Log-Penalized Least Squares, Iteratively Reweighted Lasso,
and Variable Selection for Censored Lifetime Medical Cost

Brent A. Johnson*
Department of Biostatistics and Bioinformatics, Emory University (bajohn3@emory.edu)

Qi Long
Department of Biostatistics and Bioinformatics, Emory University

Yijian Huang
Department of Biostatistics and Bioinformatics, Emory University

Kari Chansky
The Fred Hutchinson Cancer Research Center

Mary Redman
The Fred Hutchinson Cancer Research Center

Technical Report 2012-02
December 2012
Log-Penalized Least Squares, Iteratively Reweighted Lasso, and Variable Selection for Censored Lifetime Medical Cost

Brent A. Johnson∗ and Qi Long† and Yijian Huang

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta GA 30307 USA
e-mail: bajohn3@emory.edu; qlong@emory.edu; yhuang5@emory.edu

Kari Chansky and Mary Redman

The Fred Hutchinson Cancer Research Center
1100 Fairview Ave N
PO Box 19024
Seattle WA 98109, USA
e-mail: karic@crab.org; mredman@fhcrc.org

Abstract: Penalized least squares is a ubiquitous tool in statistics and its application are well-documented. Numerous investigators have investigated the tradeoffs between ℓ1-regularization versus non-concave penalties, the latter resulting in a sparser solution, smaller false positive rate and smaller prediction error when the signal-to-noise ratio is large. Authors later proposed one-step, convex approximations to non-concave penalties to avoid thorny numerical instabilities and showed that these approximations possessed the same asymptotic properties. Here, we propagate log-penalized least squares and a better numerical algorithm to compute the estimate. The algorithm is based on iteratively reweighted lasso, can be extended to any penalized likelihood problem, and can accommodate small or large data sets. Our numerical studies suggest that log-penalized least squares is as good and often better than other leading estimators when the signal-to-noise ratio is large.

Our substantive objective is to identify factors associated with censored medical cost, an important problem for patients and providers as well as legislators at the state and national level. Direct application of estimators for both uncensored and survival data lead to inconsistent estimation, in general, and result in spurious conclusions from the data. In this paper, we propose a new family of regularized estimators for lifetime censored medical cost through an extension of penalized estimating functions. Then, we show that a version of our estimator may be written as the minimizer of a convex surrogate loss function. We investigate the operating characteristics of these procedures through simulation studies and application to lifetime medical cost data due to nonsmall cell lung cancer from the Southwest Oncology Group Protocol 9509.

∗Supported in part by the National Institutes of Health grant P30 AI050409
†Supported in part by the National Institutes of Health PHS grant UL1 RR025008
Keywords and phrases: Dependent censoring, Marked point process, Regularization, Survival analysis

1. Introduction

The federal government, insurance agencies and insured individuals all have an interest in knowing what factors are associated with increased medical cost. If medical expenditures could be projected reliably, individuals could use such resource for planning purposes. This is our principal motivation in modeling censored lifetime medical cost as a function of clinical predictors in a Southwest Oncology Group (SWOG) clinical trial for treatment of nonsmall cell lung cancer. Modeling medical costs is also important in econometrics and health policy and management. Insurers use statistical models to predict cost and set insurance premiums based on subject characteristics. In March 2010, President Barack Obama signed into law, the Affordable Care Act, which stated as one of its objectives, “lower[ing] health care costs over the lifetime” of an individual (www.whitehouse.gov). Thus, modeling medical costs is an important problem in health policy, econometrics, and subject-level cost management.

Regularized estimation will serve as the cornerstone of our variable selection procedure in Section 2. Because there has been so much theoretical and numerical work in this arena over the past several years, a brief summary of the relevant literature and how the current paper fits into that literature is imperative. To this end, we begin with estimation in the linear regression model for an uncensored outcome $Y$,

\[ Y = z_1 \beta_1 + z_2 \beta_2 + \cdots + z_d \beta_d + \varepsilon_Y \]  

(1.1)

where $z = (z_1, \ldots, z_d)^T$ are regressors, $\beta = (\beta_1, \ldots, \beta_d)^T$ are regression coefficients and $\varepsilon_Y$ are mean-zero stochastic errors. The observed data are $\{(Y_i, z_i)\}_{i=1}^n$ and assumed to be a random sample of size $n$ from model (1.1). Coefficient estimation by subset selection may be written as the $\ell_0$-optimization,

\[ \min_{\beta \in \mathbb{R}^d} \frac{1}{2} \left\| Y - Z \beta \right\|_{\ell_2}^2 + \lambda \left\| \beta \right\|_{\ell_0}, \]

(1.2)

where $\ell_0$-norm is the number of non-zero elements of the coefficient vector $\beta$. Tibshirani’s (1996) least absolute shrinkage and selection operator (lasso) is the tightest convex relaxation of (1.2) and the $\ell_1$-regularized analog of (1.2). However, it is known that lasso has a tendency to overfit (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) and shrink coefficient estimates too strongly toward zero when the signal-to-noise ratio is large (Fan and Li, 2001; Zou, 2006; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). Fan and Li (2001) were the first to propose non-concave penalized least squares estimation through their smoothly clipped absolute deviation (scad) penalty and showed that this new estimator was root-$n$ consistent, would select the correct submodel with high probability, and converged to the same asymptotic distribution as if one had
known the correct submodel \textit{a priori}. Zou (2006) showed that if one weighted the $\ell_1$-regularization in Tibshirani’s lasso with weights proportional to the inverse of the absolute value of the regression coefficients from a preliminary least squares fit, then the weighted lasso (i.e. the adaptive lasso) could achieve the same asymptotic properties as scad. But, unlike scad, adaptive lasso was a convex optimization problem, could make use of existing algorithms to solve the original lasso problem, and may be preferable on numerical grounds. Although these two methods are popular in the statistics literature, there exist other lesser-known, less mature methods with the same asymptotic properties but better finite sample properties.

