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta}) \right\}^{-},$$

where $c^{-} = \max(-c, 0)$. In short, the new method can be summarized in the following three steps.

Algorithm 1.

1. Estimate $\widehat{\boldsymbol{\beta}}_I$ using observed data $\{(U_i, \delta_i, \mathbf{x}_i)\ i = 1, \ldots, n\}$,
2. Construct the pseudo responses $\zeta_i = \widetilde{Y}_i(\widehat{\boldsymbol{\beta}}_I) - \left[n^{-1} \sum_{j=1}^{n} \widetilde{Y}_j(\widehat{\boldsymbol{\beta}}_I) \right]$,
3. Lasso using synthetic data, $\{(\zeta_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$.

Because the synthetic responses are normalized in step 2 prior to step 3, there is no substantive requirement that $E(\varepsilon_1) = 0$.

2.1. The impact of the initial estimator $\widehat{\boldsymbol{\beta}}_I$

While we use the Gehan estimator to derive asymptotic results, other consistent estimators may be used. Under suitable regularity conditions, several candidate consistent initial estimators include the Gehan (1965) estimator, the log-rank estimator (Mantel, 1966), the family of $G^{\rho,\delta}$ estimators (Fleming and Harrington, 1991), the inverse-probability weighted estimator (Koul et al., 1981), doubly robust estimators (van der Laan and Robins, 2003; Tsiatis, 2006; Rubin and van der Laan, 2007), or the semi-parametric efficient estimator (Zeng and Lin, 2007). It is of sufficient interest to understand the impact of the initial estimator $\widehat{\boldsymbol{\beta}}_I$ on the efficiency of the procedure. This is a difficult problem even without regularization. We offer some insight into this problem by making the connection to multiple imputation.

Wang and Robins (1999) discuss this issue for related type-B imputation estimators with parametric imputation. Actually, Wang and Robins' (1999) type-B estimator is asymptotically equivalent to the estimator of Jin et al. (2006) for infinite imputations and parametric distribution F. Following Wang and Robins (1999, Theorem 1), we deduce that without the task of variable, that a type-B estimators has asymptotic variance $V_B \approx V_1 + V_2 + V_3$, where V_1 is the asymptotic variance of the estimator of Zeng and Lin (2007), V_2 and V_3 are two positive semidefinite matrices, and the approximation "$\approx$" is due to the Kaplan-Meier estimate of the distribution function F. The second term V_2 is attributable to the inefficiency of $\widehat{\boldsymbol{\beta}}_I$ and vanishes if the efficient estimator is used. The third term depends of the complete-data covariance and missing information but vanishes as the number of imputation increases; thus $V_3 \equiv 0$ for Buckley-James type estimators. As noted by Wang and Robins (1999), the type-B imputation estimator may or may not be more efficient than the initial estimator $\widehat{\boldsymbol{\beta}}_I$. Heuristically, the mixed message persists in regularized estimation.

2.2. Algorithmic notes

We make three important points about the proposed procedure. First, it is important to note that the new method applies to any and all lasso extensions, including adaptive lasso (alasso; Zou, 2006) and elastic net (enet; Zou and Hastie, 2005). Second, it is equally important to note that any algorithm (efficient or inefficient) which yields lasso coefficients can be used for our estimation purposes here. Third, any path-based algorithm on synthetic data produces valid coefficient paths for any given data set and the paths are completely reproducible. In particular, any `lars` (Efron et al., 2004) option (i.e. lars, forward stagewise, and stepwise regression) produces valid coefficient paths. In the sequel, we use Berwin Turlach's `lasso2` package when considering coefficient estimates for fixed τ (or λ) and the `lars` package (Efron et al., 2004) when considering coefficient paths.

In general, all penalized least squares methods require tuning of the regularization parameter. In our setup, we suggest V-fold cross-validation given by

$$\mathrm{CV}(\lambda) = \sum_{v=1}^{V} \sum_{(\zeta_k, \mathbf{x}_k) \in \mathcal{D}^v} \{\widetilde{Y}_k(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_k' \widehat{\boldsymbol{\beta}}_C^{(v)}\}^2,$$

where $\mathcal{D}$ is the full data set, $\mathcal{D}^v$ and $\mathcal{D} - \mathcal{D}^v$ are the test and training data, respectively, and $\widehat{\boldsymbol{\beta}}_C^{(v)}$ is the estimate found from the training set $\mathcal{D} - \mathcal{D}^v$.

3. Operating characteristics

In this section, we describe the operating characteristics for the estimator $\widehat{\boldsymbol{\beta}}_C$. Without loss of generality, we use the notation $\widehat{\boldsymbol{\beta}}_C$ to refer to the estimator regardless of the penalty. The asymptotic properties of penalized least squares subject to lasso constraints were first presented by Knight and Fu (2000). Knight and Fu's Theorems 1-2 suggest that the lasso estimator, for uncensored data, converges in probability and distribution, respectively, to the unique minimizers of convex functions. By examining our calibrated loss function in (6), it natural to believe that our estimator possesses similar limiting behavior. This result, subject to certain regularity conditions, turns out to be true upon careful inspection of the calibrated loss function.

3.1. Lasso shrinkage

We begin by stating several theorems regarding the limiting behavior of the regularized estimator with lasso penalty. For Theorems 1-2, we define the estimator $\widehat{\boldsymbol{\beta}}_C$ with lasso penalty as the minimizer of the random function

$$\Phi_{1,n}(\mathbf{u}) = n^{-1} \sum_{i=1}^{n} (\widetilde{Y}_i(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i' \mathbf{u})^2 + \frac{\lambda_n}{n} \sum_{j=1}^{d} |u_j|. \tag{8}$$

We also define the nonsingular matrix

$$\boldsymbol{\Sigma} = \lim_n \mathbf{X}'\mathbf{X}.$$

Theorem 1. *Assume conditions (C.1)-(C.4) and* $\lambda_n/n \to \lambda_0$, *then* $\widehat{\boldsymbol{\beta}}_C \to_p$ *argmin*(Φ_1), *where*

$$\Phi_1(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}(\mathbf{u} - \boldsymbol{\beta}_0) + \lambda_0 \sum_{j=1}^{d} |u_j|.$$

As noted in Knight and Fu (2000), if $\lambda_n = o(n)$, then Theorem 1 implies $\widehat{\boldsymbol{\beta}}_C$ is consistent. Furthermore, Theorem 1 implies that as $\lambda_n/n \to \lambda_0$, the regularized estimator shares the same limit as the lasso for uncensored data. On the one hand, the conclusion of Theorem 1 is unimpressive because we have only extended Knight and Fu's (2000) Theorem 1. However, note that we must impose much stronger conditions (C.1)-(C.4) because of censoring, not the least of which is a tail modification in (C.1). Unlike the conclusions of Knight and Fu's (2000) Theorem 1, if conditions (C.1)-(C.4) do not hold, our conclusions in Theorem 1 may no longer be true. Nevertheless, for point estimation purposes, lasso behaves as if the pseudo data are the true data. The asymptotic price of using pseudo data in Algorithm 1 is covered in Theorem 2.