The focus of this paper is simultaneous coefficient estimation and variable selection via log-penalized least squares, i.e.

$$
\min_{\beta \in \mathbb{R}^d} \frac{1}{2} \| Y - Z\beta \|_{\ell_2}^2 + \lambda \sum_{j=1}^d \log(|\beta_j|). 
$$

Log-penalized regression has received scant attention in the statistics literature even though it results in a sparse solution and can be adapted for variable selection. Logarithmic penalty is related to the smoothly clipped absolute deviation (scad; Fan and Li, 2001) in that they are both non-concave penalty functions and offer better approximations to cardinality than does the $\ell_1$-norm. In Figure 1, we illustrate the scad and logarithmic penalty. Compared with scad, logarithmic penalty offers a closer approximation to cardinality. Then, heuristically,

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Common penalty functions and their proximity to cardinality. Scad (solid blue) penalty in the left panel and logarithmic (solid blue) penalty in the right panel. Cardinality is represented by a solid vertical line.}
\end{figure}
one might believe that logarithmic penalty also offers a closer approximation to subset selection.

By now, a couple authors have discussed asymptotic properties of log-penalized regression estimators. Johnson, Lin and Zeng (2008) discuss log-penalized regression in the context of penalized estimating functions and show that it, along with several other common penalty functions, possess an oracle property. Zou and Li (2008) show that a one-step approximation to the log-penalized estimator is Zou’s (2006) adaptive lasso estimator which also possesses an oracle property. Despite optimal asymptotic properties, statisticians have found minimizing (1.3) awkward and tricky. Zou and Li (2008) offer some insight into the operational difficulty of logarithmic penalty, \( \ell_{0.01} \)-penalty, and bridge penalty. In Figure 2, adapted from Zou and Li (2008), we see that logarithmic penalty results in a discontinuous thresholding rule for orthonormal designs. Fan and Li (2001) recommend avoiding penalty functions, e.g. logarithmic penalty, that break the following rule (their rule no. 3) of a “good” penalty function:

“3. Continuity: The resulting estimator is continuous in data to avoid instability in model prediction.” (Fan and Li, 2001, p. 1349)

From Figure 2, it is clear that logarithmic penalty results in a discontinuous thresholding rule whereas scad and adaptive lasso result in continuous thresholding rules. In addition, bridge penalty results in a discontinuous thresholding rule for any power less than one (cf. Fan and Li, 2001). The impetus of rule no. 3 is numerical inaccuracy near the boundary constraints so that when one applies the “usual” numerical tricks (Tibshirani, 1996; Fan and Li, 2001; Hunter and Li, 2005; Zou and Li, 2008) to logarithmic or bridge penalty, the quadratic or linear approximation is either too crude or the algorithm simply does not converge (i.e. infinite oscillation, multiple roots).
In this paper, we extend a recent algorithm proposed by authors in the compressed sensing and signal processing literature (Candes, Wakin and Boyd, 2008; Daubechies et al., 2009; Gasso, Rakotomamonjy and Canu, 2009) that can be used to overcome the challenges in minimizing (1.3) and open the door to other complex optimization problems. The numerical algorithm amounts to iteratively reweighted lasso (Gasso, Rakotomamonjy and Canu, 2009) and the technique can be used to manipulate any discontinuous thresholding rule with a singularity at the origin into a stable variable selection procedure. This recent development debunks the myth in statistics that penalties resulting in discontinuous thresholding rules ought to be avoided.

Our contributions to the literature are three-fold: (a) to advocate and propagate an iteratively reweighted lasso procedure (Gasso, Rakotomamonjy and Canu, 2009) for log-penalized regression in large- or small-scale problems, (b) to extend this technique from penalized least squares to a semi-parametric estimator for a problem with induced dependent censoring, and (c) to apply the technique to censored lifetime medical cost data from SWOG 9509. Lifetime medical cost is an example of a marked outcome, an endpoint observed at the time of failure but not when the failure event is censored. Although the statistical analysis of censored medical cost has received substantial attention in the literature (cf. Lin, Feuer and Wax, 1997; Huang and Louis, 1998; Bang and Tsiatis, 2000; Huang, 2002; Jain and Strawderman, 2002; Huang and Lovato, 2002), there is only one other paper on variable selection for censored medical cost data. Approximately one decade ago, Jain and Strawderman (2002) modeled the log hazard function of censored lifetime cost as a function of covariates using piecewise linear splines and their tensor products. They employed inverse weighted estimating equations to account for the induced informative censoring and introduced a stepwise deletion procedure to add or delete basis functions in their model so as to improve the finite sample properties of their estimator. Here, rather than build on the inverse weighting framework, our family of estimation procedures is built on the class of weighted logrank estimators via Huang’s (2002) calibration regression for censored lifetime medical cost. In addition, the family of \( \ell_1 \)-regularized coefficient estimators proposed here is quite different than stepwise deletion.