Theorem 2. *Assume (C.1)-(C.4),* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. *Then,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0) \to_d \textit{argmin}(\Phi_2),$$

where

$$\Phi_2(\mathbf{u}) = -2\mathbf{u}'\mathbf{W}_1 - 2\mathbf{u}'\mathbf{W}_2 + \mathbf{u}'\Sigma\mathbf{u} + \lambda_0 \sum_{j=1}^{d} \left[u_j \, sgn(\beta_{j0}) I(\beta_{j0} \neq 0) + |u_j| I(\beta_{j0} = 0)\right],$$

$\mathbf{W}_1$ *has a* $N(0, \mathbf{V})$, $\mathbf{W}_2$ *has a* $N(0, \mathbf{D})$, *where* $\mathbf{V}$ *and* $\mathbf{D}$ *are defined in the Appendix.*

While Theorem 1 suggests an asymptotically negligible effect for using pseudo data in lasso-type estimation, the story is quite different in Theorem 2. The basic form of the asymptotic distribution matches that of Knight and Fu's (2000) Theorem 2; however, our random variables $\mathbf{W}_1$ and $\mathbf{W}_2$ are quite different.

Hence, we have argued for the consistency and asymptotic normality of the calibrated estimator with lasso penalty. Note, if we had defined the estimator iteratively as in penalized Buckley-James estimators, the conclusions of Theorems 1-2 would not follow (at least, not by the proofs given in the Appendix). The latter iterative definition of penalized Buckley-James estimating function leads, in principle, to the problems discussed in Johnson et al. (2008).

3.2. Oracle properties

To make claims about the oracle behavior of the estimator, we first state a preliminary result about the unpenalized, calibrated estimator by Jin et al. (2006). Define the unpenalized estimator $\widehat{\boldsymbol{\beta}}_U$ as the solution to $0 = \mathbb{S}^U(\boldsymbol{\beta})$, where,

$$\mathbb{S}^U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \{\widetilde{Y}_i(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'\boldsymbol{\beta}\}.$$

The following Lemma from Jin et al. (2006) gives the consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_U$.

Lemma 1. *Under conditions (C.1)-(C.4),* $\|\widehat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ *and* $n^{1/2}(\widehat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0)$ *converges to a mean-zero Gaussian random vector with covariance* $\boldsymbol{\Omega}$.

Using the results of Lai and Ying (1991), Jin et al. (2006) show that

$$\widehat{\boldsymbol{\beta}}_U \cong (\mathbf{I}_d - \boldsymbol{\Sigma}^{-1}\mathbf{A})\widehat{\boldsymbol{\beta}}_I + (\boldsymbol{\Sigma}^{-1}\mathbf{A})\widehat{\boldsymbol{\beta}}_{\mathrm{BJ}},$$

where $\mathbf{I}_d$ is the identity matrix of dimension d. In other words, the estimator $\widehat{\boldsymbol{\beta}}_U$ lies on the line-segment between the Gehan and Buckley-James estimator. Because both $\boldsymbol{\Sigma}$ and $\mathbf{A}$ are assumed nonsingular, the consistency of $\widehat{\boldsymbol{\beta}}_U$ follows immediately.

Theorem 3 states the asymptotic result for the estimator with alasso penalty (Zou, 2006) including the existence of an $n^{1/2}$-consistent estimator, the sparsity of the estimator and the asymptotic normality of the estimator. Let $\mathcal{A}$ denote the indices of the predictors in the true model, i.e. $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$. Define the calibrated estimator $\widehat{\boldsymbol{\beta}}_C$ with alasso penalty as the minimizer of the following objective function:

$$\Phi_{3,n}(\mathbf{u}) = n^{-1}\sum_{i=1}^{n}(\widetilde{Y}_i(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'\mathbf{u})^2 + \frac{\lambda_n}{n}\sum_{j=1}^{d}\pi_j|u_j|. \tag{9}$$

where $\pi_j = 1/|\widehat{\beta}_{I,j}|$ for $j = 1, \ldots, d$. Recall, $\widehat{\boldsymbol{\beta}}_I$ is root-n consistent for $\boldsymbol{\beta}_0$. Using the Gehan estimator $\widehat{\boldsymbol{\beta}}_I$ in the definition of weights π_j is novel but also consistent with the literature (Zou, 2006; Zhang and Lu, 2007; Wang et al., 2007; Johnson et al., 2008).

Theorem 3. *Let* $\widehat{\boldsymbol{\beta}}_C$ *be defined as in* (9) *and assume the regularity conditions (C.1)-(C.4). If* $\sqrt{n}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, *then* $\|\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, $\lim_n P(\widehat{\boldsymbol{\beta}}_{C,j} = 0) = 1$, *for every* $j \notin \mathcal{A}$, *and*

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}}_{C,\mathcal{A}} - \boldsymbol{\beta}_{0,\mathcal{A}}\right) \to_d N(0, \boldsymbol{\Omega}_{\mathcal{A}}),$$

where $\boldsymbol{\Omega}$ *is defined in Lemma 1 and* $\boldsymbol{\Omega}_{\mathcal{A}}$ *is the sub-matrix containing only the elements of* $\boldsymbol{\Omega}$ *whose indices belong to* $\mathcal{A}$.

Recently, Johnson (2008) argued that, at least approximately, his penalized Buckley-James estimator possessed an oracle property with non-concave penalty (Fan and Li, 2001). The estimator in Theorem 3 possesses the same asymptotic properties as Johnson's (2008) estimator under similar conditions but $\widehat{\boldsymbol{\beta}}_C$ is the global minimizer of a convex loss function.

3.3. Grouping effect

Here, we illustrate that our estimator with elastic net penalty possesses the small sample *grouping effect* property (Zou and Hastie, 2005, Theorem 1). The grouping effect of the elastic net asserts that the regression coefficient estimates for highly correlated predictors will be nearly identical. The proposed naïve elastic net estimator for censored data is defined as the minimizer of the following naïve elastic net criterion:

$$\|\widetilde{\mathbf{Y}}(\widehat{\boldsymbol{\beta}}_I) - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{d} |\beta_j| + \lambda_2 \sum_{j=1}^{d} \beta_j^2, \tag{10}$$

where λ_1 and λ_2 are user-defined regularization parameters. Let $\widehat{\boldsymbol{\beta}}_C(\lambda_1, \lambda_2)$ be the minimizer of (10) using regularization parameters (λ_1, λ_2) and $\widehat{\beta}_{C,j}(\lambda_1, \lambda_2)$ denote j-th element of $\widehat{\boldsymbol{\beta}}_C(\lambda_1, \lambda_2)$. Suppose that the product $\widehat{\beta}_{C,j}(\lambda_1, \lambda_2) \cdot \widehat{\beta}_{C,k}(\lambda_1, \lambda_2) > 0$. Define the scaled difference between coefficient estimates

$$\Delta_{j,k}(\lambda_1, \lambda_2) = \frac{1}{\|\widetilde{\mathbf{Y}}(\widehat{\boldsymbol{\beta}}_I)\|_1} \left| \widehat{\beta}_{C,j}(\lambda_1, \lambda_2) - \widehat{\beta}_{C,k}(\lambda_1, \lambda_2) \right|.$$

Using the same arguments as in Zou and Hastie (2005, Appendix) and the calibrated loss function in (10), one can show that

$$\Delta_{j,k}(\lambda_1, \lambda_2) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho_{jk})}$$

where $\rho_{jk} = \text{corr}(x_j, x_k)$, the correlation between predictors x_j and x_k. In other words, the calibrated estimator with elastic net penalty possesses the grouping effect property, albeit on a slightly different scale than its uncensored data counterpart.