We perform simulation studies in Section 5 to evaluate the performance of several related procedures in the \( n > d \) paradigm, which is the paradigm of interest for our censored medical cost application. Because our literature review is not specific to censored data and we report results from simulation studies for both censored and uncensored data, our conclusions may appeal an audience beyond the specific medical cost application. The conclusions are as follows:

- Iteratively reweighted \( \ell_1 \) minimization is different than one-step approximations and our numerical studies indicate that estimates computed by iteratively reweighting leads to a smaller false positive error rate and a smaller prediction error;
- Even when the asymptotic distribution among “oracle” procedures is the same, not all such procedures perform equally well in finite samples. We
found that bridge and logarithmic penalty performed somewhat better than scad and adaptive lasso among penalized least squares estimators in complete, uncensored data simulations. Our findings in the $n > d$ paradigm agree with and complement other numerical studies in the signal processing and compressed sensing literature for the high-dimensional case (cf. Candes, Wakin and Boyd, 2008; Daubechies et al., 2009; Gasso, Rakotomamonjy and Canu, 2009);

- Weighting $\ell_1$ minimizers are powerful tools that many authors have explored before us, but some of these authors have taken unfortunate turns leading to suboptimal weights. We observed that when the optimal weights are used, the weighted lasso performs as well as or better than the Oracle, on average. Poor weighting schemes can lead to estimators that perform worse than the unweighted $\ell_1$-type estimator and, hence, have an undesirable effect;

- In the presence of marked outcome data, the trends were generally the same as in the uncensored data case but with bridge regression (i.e. $\ell_{0.5}$-regression) performing slightly better than its competitors.

2. Methods for Censored Lifetime Medical Cost

2.1. Notation

For $i = 1, \ldots, n$, let $Y_i$ be a mark variable of interest observed at the log-transformed event time $T_i$, and $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})^T$ is a $d$-vector of independent regressors. Write the joint regression model,

$$
Y_i = \sum_{j=1}^{d} z_{ij} \beta_j + \varepsilon_{Y,i}, \quad T_i = \sum_{j=1}^{d} z_{ij} \vartheta_j + \varepsilon_{T,i},
$$

where the unknown parameters include the regression coefficients on the mark scale, $\beta = (\beta_1, \ldots, \beta_d)^T$, as well as regression coefficients on the time scale, $\vartheta = (\vartheta_1, \ldots, \vartheta_d)^T$. We assume that the bivariate random error vector $\varepsilon_i = (\varepsilon_{Y,i}, \varepsilon_{T,i})^T$ is independent and identically distributed according to an unspecified bivariate distribution function. In mark variable regression, the observed data are $\{\tilde{Y}_i, \tilde{T}_i, \delta_i, \mathbf{z}_i\}_{i=1}^{n}$, where $\tilde{Y}_i = Y_i \cdot I(T_i \leq C_i)$, $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, for a random censoring variable $C_i$, e.g. administrative censoring at the end of clinical study. For some applications, if may be reasonable to assume that the mark $Y_i$ is linearly related to regressors exactly as it is written in (2.1). However, because of the right-skewed nature of patient medical cost data, we assert a log-linear relationship such that $Y_i$ in (2.1) pertains to transformed medical cost on natural logarithmic scale.

2.2. Coefficient Estimation

Define the at-risk process $R_i(u, \vartheta) = I(\tilde{T}_i - \mathbf{z}_i^T \vartheta \geq u)$ and counting process $N_i^{\tilde{T}}(u, \vartheta) = I(\tilde{T}_i - \mathbf{z}_i^T \vartheta \leq u, \delta_i = 1)$. Let $\Lambda^{\tilde{T}}(\cdot)$ be the cumulative baseline
hazard function of $\varepsilon_T$ in (2.1). Following standard martingale theory, we have that

$$M^\varepsilon_T(t) = N^\varepsilon_T(t, \vartheta) - \int_{-\infty}^t R_i(s, \vartheta) \, d\Lambda^\varepsilon_T(s),$$

is a local martingale with respect to the filtration

$$\mathcal{F}_t = \sigma\left\{ I(\tilde{T}_i - z^T_i \vartheta \leq u, \delta_i = 1), I(\tilde{T}_i - z^T_i \vartheta \leq u, \delta_i = 0), \right.$$\nonumber

$$
Y_i \cdot I(\tilde{T}_i - z^T_i \vartheta \leq u, \delta_i = 1), z_i, u \leq t, i = 1, \ldots, n \}.$$