3.4. Statistical Inference

Statistical inference for regularized estimators is tricky, even for ordinary lasso and penalized least squares estimators (Tibshirani, 1996; Osborne et al., 2000). Fan and Li (2001) suggest drawing approximate inference through local quadratic approximation for the regression coefficient estimates in the active set, i.e. $\widehat{\boldsymbol{\beta}}_{C,\mathcal{A}}$. For estimators that achieve an oracle property, this inferential procedure is based on an asymptotic argument where the standard errors for inactive coefficients

are set to zero (Fan and Li, 2001; Zou, 2001; Johnson et al., 2008). For lasso, elastic net, and all other penalties, this inferential procedure is *ad hoc*, although authors report reasonable behavior of the "approximate" standard errors. Under the assumption that these standard errors are indeed valid across the spectrum of penalized estimators, one can obtain standard error estimates for the one-step, calibrated estimators using the resampling technique described by Jin et al. (2006).

4. Examples

4.1. Comparison of one-step versus fully iterative estimators

In multiple linear regression, Jin et al. (2006) describe an iterative algorithm that partially motivated the calibrated estimator $\widehat{\boldsymbol{\beta}}_C$. Here, we investigate small sample differences between an iterative estimator compared to the one-step estimator. We simulated 500 data sets from the normal-theory linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i,\ i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (1, 1/4, 0, 1/2)'$, $\sigma = 1$, and ε_i are independent and identically distributed standard normal. Here, we consider a fixed design $\mathbf{x}$ with correlation between the j-th and k-th components of x equal to $0.5^{|j-k|}$. The censoring distribution is assumed to be uniform$(0, \tau)$, where τ yielding 20% censoring. In this particular simulation study, we only consider the lasso penalty. We compare the Monte Carlo average and standard deviation of the lasso on the unobserved (true) data, the calibrated and a second fully iterative estimator. We considered a range values of the regularization parameter an order of magnitude apart but only present the extreme values, $\lambda_1 = n^{-1/2}$ and $\lambda_2 = n^{-2}$. Note that a regularization parameter $\lambda = n^{-1}$ corresponds roughly to the rule-of-thumb regularization parameter for Akaike Information Criterion (AIC) in uncensored data (Tibshirani, 1996; Wang et al., 2007). The simulation results for three sample sizes are presented in Table 1.

Our simulation studies indicate that for appropriate choices of the regularization parameter λ, the Monte Carlo average of the one-step estimator approaches the lasso in uncensored data as $n \to \infty$ and $\lambda \to 0$, which suggests the conclusions of Theorem 1 are correct. Note the absolute difference in $\widehat{\boldsymbol{\beta}}_1$ between the one-step estimator and uncensored lasso is about 0.05 for λ_1 at $n = 50$ and about 0.03 for λ_2 at $n = 100$. Small sample differences in Monte Carlo averages between the one-step and uncensored data estimator is partially due to the tail modification of the the pseudo data (See condition (C.1) in the Appendix). However, the Monte Carlo average of the iterative estimator is nearly identical to the uncensored lasso for the particular values of λ, especially as n increases. We also note that for fixed λ, the Monte Carlo standard deviation for both the one-step and iterative estimators is about the same. Thus, in small samples, we found no evidence to suggest that an iterative estimator is significantly more precise than the one-step estimator.

TABLE 1
Comparing the differences of the proposed estimation procedures. Table entries (×10E-3) are equal the mean and standard deviation of lasso estimates over 500 Monte Carlo data sets for $\lambda_1 = n^{-1/2}$ and $\lambda_2 = n^{-2}$.

	One-step		Iterative		True data	
	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2
$n = 50, \sigma = 1$						
$\widehat{\beta}_1$	901 (130)	956 (133)	935 (126)	999 (132)	948 (124)	1002 (128)
$\widehat{\beta}_2$	182 (147)	252 (189)	175 (142)	262 (188)	182 (141)	259 (180)
$\widehat{\beta}_3$	66 (113)	13 (242)	46 (91)	-10 (239)	54 (95)	-13 (227)
$\widehat{\beta}_4$	329 (186)	493 (225)	314 (179)	500 (224)	328 (180)	499 (219)
$n = 75, \sigma = 1$						
$\widehat{\beta}_1$	905 (116)	961 (121)	942 (116)	1000 (120)	951 (114)	1001 (117)
$\widehat{\beta}_2$	135 (123)	162 (150)	194 (134)	247 (153)	197 (132)	247 (149)
$\widehat{\beta}_3$	80 (107)	81 (175)	44 (84)	9 (177)	45 (87)	5 (170)
$\widehat{\beta}_4$	316 (157)	442 (179)	340 (154)	497 (180)	353 (153)	498 (175)
$n = 100, \sigma = 1$						
$\widehat{\beta}_1$	918 (110)	968 (114)	948 (110)	1005 (114)	957 (107)	1004 (110)
$\widehat{\beta}_2$	188 (122)	216 (142)	201 (121)	249 (142)	211 (118)	248 (135)
$\widehat{\beta}_3$	49 (76)	40 (145)	32 (63)	6 (146)	34 (64)	7 (140)
$\widehat{\beta}_4$	396 (124)	484 (136)	379 (120)	494 (135)	400 (118)	495 (131)

4.2. Illustration of Oracle Properties

In this simulation exercise, we illustrate that the calibrated estimator $\widehat{\boldsymbol{\beta}}_C$ possesses the so-called oracle properties (cf. Fan and Li, 2001; Zou, 2006; Wang et al., 2008; Johnson et al., 2008). Here, we simulated 200 data sets from the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i, \ i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (3, 3/2, 0, 0, 2, 0, 0, 0)'$, and $\mathbf{x}_i$ and ε_i are independent standard normal with correlation between the j-th and k-th components of $\mathbf{x}$ equal to $0.5^{|j-k|}$. This model has been considered elsewhere in the model selection literature (cf. Tibshirani, 1996; Fan and Li, 2001). We compare the model error $\text{ME} \equiv (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'E(\mathbf{x}\mathbf{x}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ of the lasso and adaptive lasso (alasso) estimator for both the one-step and fully iterative estimator. The median model error (MME) over 200 Monte Carlo data sets is reported for each simulation scenario. We also compare the average numbers of regression coefficients that are correctly (C) or incorrectly (I) shrunk to 0. Furthermore, we consider two different strategies for parameter tuning: one based on cross-validation and a second based on the rule-of-thumb $\widehat{\lambda} = \log(n)/n$ (Wang et al., 2007). All simulation results are presented in Table 2, where *oracle* pertains to the situation in which non-zero coefficients are known *a priori*.