The filtration $\mathcal{F}_t$ contains all the survival information up to and including time $t$, and we can appeal to classic counting process theory (cf. Andersen et al., 1993) to construct a consistent estimator of the regression coefficient vector $\vartheta$ on the time-scale. In particular, the usual weighted logrank estimating function (Tsiatis, 1990; Wei, Ying and Lin, 1990) for $\vartheta$ is

$$S^\varepsilon_T(\vartheta) = \sum_{i=1}^n \int_{-\infty}^\infty W^\varepsilon_T(u, \vartheta) \{ z_i - \bar{z}(u, \vartheta) \} \, dN^\varepsilon_T(u, \vartheta),$$

where $W^\varepsilon_T(t, \vartheta)$ is a $\mathcal{F}_t$-predictable, non-negative weight function and the weighted average of regressors, $\bar{z}(u, \vartheta) = \sum_j z_j R_j(u, \vartheta) / \sum_j R_j(u, \vartheta)$. We define the marked process as the product,

$$N^{\varepsilon_T}_i(t, \vartheta) = N^\varepsilon_T(t, \vartheta) \psi(Y_i - z^T_i \beta), \quad (2.2)$$

where $\psi(\cdot)$ is a known continuous and strictly monotone (Huang, 2002). This leads to Huang’s estimator for $\beta$, that is, the solution to $0 = S^\varepsilon_T(\vartheta)$, where

$$S^\varepsilon_T(\vartheta) = \sum_{i=1}^n \int_{-\infty}^\infty W^\varepsilon_T(u, \vartheta) \{ z_i - \bar{z}(u, \vartheta) \} \, dN^{\varepsilon_T}_i(u, \vartheta). \quad (2.3)$$

Note, that estimating function $S^\varepsilon_T(\vartheta)$ in (2.3) looks very much like the weighted logrank estimating function $S^\varepsilon_T(\vartheta)$ but there are some important differences. First, the weight function $W^\varepsilon_T(u, \vartheta)$ can be a function of the time- and mark-scale parameters. Second, the jump-size of the marked process $N^{\varepsilon_T}_i(u, \vartheta)$ is stochastic unlike standard applications of counting process theory for survival analysis (cf. Kalbfleisch and Prentice, 2002). Thus, in general, solving $0 = S^\varepsilon_T(\vartheta)$ can be a difficult task numerically due to choice of weight functions $W^\varepsilon_T$ and $\psi$.

### 2.3. Local Least Squares

To induce sparsity in coefficient estimation, one possible approach is to augment the estimating function $S^\varepsilon_T(\vartheta)$ with a penalty function that imposes a jump
discontinuity at zero (cf. Fu, 2003; Johnson, Lin and Zeng, 2008). However, a particular version of the score $S_{cy}(\theta)$ will have better numerical properties. In the special case where $\psi(t) = t$ in (2.2) and $W_{cy}(u, \theta)$ is not a function of $\beta$, the estimating function $S_{cy}(\theta)$ simplifies significantly. Write the simplified weight function in (2.3) as $W_{cy}(u, \theta) = W_{cy}(u, \vartheta)$, because we still assume the weight function may depend on the time-scale coefficient vector $\vartheta$. In this case, solving the estimating equation, $0 = S_{cy}(\theta)$, is equivalent to solving the linear system,

$$M(\vartheta)\beta = v(\vartheta),$$

where

$$v(\vartheta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} W_{cy}(u, \vartheta) Y_i \{z_i - \bar{z}(u, \vartheta)\} \, dN_{i}^{\varepsilon T}(u, \vartheta),$$

$$M(\vartheta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} W_{cy}(u, \vartheta) \{z_i - \bar{z}(u, \vartheta)\} \otimes^2 dN_{i}^{\varepsilon T}(u, \vartheta),$$

where $\otimes^2 = vv^T$. Indeed, Huang’s (2002) original estimator is defined:

$$\hat{\beta}_{\text{Full}} = \{M(\vartheta^*)\}^{-1}v(\vartheta^*),$$

where $\vartheta^*$ is a $p$-dimensional estimate of the time-scale regression coefficients $\vartheta$. In Huang (2002) and in this paper, $\vartheta^*$ is assumed to be root-$n$ consistent and computed as the solution to the weighted log rank estimating equations on the time-scale, $0 = S_{cy}(\vartheta)$; see Section 6 for comments on estimation in high dimensions. To define our regularized estimator, we rewrite the solution to the linear system (2.4) as the minimum of squared errors. Define the vector $V(\vartheta) = \{A^T(\vartheta)\}^{-1}v(\vartheta)$ where $A(\vartheta)$ is the Choleski decomposition of $M(\vartheta)$, i.e. $A^T(\vartheta)A(\vartheta) = M(\vartheta)$. Then, our log-penalized coefficient estimator is defined as the minimizer

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^d} L(\beta, \vartheta^*) + \lambda \sum_{j=1}^{d} \log(|\beta_j|),$$

with surrogate loss function,

$$L(\beta, \vartheta^*) = \frac{1}{2} \{V(\vartheta^*) - A(\vartheta^*)\beta\}^T \{V(\vartheta^*) - A(\vartheta^*)\beta\}.$$

The asymptotic properties of the log-penalized estimator $\hat{\beta}$ are described in a Web Appendix. In Section 3, we show how to effectively solve for logarithmic penalty through iteratively reweighted lasso.