First, we note that using the rule-of-thumb $\widehat{\lambda}$ leads to very poor performance in both the calibrated estimator and fully iterative estimators regardless of the

TABLE 2

Simulation results on model selection. Table entries include the median model error (MME), average number of correct (C) and incorrect (I) zeros over 200 Monte Carlo data sets. For each of lasso and alasso estimators, the regularization parameter is either set equal to a rule-of-thumb value or selected after cross-validation.

		Iterative			One-step		
Method		MME (×100)	C	I	MME (×100)	C	I
$n = 50, \sigma = 3$							
lasso	$\widehat{\lambda}$	143.77	0.74	0	152.03	0.64	0
	CV	193.85	3.11	0.06	179.09	2.60	0.04
alasso	$\widehat{\lambda}$	134.10	2.31	0.02	145.59	2.13	0.02
	CV	126.47	4.25	0.10	125.70	3.92	0.08
$n = 50, \sigma = 1$							
lasso	$\widehat{\lambda}$	18.21	2.11	0	20.97	1.52	0
	CV	22.24	2.46	0	25.86	2.19	0
alasso	$\widehat{\lambda}$	10.37	4.58	0	12.26	4.33	0
	CV	11.80	4.37	0	13.71	4.24	0
$n = 75, \sigma = 3$							
lasso	$\widehat{\lambda}$	93.72	0.58	0	106.81	0.52	0
	CV	161.32	3.30	0.01	130.48	2.58	0
alasso	$\widehat{\lambda}$	79.27	2.26	0	92.27	2.10	0
	CV	79.60	4.32	0.04	77.98	4.05	0.04
$n = 75, \sigma = 1$							
lasso	$\widehat{\lambda}$	12.52	1.77	0	13.43	1.38	0
	CV	18.68	2.58	0	18.14	2.27	0
alasso	$\widehat{\lambda}$	6.71	4.63	0	7.92	4.42	0
	CV	6.79	4.55	0	8.35	4.36	0

penalty. This is due, in part, to the fact that $\widehat{\lambda}$ was proposed for uncensored data, not censored data. One can see that tuning based on $\widehat{\lambda}$ produces models that are too complex, on average, suggesting that $\widehat{\lambda}$ is too small in censored data applications. Cross-validating λ leads to estimates with the desired operating characteristics. By now, it is well-known that lasso performs better than alasso when the signal-to-noise ratio is low while the opposite is true when the signal-to-noise ratio is large. This phenomenon can be observed in Table 2. We observe that the there may be some advantage (i.e. smaller model error and model complexity) in the iterative estimator over the one-step as the sample size increases. However, the convergence of the iterative estimator is not guaranteed. Nevertheless, with careful tuning of the regularization parameter λ for the iterative estimator, both the one-step and the iterative estimator possess similar operating characteristics, on average; however, the one-step estimator does so with far less computational cost.

4.3. Mayo Primary Biliary Cirrhosis Study

We consider the Mayo primary biliary cirrhosis (PBC) data (Fleming and Harrington, 1991, Appendix D.1). The data contains information about the survival time and prognostic variables for 418 patients who were *eligible* to participate in a randomized study of the drug D-penicillamin. Of 418 patients who met standard eligibility criteria, a total of 312 patients participated in the randomized portion of the study. The study investigators used stepwise deletion to build a Cox proportional hazards model for the natural history of PBC (Dickson et al., 1989). We perform model selection with ten predictors including age, log(albumin), log(alkaline phosphatase), ascites, log(bilirubin), edema, hepatomegaly, log(protime), sex and spiders. This is the same data set used by Johnson (2008) but his method estimated regression coefficients through local quadratic approximation (Tibshirani, 1996; Fan and Li, 2001), which is very different than the algorithm presented here. Coefficient paths for each of the ten predictors are displayed in Figure 1.

Figure 1 illustrates the coefficient paths for each of lasso, alasso, and elastic net (enet). In panels (a)-(c), we present coefficient paths from both Berwin Turlach's `lasso2` package as well as the piecewise-linear `lars` coefficient paths (Efron et al., 2004) in panels (d)-(f). Finally, we cross-validated each of the three estimators using the BIC criterion. The final estimated regression coefficients using the optimal regularization parameter (not shown here but in an earlier technical report) are similar for lasso and elastic net coefficient estimates. However, the alasso coefficient estimates are rather different than the other two estimators; in particular, the final alasso model has two fewer coefficients than either lasso or elastic net.

4.4. Lung cancer microarray data

In 2003, Research Triangle Park (North Carolina, U. S. A.) hosted the third international conference on the Critical Assessment of Microarray Data Anal-

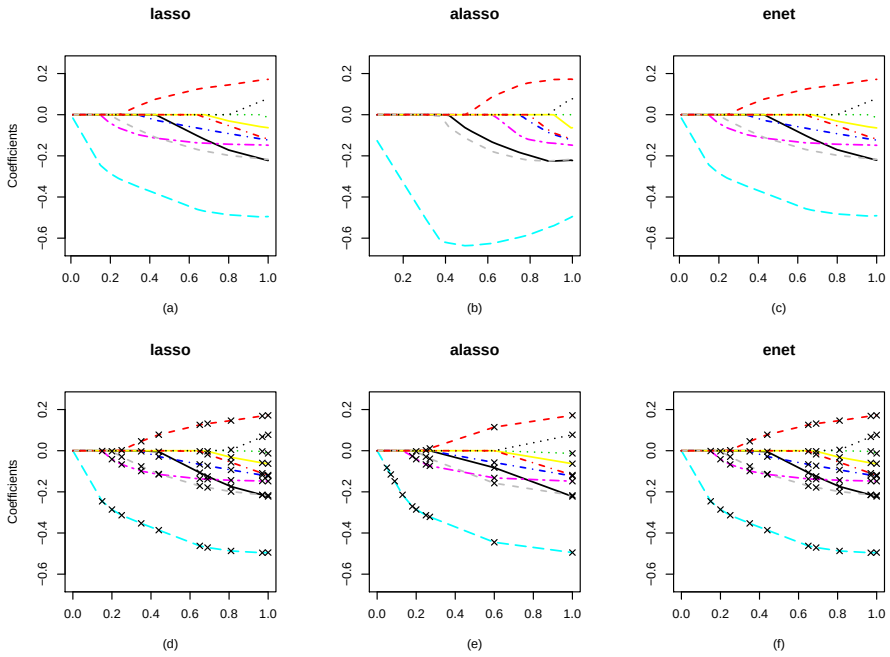

FIG 1. *Coefficient paths for estimation and variable selection in the semiparametric linear regression model using calibrated one-step estimation and data from the Mayo primary biliary cirrhosis study. Panels (d)-(f) are piece-wise linear coefficient paths with a cross (x) marking changes in the active set $\mathcal{A}(\lambda)$; panels (a)-(c) are approximate coefficient paths fit over a range of the regularization parameter τ.*

ysis (CAMDA). CAMDA offers researchers from all areas the opportunity to brainstorm new methodologies for the analysis of microarray data. We analyze part of the CAMDA 2003 challenge data set for which survival endpoints and microarray data are both available. Specifically, our original data set consists of 124 Harvard samples, analyzed using Affymetrix microarray chips, and 1034 genes while our actual data analysis uses a smaller subset of genes (see below). Additional details can be found on the CAMDA 2003 website (http://www.camda.duke.edu/camda03/datasets/index.html).