3. Computational Methods

3.1. Fundamental Tools from Compressed Sensing

Over the last three decades, there has been a largely independent and parallel interest among engineers, computer scientists, and mathematicians in solving...
an underdetermined linear systems of equations, say,
\[ \mathbf{Z}\beta = \mathbf{Y}, \]  
(3.1)
where \( \mathbf{Y} \in \mathbb{R}^n \), \( \mathbf{Z} \) is an \( n \times d \) matrix with \( n < d \) and \( \beta \in \mathbb{R}^d \) are parameters to be estimated. Because the system is underdetermined, there are infinitely solutions that satisfy (3.1). Assuming that the true vector of unknowns \( \beta \) is itself sparse leads one to believe the \( \ell_1 \) minimization approach could be a reasonable solution to (3.1), say
\[
\min_{\beta \in \mathbb{R}^d} \| \beta \|_{\ell_1} \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{Z}\beta. 
\]
(3.2)
Recently, Daubechies et al. (2009) outlined the historical development of the \( \ell_1 \) minimization solution to (3.1) and traced the earliest applications back to the late-1970s although the first theoretical justification did not come until the late-1980s. Recently, Candes, Wakin and Boyd (2008) offered the following improvement:
\[
\min_{\beta \in \mathbb{R}^d} \| W\beta \|_{\ell_1} \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{Z}\beta, 
\]
(3.3)
where \( W = \text{diag}\{w_1, \ldots, w_d\} \) and
\[
w_j = \begin{cases} 
\frac{|\beta_0|}{\beta_j}, & j \in \mathcal{A}, \\
\infty, & j \in \mathcal{A}^c,
\end{cases}
\]
(3.4)
where the active set is defined \( \mathcal{A} = \{j|\beta_0 \neq 0, j = 1, \ldots, d\} \). However, because the optimal weights are unknown, (3.3) cannot be solved directly. So, Candes, Wakin and Boyd (2008) offered the following approximation. By adding a small error, one can rewrite the optimization problem in (3.3) as
\[
\min_{\beta \in \mathbb{R}^d} \sum_{j=1}^d \log(|\beta_j| + \epsilon) \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{Z}\beta, 
\]
which can then be iteratively solved. Given current iterates \( \beta_j^{(k-1)} \), at the \( k \)-th iteration, one solves the weighted \( \ell_1 \) minimization,
\[
\min_{\beta \in \mathbb{R}^d} \sum_{j=1}^d \left( \frac{1}{|\beta_j^{(k-1)}| + \epsilon} \right) |\beta_j| \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{Z}\beta. 
\]
(3.5)
The rationale behind (3.5) is that because the logarithm function is concave, one can improve on a guess to the solution by locally minimizing a linearized function centered at the current guess and repeating the process (Candes, Wakin and Boyd, 2008). This is the essence of MM algorithms which have a long history of application in computer science, statistics and applied mathematics. In a similar spirit, Daubechies et al. (2009) suggested iteratively reweighting \( \ell_2 \) minimization for sparse recovery.
3.2. Iteratively Reweighted Lasso

Now, we apply the principles embodied in Candes, Wakin and Boyd (2008) to penalized least squares. Gasso, Rakotomamonjy and Canu (2009) published a variation of this algorithm first under the title of “DC programming,” where “DC” stands for difference in convex functions. Rewrite the log-penalized estimator in (1.3) as the perturbed expression

$$\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2} \left\| Y - Z\beta \right\|_2^2 + \lambda \sum_{j=1}^{d} \log(|\beta_j| + \epsilon) \right\}. \tag{3.6}$$

Then, taking a first-order Taylor-series approximation of the logarithm penalty about the current value $\beta_j = \beta_j^*$, we obtain the over-approximation of (3.6),

$$\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2} \left\| Y - Z\beta \right\|_2^2 + \lambda \sum_{j=1}^{d} \left( \frac{1}{|\beta_j^*| + \epsilon} \right) |\beta_j| \right\}. \tag{3.7}$$

Note, the linear approximation in (3.7) is analogous to one proposed in Zou and Li (2008) except that Zou and Li (2008) did not make use of the perturbation $\epsilon$ and they argued strongly in favor of a one-step approximation. Instead, we offer the following algorithm to generate the entire coefficient path for log-penalized least squares. Assume all the covariates have been standardized to have mean zero and unit variance. Without loss of generality, the span of regularization parameters is $\lambda_{\text{min}} = \lambda_0 < \cdots < \lambda_L = \lambda_{\text{max}}$, for $l = 0, \ldots, L$ (cf. Friedman et al., 2007).

Algorithm 1.

(A.1) Start with $\lambda_{\text{max}} = \max_{1 \leq j \leq d} |z_j^T Y|/n$. Set $l = L$ and $\beta_{(l)} = 0$.

OUTER LOOP:

(B.0) Set $t = 0$ and $\overline{\beta}^{[0]} = \beta_{(l)}$;

(B.1) Increment $t = t + 1$ and $\overline{\beta}^{[t+1]} = \overline{\beta}^{[t]}$;

(B.2) Update the weights: $\overline{w}_j = 1/(|\overline{\beta}^{[t]}_j| + \epsilon)$, $j = 1, \ldots, d$;

INNER LOOP: Solve the Karush-Kuhn-Tucker (KKT) conditions for fixed $\overline{w}_j$:

$$z_j^T (Y - Z\beta) - \lambda_t \overline{w}_j \text{sgn}(\beta_j) = 0, \quad \text{if } j \in A,$$

$$|z_j^T (Y - Z\beta)| < \lambda_t \overline{w}_j, \quad \text{if } j \notin A,$$

(B.3) Goto step (B.1);

(B.4) Repeat steps (B.1)–(B.3) until $\| \overline{\beta}^{[t+1]} - \overline{\beta}^{[t]} \|_{\ell_\infty} < \tau$. The estimate $\beta_{(l)}$ is the limit point of the outer loop, $\overline{\beta}^{[\infty]}$;
(A.2) Decrement \( l = l - 1 \) and \( \lambda_l = \lambda_{l-1} \). Return to (B.0) using \( \beta_{(l)} \) as a warm start.

In small-scale problems where \( n > d \), we set \( \lambda_{\text{min}} = 0 \). However, when \( n < d \), one attains the saturated model long before the regularization parameter reaches zero. In this case, we suggest \( \lambda_{\text{min}} = 10^{-4} \lambda_{\text{max}} \) (cf. Friedman et al., 2007). We used \( \tau = \epsilon = 10^{-3} \) as in Gasso, Rakotomamonjy and Canu (2009).

Remark 1. Algorithm 1 is very much like majorization-minorization (MM) algorithms, a generalization of EM algorithms for missing data problems and a staple of modern statistical computation. Hunter and Li (2005) were the first to advocate MM algorithms for variable selection when \( n > d \) while Candès, Wakin and Boyd (2008) was the first to suggest an MM algorithm for the logarithmic penalty regardless of the dimension of \( d \) relative to the sample size \( n \) (see also, Gasso, Rakotomamonjy and Canu, 2009; Daubechies et al., 2009). A subtle but important point is to notice how Algorithm 1 differs from the algorithm described in Hunter and Li (2005). The use of MM algorithms in Hunter and Li (2005) is to justify a quadratic over-approximation to penalty functions with singularities at the origin. In the inner loop of Algorithm 1, we avoid such approximation by solving the KKT conditions precisely and efficiently (Osborne, Presnell and Turlach, 2000; Efron et al., 2004). Our use of MM algorithms in Algorithm 1 is to justify the local linear approximation to logarithmic penalty in the outer loop.

Remark 2. Local linear approximations to logarithmic and bridge penalty were also proposed by Zou and Li (2008). Algorithm 1 differs from that proposed by Zou and Li (2008) in that it constructs the entire coefficient path, even when \( n < d \), whereas the procedure by Zou and Li (2008) computes the coefficient estimates for a fixed regularization parameter \( \lambda \) starting with a root-\( n \) consistent estimator of the true coefficient \( \beta_0 \) at the initial step. Moreover, even with the same initial estimate, the limit point from Algorithm 1 will differ from Zou and Li (2008) after one iteration.

Remark 3. In the more general case where \( n < d \), we used coordinate-wise optimization (cf. Friedman et al., 2007) to compute the entire regularized coefficient path via Algorithm 1. Gasso, Rakotomamonjy and Canu (2009) also suggested that one could adopt coordinate-wise optimization but, in the end, they settled on another algorithm. Our experience is that coordinate-wise optimization works well in practice and converges very quickly.

Remark 4. In the case of marked outcomes, Algorithm 1 may be applied using the pseudo-response vector \( V(\vartheta_*) \) and pseudo-design matrix \( A(\vartheta_*) \). All other algorithmic details remain unchanged.

Although there is no theory guaranteeing convergence of MM algorithms, Gasso, Rakotomamonjy and Canu (2009) showed convergence for penalty func-
tions that could be represented as a difference of two convex functions, a result which applies to logarithmic and bridge penalty. We tested Algorithm 1 in several simulated data sets for $n > d$ and, in our experience, it always converged in these cases. Other authors (Candes, Wakin and Boyd, 2008; Gasso, Rakotomamonjy and Canu, 2009; Daubechies et al., 2009) have tested the high-dimensional log-penalized estimator and we refer the interested reader to elsewhere in the literature for a discussion of this topic.

3.3. Parameter Tuning

For tuning the regularization parameter in a least squares framework for uncensored outcome data, we used an ordinary BIC statistic,

$$
\text{BIC}(\lambda) = \frac{1}{n} \| Y - Z\hat{\beta} \|_2^2 + \log(n)\hat{d}(\lambda),
$$

where $\hat{d}(\lambda)$ is the number of non-zero coefficient estimates in $\hat{\beta}$. For parameter tuning in a marked outcome framework, we used a type of dispersion statistic. Let $\Sigma_c(\theta) = (\Sigma_{Y_i}^T, \Sigma_{\epsilon_i}^T)^T$ and $\hat{\beta}$ be a regularized estimate as a result of the procedure described in Algorithm 1. Then, under Conditions A-E in Huang (2002, Appendix), $n^{-1/2}\Sigma_c(\theta_0)$ converges in distribution to a mean-zero, normal random variable with covariance $\Omega(\theta_0)$, with

$$
\Omega(\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \frac{W_{e_i}^T(u, \theta)\psi(Y_i - z_i^T\beta)}{W_{e_i}^T(u, \theta)} \right\} \otimes \left[ \frac{\sum_{j=1}^{n} z_j^T R_j(u, \theta)}{\sum_{j=1}^{n} R_j(u, \theta)} - \{\tilde{z}(u, \theta)\} \right] dN_{e_i}(u, \theta),
$$

where $A \otimes B$ is the kronecker product of matrices $A$ and $B$. If we let $\hat{\Omega}(\theta)$ denote the empirical estimator for $\Omega(\theta)$ and $\hat{\theta}_{\text{Full}} = (\hat{\beta}_{\text{Full}}^T, \hat{\vartheta}_{\text{Full}}^T)^T$, then our dispersion statistic is defined as