In this data set, the survival endpoint is the natural logarithm of time-to-death (in months) , which may be right-censored at the end of the study followup period. Our analysis consists of 94 gene expression levels on 124 samples. The 94 genes represents an important subset of the original 1034 genes chosen through univariate Gehan regressions on each gene separately and selecting those genes with a Wald test statistic greater than 2.25 in absolute value. This initial preprocessing step is ad hoc and not included as part of the original estimation scheme in Algorithm 1 but also not uncommon in the microarray literature (cf. Wang et al., 2008, and references therein). Available software for simultaneous testing (e.g. local FDR, Efron, 2005) suggests that fewer than 94 genes are important.

We fit the calibrated estimator with elastic net penalty (Zou and Hastie, 2005) to the reduced lung cancer microarray data. We then constructed two and three risk groups from quantiles of the elastic net predictions and then estimate the stratified survivor function $P(y_1 \geq t)$ using the Kaplan-Meier estimator. The estimated survivor functions are presented in Figure 2 and suggest the elastic net predictions on the 94 significant genes yield reasonable separation. The score (i.e. logrank) test statistics are 22.9 and 24.7 on one and two degrees of freedom for the two and three group analyses, respectively. Wang et al. (2008) present a similar analysis of their microarray data based on penalized Buckley-James estimators with elastic net penalty.

5. Remarks

Penalized Buckley-James statistics seem like a natural way of extending penalized least squares to censored data. Unfortunately, the unreliability of the estimator has not supported its use in practice. Instead, we propose a calibrated lasso-type estimator that offers investigators a new method for simultaneous estimation and variable selection in the linear regression model (1) with censored data. The hypotheses of the new estimator are similar to penalized Buckley-James statistics, but our estimator possesses better small sample properties. The small and large sample properties of the one-step calibrated estimator illustrates its good performance.

The success of our method is due to a principled initial value (Jin et al., 2006) which provides a solid footing for local estimation. The subsequent use of normalized pseudo response data leads to a reliable estimator which can be written as the unique solution of a convex loss function. While others have considered similar lasso-type estimation in the AFT model, none appear quite as simple,

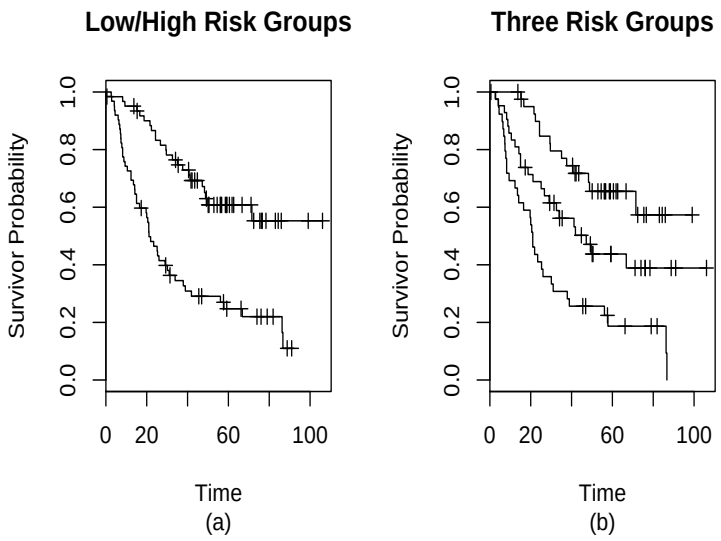

FIG 2. *Stratified Kaplan-Meier curves for the lung cancer microarray data and elastic net predicted values.*

straightforward, or generalizable as the proposed approach. Our procedure is accomplished through three easy steps: find a consistent initial estimate for $\boldsymbol{\beta}_0$ in the linear model (1), construct synthetic responses, and compute the lasso optimization problem (2). Because there are no additional iterative steps, this fundamentally sets it apart from other penalized Buckley-James statistics (e.g. Wang et al., 2008). Technically, we place modest assumptions on the tail of the error distribution but these are the same regularity conditions as in Jin et al. (2006).

Rather the estimate coefficient paths based on imputation and least-squares principle, one may choose a rank-based approach. Johnson (2008a) first studied penalized rank-based estimators for censored data. Subsequently, Cai et al. (2008) provided an elegant path algorithm by applying the technique of Zhu et al. (2004); for alternative estimation, see Johnson (2008b). It is important to note that the rank-based setup is very different than the least-squares setup proposed here. In particular, the proposed one-step estimator reduces to Tibshirani's (1996) lasso as the proportion of censored observations approaches zero. The regularized Gehan estimator reduces to the regularized Wilcoxon estimator (Johnson and Peng, 2008) for uncensored data. So the method proposed here is a direct extension of Tibshirani's (1996) lasso estimator to censored data. The argument for supremacy between rank-based or least-squares methods is difficult to judge and subject to debate. An improved estimator over both proposals is the semi-parametric efficient estimator in the AFT model (Lin and Zeng, 2007). A potential difficulty with the semi-parametric efficient estimator

is that it requires smoothing the hazard function. The beauty of the proposed approach here is its simplicity.

In statistics, as in life, there is usually a trade-off for such simplicity. Although the unpenalized, one-step, synthetic data calibrated estimator (Jin et al., 2006) is a consistent estimator of the regression coefficient in model (1), it is not an efficient one. This is partially due to an inefficient initial estimator (i.e. Gehan), the approximation of the error distribution, and the Buckley-James imputation or transformation. In variable selection, inefficiency translates into larger model error. There exist other synthetic data techniques which fall under the heading "censoring unbiased transformations" (cf. Fan and Gijbels, 1996; Rubin and van der Laan, 2007, and references therein). These methods are similar to inverse weighting (Koul et al., 1981) in that they model the censoring or conditional censoring distribution given covariates, however, they are significantly more precise than the method of Koul et al. (1981). Fan and Gijbels refer to the Buckley-James transformation as the "best restoration" in the sense that it minimizes squared error loss; then, the proposed one-step estimator may be interpreted as a one-step approximation to the best restoration. Compared with censoring unbiased transformations, the method here does not require a model for the (conditional) censoring distribution and the technical requirements on the censoring distribution are the same as in Ritov (1990) and Lai and Ying (1991). At the same time, as this method was primarily developed to address concerns with the "inexactness" of coefficient estimates in the AFT model using Buckley-James estimation procedures, any one-step regularized estimator on synthetic data could claim similar improvement.