$$
X(\lambda) = n^{-1} \left( \begin{array}{c} \Sigma_{e_i} \left( \hat{\beta}_{\lambda}, \vartheta_{\lambda} \right) \\ \Sigma_{e_i} \left( \vartheta_{\lambda} \right) \end{array} \right)^T \hat{\Omega}^{-1} \left( \begin{array}{c} \Sigma_{e_i} \left( \hat{\beta}_{\text{Full}} \right) \\ \Sigma_{e_i} \left( \vartheta_{\text{Full}} \right) \end{array} \right),
$$

which, under suitable regularity assumptions (cf. Wei, Ying and Lin, 1990), $X(\lambda)$ follows a $\chi^2$ distribution with degrees of freedom equal to the number of non-zero coefficient estimates in $(\hat{\beta}_{\lambda}, \vartheta_{\lambda})$. Then, an BIC-type criterion can be defined

$$
\text{BIC}(\lambda) = X(\lambda) + \log(n)\hat{d}(\lambda),
$$

where $\hat{d}(\lambda)$ is an estimate of the degrees of freedom for the fitted model; here, we approximate $\hat{d}(\lambda)$ with the cardinality of the estimated active set on the mark-scale. We define an analogous AIC-type criterion by substituting $\log(n)$ on the right-hand side of the above expression with two. We note that Jain and Strawderman (2002) define a similar statistic to $X(\lambda)$ in their stepwise deletion procedure for modeling the hazard of censored medical cost.
4. Analysis of the SWOG 9509 Data

The randomized Southwest Oncology Group (SWOG) 9509 trial was designed to investigate Paclitaxel plus Carboplatin versus Vinorelbine plus Cisplatin therapies in untreated patients with advanced nonsmall cell lung cancer. The primary study endpoint was survival time (Kelly et al., 2001) and subsequent secondary analyses considered lifetime medical costs (Huang and Lovato, 2002; Huang, 2002). For each of 408 eligible study participants, the “lifetime medical cost” endpoint was computed from resource utilization metrics, including medications, medical procedures, different treatments on- and off-protocol, and days spent in the outpatient or inpatient clinic. The cost incurred for each type of resource used was computed using national databases and were standardized to 1998 US dollars (Huang, 2002). Resource utilization was measured at 3, 6, 12, 18, and 24 month clinic visits. Both time and cost are modeled on the natural logarithmic scale.

We considered 18 baseline variables as main effects in the joint regression model (2.1). The regressors are treatment arm (trt, 1=Paclitaxel plus Carboplatin, else 0), gender (sex), progression status (progstat), SWOG performance status (ps), clinical stage (stage), IIIB by pleural effusion, weight (kg), height (cm), creatinine (creat), albumin (g/dl), calcium (mg/dl), serum lactate dehydrogenase (ldh, U/l), alkaline phosphatase (alkptase), bilirubin (mg/dl), white blood cell count (wbc, cells/microliter), platelet count (platelet, cells/microliter), hemoglobin level (hgb, g/dl), and age (years). Serum lactate dehydrogenase (ldh), alkaline phosphatase (alkptase), and bilirubin are all derived binary random variables, with one indicating that the patient’s measurement exceeded the upper limit of normal (ULN).

The sample size of our cost data was 398 participants and excluded 10 study participants with insufficient documentation. This cost data set was merged with the data set of baseline variables. Then, we excluded all those patients missing any of the 18 covariates. This led to a final sample size of \( n = 343 \). As a final step of preprocessing, each of 18 baseline variables was standardized to have mean zero and unit variance. The scale of the coefficients has no effect on model selection but facilitates displaying coefficient estimates for regressors with measurement scales of varying order. The model selection procedure was performed for each of the Gehan and logrank weight function, with the same weight function used in the time- and mark-scale score function. The results of our model selection procedures are presented in Table 1.

The results in Table 1 lead to several interesting findings. First, the selected variables in the Gehan and logrank weights across lasso and log regularizations, generally agree although the magnitude of the coefficients in the active sets differ. The seven strongest predictors are treatment, progression status, SWOG performance score, albumin level, lactate dehydrogenase, white blood cell count, and hemoglobin level. No fewer than five variables are selected by AIC but ignored by BIC using the same weight-penalty combination. Not surprisingly, we find that progression status is the most informative predictor of higher medical cost followed by white blood cell count. It is somewhat interesting to observe
how treatment, an unimportant variable on the time-scale as reported in Kelly et al. (2001), is moderately important on the cost-scale. The relative importance of treatment differs by penalty and weight combination with Gehan weight inflating its importance while log-penalty, logrank weight downweighting its importance. Nevertheless, the treatment variable underscores the dual roles that covariates play in joint models and that it is impossible to determine which covariates will be important on the cost-scale based solely on their prognostic value for the survival endpoint.