Appendix: Large Sample Theory

Regularity conditions

Throughout, we assume the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ (i = 1, \ldots, n),$$

assume the true coefficient vector $\boldsymbol{\beta}_0$ is an interior point of a compact parameter space $\mathbb{B}$, and that $(U_i, \delta_i, \mathbf{x}_i)$ are independent replicates from the underlying distribution of $(y, C, \mathbf{x})$. For completeness, we define the original, unpenalized Buckley-James estimating function $\mathbb{S}(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i(\widetilde{Y}_i(\boldsymbol{\beta}) - \mathbf{x}_i'\boldsymbol{\beta})$, and the following $d \times d$ matrices:

$$\begin{aligned}
\mathbf{A} &= \int \mathrm{Var}(\mathbf{x}|C - \mathbf{x}'\boldsymbol{\beta}_0 \geq u)\,\{u - E[\varepsilon|\varepsilon > u]\} \times \\
&\qquad \left[-\frac{\dot{f}(u)}{f(u)} + E\left\{\frac{\dot{f}(\varepsilon)}{f(\varepsilon)}|\varepsilon > u\right\}\right] P(C - \mathbf{x}'\boldsymbol{\beta}_0 \geq u)\, dF(u), \\
\mathbf{V} &= \int \mathrm{Var}(\mathbf{x}|C - \mathbf{x}'\boldsymbol{\beta}_0 \geq u)\,\{u - E[\varepsilon|\varepsilon > u]\}^2 P(C - \mathbf{x}'\boldsymbol{\beta}_0 \geq u)\, dF(u),
\end{aligned}$$

where $\dot{f}(t) = (d/dt)f(t)$, $f(t) = (d/dt)F(t)$. Finally, we adopt the following regularity conditions:

C.1. There exists a constant c_0 such that $P(U - \mathbf{x}'\boldsymbol{\beta} < c_0) < 1$ for all $\boldsymbol{\beta}$ in some neighborhood of $\boldsymbol{\beta}_0$.
C.2. The random variable $\mathbf{x}$ has compact support.
C.3. F has finite Fisher information for location.
C.4. The asymptotic slope matrix $\mathbf{A}$ is nonsingular.

It is important to note that (C.1)-(C.4) are Ritov's (1990) regularity conditions to show the existence of a root-n consistent root to the unpenalized, Buckley-James estimating equations, $0 \cong \mathbb{S}(\widehat{\boldsymbol{\beta}}_{\mathrm{BJ}})$ (where $\cong$ denotes a difference of order $o_p(n^{-1/2})$), and to demonstrate that $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{BJ}} - \boldsymbol{\beta}_0)$ converges in distribution to a mean-zero, Gaussian random vector with covariance matrix $\mathbf{A}^{-1}\mathbf{V}\mathbf{A}^{-1}$. Alternatively, we could impose Lai and Ying's (1991) technical conditions to prove Theorems 1-3.

Proof of Theorem 1

Define $\widehat{\boldsymbol{\beta}}_{(0)} = \mathrm{argmin}(\Phi_{0,n})$, where

$$\Phi_{0,n}(\mathbf{u}) = n^{-1}\sum_{i=1}^{n}(\widetilde{Y}_i(\boldsymbol{\beta}_0) - \mathbf{x}_i'\mathbf{u})^2 + \frac{\lambda_n}{n}\sum_{j=1}^{d}|u_j|. \tag{A.1}$$

Note the difference between $\Phi_{0,n}(\mathbf{u})$ and $\Phi_{1,n}(\mathbf{u})$ is that the former defines the pseudo response as a function of the true regression coefficient $\boldsymbol{\beta}_0$ while the latter uses the Gehan estimate $\widehat{\boldsymbol{\beta}}_I$. We will first show that $\widehat{\boldsymbol{\beta}}_{(0)} \to_p \mathrm{argmin}(\Phi_1)$ and then show $\widehat{\boldsymbol{\beta}}_C$ converges to the same limit.

Expanding the calibrated squared error loss in $\Phi_{0,n}(\mathbf{u})$, we have

$$n^{-1}\sum_{i=1}^{n}(\widetilde{Y}_i(\boldsymbol{\beta}_0) - \mathbf{x}_i'\mathbf{u})^2 =$$
$$n^{-1}\sum_{i=1}^{n}(\widetilde{Y}_i(\boldsymbol{\beta}_0) - \mathbf{x}_i'\boldsymbol{\beta}_0)^2 + 2n^{-1}(\boldsymbol{\beta}_0 - \mathbf{u})\mathbb{S}(\boldsymbol{\beta}_0) + (\mathbf{u} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}(\mathbf{u} - \boldsymbol{\beta}_0) \tag{A.2}$$

where $\mathbb{S}(\boldsymbol{\beta}_0)$ was defined above. Using a Taylor series expansion, integration-by-parts (cf. Lai and Ying, 1991), and the martingale representation of the Kaplan-Meier estimator (Fleming and Harrington, 1991), one can show that

$$n^{-1}\sum_{i=1}^{n}(\widetilde{Y}_i(\boldsymbol{\beta}_0) - \mathbf{x}_i'\boldsymbol{\beta}_0)^2 \cong n^{-1}\sum_{i=1}^{n}\xi_{1i} \to_p E\xi_{11}, \tag{A.3}$$

where $\xi_{11}, \ldots, \xi_{1n}$ are iid random variables with $E|\xi_{11}| < \infty$. For the second expression in (A.2), note that $n^{-1/2}\mathbb{S}(\boldsymbol{\beta}_0) \to_d N(0, \mathbf{V})$ by conditions (C.1)-(C.3) and Theorem 5.1 of Ritov (1990). Thus, as $n \to \infty$, $\{(\boldsymbol{\beta}_0 - \mathbf{u})/\sqrt{n}\} \cdot$

$n^{-1/2}\mathbb{S}(\boldsymbol{\beta}_0) \to_p 0$ because $n^{-1/2}\mathbb{S}(\boldsymbol{\beta}) = O_p(1)$. Therefore, $\Phi_{0,n}(\mathbf{u}) \to_p \Phi_1(\mathbf{u})$ for every $\mathbf{u}$. Because $\boldsymbol{\Sigma}$ is positive definite, $\Phi_1(\mathbf{u})$ has a unique minimizer. By the convexity of $\Phi_{0,n}(\mathbf{u})$ and epiconvergence (Geyer, 1994), $\widehat{\boldsymbol{\beta}}_{(0)} = \text{argmin}(\Phi_{0,n}) \to_p \text{argmin}(\Phi_1)$.

Now, replace $\boldsymbol{\beta}_0$ with the root-n consistent estimate $\widehat{\boldsymbol{\beta}}_I$ in the pseudo response $\widetilde{Y}_i(\boldsymbol{\beta}_0)$ in (A.1) and consider the expansion of the calibrated squared error loss as in (A.2). The third expression in (A.2) is the same while the first expression does not depend on $\mathbf{u}$ but, nevertheless, converges in probability to the same limit in (A.3). The second expression is

$$-2n^{-1}(\boldsymbol{\beta}_0 - \mathbf{u}) \cdot \left\{ \sum_{i=1}^{n} \mathbf{x}_i \left(\widetilde{Y}(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'\boldsymbol{\beta}_0 \right) \right\}.$$

By conditions (C.1)-(C.3), in a neighborhood of $\boldsymbol{\beta}_0$,

$$\sum_{i=1}^{n} \mathbf{x}_i \left(\widetilde{Y}(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'\boldsymbol{\beta}_0 \right) = \mathbb{S}(\boldsymbol{\beta}_0) - n\mathbf{A}(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0) + o_p(n^{1/2}). \tag{A.4}$$

Multiplying the right-hand side of (A.4) by $n^{-1/2}$, the first term is $O_p(1)$, the third term is $o_p(1)$, and the second term is equal to $\mathbf{A} \cdot \sqrt{n}(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0) = O_p(1)$. Hence, (A.4) $= O_p(n^{1/2})$ and the rest of the proof follows straightforwardly from the above paragraph. Therefore, $\widehat{\boldsymbol{\beta}}_C \to_p \text{argmin}(\Phi_1)$, as desired.

□

Proof of Theorem 2

Rewrite the random function $\Phi_{1,n}$ as

$$\eta_n(\mathbf{u}) = \sum_{i=1}^{n} [\widetilde{Y}_i(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'(\boldsymbol{\beta}_0 + \frac{\lambda_n}{n}\mathbf{u})]^2 + \sum_{j=1}^{d} |\beta_{j0} + \frac{\lambda_n}{n} u_j|.$$

Note that $\widehat{\mathbf{u}}$ minimizes the difference $\eta_n(\mathbf{u}) - \eta_n(0)$ and that the latter term does not depend on $\mathbf{u}$. Now, define $\Phi_{2,n}(\mathbf{u}) = \eta_n(\mathbf{u}) - \eta_n(0)$, where:

$$\begin{aligned} \Phi_{2,n}(\mathbf{u}) \quad = \quad & \mathbf{u}' \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right) \mathbf{u} - 2n^{-1/2} \left\{ \sum_{i=1}^{n} \mathbf{x}_i \left(\widetilde{Y}(\widehat{\boldsymbol{\beta}}_I) - \mathbf{x}_i'\boldsymbol{\beta}_0 \right) \right\} \cdot \mathbf{u} \\ & + \lambda_n \sum_{j=1}^{d} (|\beta_{j0} - n^{-1/2}u_j| - |\beta_{j0}|). \end{aligned}$$

Again, we have an expression like (A.4). Using integration by parts and the martingale representation of the Kaplan-Meier estimator, one can show that, almost surely,

$$\text{(A.4)} = \mathbb{S}(\boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0) \sum_{i=1}^{n} \int_{-\infty}^{\infty} \boldsymbol{\xi}_{2i}(t)\, dM_i(t) + o(n^{1/2} + \|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\|),$$

(Lai and Ying, 1991) where $\boldsymbol{\xi}_{2i}(t)$ are independent random vectors and

$$M_i(t) = \delta_i I\{e_i(\boldsymbol{\beta}_0) \leq t\} - \int_0^t I\{e_i(\boldsymbol{\beta}_0) \geq t\}\, \lambda(s)\, ds.$$

From Tsiatis (1990), it was shown that

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0) = n^{-1/2} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \boldsymbol{\gamma}_i(t)\, dM_i(t) + o_p(1),$$

for nonrandom functions $\boldsymbol{\gamma}_i(t)$. If $\bar{\boldsymbol{\xi}}_2(t) = \lim_n n^{-1} \sum_{i=1}^n \boldsymbol{\xi}_{2i}(t)$, then it follows that

$$(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0) \sum_{i=1}^{n} \int_{-\infty}^{\infty} \boldsymbol{\xi}_{2i}(t)\, dM_i(t) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{\bar{\boldsymbol{\xi}}_2(t) \cdot \boldsymbol{\gamma}_i(t)\right\}\, dM_i(t) + o_p(n^{1/2}).$$

Multiplying both sides by $n^{-1/2}$, we have that

$$n^{-1/2} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{\bar{\boldsymbol{\xi}}_2(t) \cdot \boldsymbol{\gamma}_i(t)\right\}\, dM_i(t) \to_d \mathbf{W}_2,$$

where, by the martingale CLT, $\mathbf{W}_2 \equiv N(0, \mathbf{D})$ and

$$\mathbf{D} = \int_{-\infty}^{\infty} E\left[\left\{\bar{\boldsymbol{\xi}}_2(t) \cdot \boldsymbol{\gamma}_i(t)\right\}^{\otimes 2} I\{e_i(\boldsymbol{\beta}) \geq t\}\right] \lambda(t)\, dt,$$

$\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$. Then, for every $\mathbf{u}$, $\Phi_{2,n}(\mathbf{u}) \to_p \Phi_2(\mathbf{u})$. The conclusion then follows from the convexity of $\Phi_{2,n}$ and the fact that Φ_2 has a unique minimizer (Geyer, 1994; Knight and Fu, 2000, Theorem 2).

□

Proof of Theorem 3

Append the alasso penalty to the estimating function $\mathbb{S}^U(\boldsymbol{\beta})$ from Jin et al. (2006) to define the following system of penalized estimating equations:

$$\Psi^C(\boldsymbol{\beta}) = \mathbb{S}^U(\boldsymbol{\beta}) - n\lambda \cdot \boldsymbol{\pi}\mathbf{s}(\boldsymbol{\beta}),$$

and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)'$. Johnson et al. (2008) recently showed that subject to two regularity conditions, carefully defined zero-crossings of $\Psi^C(\boldsymbol{\beta})$ possess the oracle property and all conclusions of Theorem 3 follow. First, we must show that

J.1. There exists a nonsingular matrix $\mathbf{A}$ such that for any given constant M,

$$\sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \leq M n^{-1/2}} |n^{-1/2}\boldsymbol{\Psi}(\boldsymbol{\beta}) - n^{-1/2}\boldsymbol{\Psi}(\boldsymbol{\beta}_0) - n^{1/2}\mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| = o_p(1).$$

Furthermore, $n^{-1/2}\boldsymbol{\Psi}(\boldsymbol{\beta}_0) \to_d N(0, \mathbf{V})$, for $\mathbf{V}$ a $d \times d$ matrix.

Condition J.1 follows from conditions (C.1)-(C.4), Theorem 5.1 of Ritov (1990), and arguments similar to the Appendix of Jin et al. (2006). Key steps include integration by parts, the martingale representation of the Kaplan-Meier estimator, and the martingale CLT. The second technical condition is:

J.2. The penalty function $q_{\lambda_n}(|\theta|) = \lambda_n \pi$ possesses the following properties:

(i) For non-zero fixed θ, $\lim n^{1/2} q_{\lambda_n}(|\theta|) = 0$ and $\lim q'_{\lambda_n}(|\theta|) = 0$;

(ii) For any $M > 0$, $\lim \sqrt{n} \inf_{|\theta| \leq M n^{-1/2}} q_{\lambda_n}(|\theta|) \to \infty$.

Condition J.2 follows from the definition of $\pi = |\theta|^{-1}$ and the root-n consistency of the Gehan estimator and is described in detail in Remark 3(c) of Johnson et al. (2008). Therefore, by Theorem 1 of Johnson et al. (2008), the calibrated one-step estimator with alasso penalty possesses the oracle property and its solution given by (9) is exact.

□

References

[1] Breiman, L. (1995). Better Subset Selection using the Nonnegative Garotte. *Technometrics*, **37**, 373–384.

[2] Buckley, J. and James, I. (1979). Linear Regression with Censored Data, *Biometrika*, **66**, 429–436.

[3] Bühlmann, P. and Hothorn, T. (2008). Boosting Algorithms: Regularization, Prediction, and Model Fitting (with Discussion), *Statistical Science*, **22**, 477–522.

[4] Cai, T., Huang, J. and Tian, L. (2008). Regularized estimation for the accelerated failure time model. *Biometrics*, (In press).

[5] Cox, D. R. (1972). Regression Models and Life-Tables (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **34**, 187–202.

[6] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, London: Chapman and Hall.

[7] Datta, S., Le-Rademacher, J. and Datta, S. (2007) Predicting Survival from Microarray Data by Accelerated Failure Time Modeling using Partial Least Squares and Lasso. *Biometrics*, **63**, 259–271.

[8] Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworth, A. (1989). "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," *Hepatology*, **10**, 1–7.

[9] Efron, B. (2005). Local False Discovery Rates, Technical Report, Department of Statistics, Stanford University.

[10] Efron, B., Hastie, T., Johnstone, I. M. and Tibshirani, R. (2004). Least Angle Regression (with discussion). *The Annals of Statistics*, **32**, 407–499.

[11] Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. London: CRC Press.

[12] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized

Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

[13] Fleming, T. A. and Harrington, D. P. (1991). *Counting Processes and Survival Analyses*. New York: Wiley.

[14] Freund, Y. and Schapire, R. (1995). A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, Springer, Berlin.

[15] Friedman, J. and Popescu, B. (2004). Gradient Directed Regularization. *Technical Report*. Stanford University.

[16] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics*, **1**, 302–332.

[17] Fu, W. J. (2003). Penalized Estimating Equations. *Biometrics*, **35**, 109–148.

[18] Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Single-Censored Samples. *Biometrika*, **90**, 341–353.

[19] Geyer, C. J. (1994). On the asymptotics of Constrained M-Estimation. *The Annals of Statistics*, **22**, 1993–2010.

[20] Huang, J., Ma, S. and Xie, H. (2006). Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. *Biometrics*, **62**, 813–820.

[21] Jin, Z., Lin, D. Y. and Ying, Z. (2006). On Least-Squares Regression with Censored Data. *Biometrika*, **93**, 147–162.

[22] Johnson, B. A. (2008a). Variable Selection in Semiparametric Linear Regression with Censored Data, *Journal of the Royal Statistical Society*, Ser. B, **70**, 351–370.

[23] Johnson, B. A. (2008b). Estimation in the ℓ_1-Regularized Accelerated Failure Time Model. Technical Report, Emory University.

[24] Johnson, B. A. and Peng, L. (2008). Rank-based Variable Selection, *Journal of Nonparametric Statistics*, Ser. B, **20**, 241–252.

[25] Johnson, B. A., Lin, D. Y. and Zeng D. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, **103**, 672–680.

[26] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed., Hoboken, Wiley.

[27] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, **28**, 1356–1378.

[28] Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression Analysis with Randomly Right-Censored Data. *The Annal of Statistics*, **9**, 1276–1288.

[29] Lai, T. L. and Ying, Z. (1991), Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. *The Annals of Statistics*, **19**, 1370–1402.

[30] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its considerations. *Cancer Chemo. Rep.*, **50**, 163–170.

[31] Miller, R. G. and Halpern, J. (1982). Regression with Censored Data.

Biometrika, **69**, 521–531.

[32] Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the Lasso and its Dual. *Journal of Computational and Graphical Statistics*, **9**, 319–337.

[33] Prentice, R. L. (1978), Linear Rank Tests with Right-Censored Data. *Biometrika*, **65**, 167–179.

[34] Reid, N. (1994). A Conversation with Sir David Cox. *Statistical Science*, **9**, 439–455.

[35] Ritov, Y. (1990). Estimation in a Linear Regression Model with Censored Data. *The Annals of Statistics*, **18**, 303–328.

[36] Rubin, D. and van der Lann, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, **3**, 2007.

[37] Schmid, M. and Hothorn, T. (2008). Flexible Boosting of Accelerated Failure Time Models. Technical Report 018-2008, Department of Statistics, University of Munich.

[38] Tibshirani, R. J. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, Ser. B, **58**, 267–288.

[39] Tibshirani, R. J. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385–395.

[40] Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society*, Ser. B, **67**, 91–108.

[41] Tsiatis, A. A. (1990). Estimating Regression Parameters using Linear Rank Tests for Censored Data. *The Annals of Statistics*, **18**, 354–372.

[42] Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer: New York.

[43] van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer: New York.

[44] Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation. *Journal of the American Statistical Association*, **102**, 1039–1048.

[45] Wang, H., Li, G. and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection through the Lad-Lasso. *Journal of Business and Economic Statistics*, **11**, 1–6.

[46] Wang, S., Nan, B., Zhu, J. and Beer, D. G. (2008). Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates. *Biometrics*, **64**, 132–140.

[47] Wang, N. and Robins, J. M. (1999). Large-sample Theory for Parametric Multiple Imputation Procedures. *Biometrika*, **85**, 935–948.

[48] Zeng, D. and Lin, DY. (2007). Efficient Estimation for the Accelerated Failure Time Model. *Journal of the American Statistical Association*, **102**, 1387–1396.

[49] Zhang, H. H. and Lu, W. (2007). Adaptive-Lasso for Cox's Proportional Hazards Model. *Biometrika*, **94**, 691–703.

[50] Zhu, J., Hastie, T., Rosset, S. and Tibshirani, R. (2004). 1-Norm Support Vector Machines. *Neural Information Processing Systems*, **16**.

[51] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418–1429.
[52] Zou, H. (2008). A Note on Path-Based Variable Selection in the Penalized Proportional Hazards Model. *Biometrika*, **95**, 241–247.
[53] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, Ser. B, **67**, 301–320.
[54] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the 'Degrees of Freedom' of the Lasso. *The Annals of Statistics*, **35**, 2173–2192.