5. Simulation Studies

In our simulation studies, we present results from nine regularized estimators, in addition to the full model estimator and the oracle. The nine regularized estimators include familiar procedures such as lasso (Tibshirani, 1996), hard thresholding, scad (Fan and Li, 2001), bridge regression ($\ell_0.5$; Frank and Friedman, 1993), and adaptive lasso (AL(Z); Zou, 2006). We also implement two other versions of adaptive lasso. We include the adaptive lasso estimator by (AL(H); Huang, Ma and Zhang, 2008) which estimates fixed weights through univariate simple linear regressions for each covariate one-at-a-time. The optimal adaptive lasso (AL(Opt)) uses fixed weights $w_j = 1/(|\beta_0| + \epsilon)$. Finally, we also include the multi-stage adaptive lasso (MSAL; Bühlmann and Meier, 2008) which iteratively reweights lasso but re-tunes the procedure at every stage.

5.1. Least Squares

In the first simulation set of simulations, we considered penalized least squares estimators with no censoring whatsoever. The results from these studies allow us a cleaner comparison of various methods without artifacts of induced censoring and the marked outcome framework. We simulate outcomes $Y$ according to a linear model with mean-zero normal errors and variance $\sigma^2$. The covariate vector $z$ is multivariate normal with mean zero, unit variance, and correlation $\text{corr}(z_j, z_k) = \rho^{1-|j-k|}$. The data are generated independently for $n = 85$ samples. There are a total of $d = 40$ covariates, 22 of which have zero coefficient and no effect on outcome. The 18 non-zero coefficients are as follows: “3” in positions 16, 17, 18; “2” in positions 6, 23, 24, 25, 32, 36; “3/2” in position 2; “1” in positions 1, 10, 11, 12, 27, 28, 33, 35. The following four statistics are used to compare the various procedures: $\ell_2$ error, median model error (ME), $\text{ME} = (\hat{\beta} - \beta_0)^T \text{E}(zz^T)(\hat{\beta} - \beta_0)$, average number of non-zero coefficients incorrectly set to zero (false negative), and average number of coefficients whose true value is zero but whose estimate is non-zero (false positive). Results from 100 Monte Carlo are displayed in Table 2 for $\sigma^2 \in \{2, 9\}$ and $\rho \in \{0.25, 0.75\}$. To compare with existing procedures in the literature based on MM algorithms (Hunter and Li, 2005), the algorithm for log-penalized least squares is also based on a nested MM algorithm which computes estimator for a fixed regularization
parameter rather than the entire solution path. The range of regularization parameters is 19 values in the interval $[10^{-4}, 1.6] \times \hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is the root mean-squared error based on the full model. The values are computed on a log-scale and then transformed back; an identical set of values is used for scad, bridge, hard, log, MSAL, and all adaptive lasso estimators. All regularized estimators were tuned using BIC criteria.

The results in Table 2 tell several interesting stories. First, logarithm penalty has the smallest false positive (FP) rate in three out of four scenarios among estimators that can be computed using the observed data (thus, ruling out AL(Opt)). In the one instance where log does not have the smallest FP rate, bridge regression has the smallest FP error rate (i.e. when $\sigma^2 = 9$ and $\rho = 0.75$). In general, lasso has the smallest false negative (FN) rate meaning that if lasso fails to include a variable in the final model, the corresponding true regression coefficient is zero with high probability. Interestingly, when the data are highly correlated, AL(H) has a false negative rate similar to lasso and a false positive worse than lasso. Hence, simple weighting schemes can adversely affect the performance of adaptive lasso. In some cases, AL(Z) has better prediction error than bridge and logarithm penalty, but it comes at the expense of an unnecessarily complex model. In the situations we considered, MSAL and scad were competitive with log, bridge, and AL(Z) but not the best performers. Finally, with optimal weights, we see that AL(Opt) has better prediction error and $\ell_2$ error than the oracle and the FP and FN error rates are of the same order as the oracle. Presumably, the prediction error of AL(Opt) is slightly better than the oracle because it retains one of two noise variables for improved finite sample performance and the slightly higher FP error rate is due to the perturbation.

5.2. Marked Outcomes

The second set of simulations is specific to the marked variable application. Here, we simulated data according the joint model

$$
\begin{pmatrix}
Y_i \\
T_i
\end{pmatrix} = z_i^T \begin{pmatrix}
\beta \\
\vartheta
\end{pmatrix} + \begin{pmatrix}
\varepsilon_{Y,i} \\
\varepsilon_{T,i}
\end{pmatrix},
$$

where the errors $(\varepsilon_{Y,i}, \varepsilon_{T,i})$ are bivariate normal with mean zero and covariance $\Sigma$,

$$
\Sigma = \Gamma \Gamma^T = \begin{pmatrix}
\sigma_{\varepsilon_Y} & 0 \\
0 & \sigma_{\varepsilon_T}
\end{pmatrix} \begin{pmatrix}
1 & 0.25 \\
0.25 & 1
\end{pmatrix} \begin{pmatrix}
\sigma_{\varepsilon_Y} & 0 \\
0 & \sigma_{\varepsilon_T}
\end{pmatrix}.
$$

The design matrix and coefficients on the mark-scale are exactly as in Section 5.1 except with a sample size of $n = 150$. The regression coefficients on the time-scale are $\vartheta_j = 0.5$ for $j = 1, \ldots, d$. In this setting, the censoring random variable was independently simulated, $C_i \sim \text{Un}(0, 6)$. Then, the observed data is defined as $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, and $\tilde{Y}_i = \delta_i Y_i + (1 - \delta_i)(-100)$. The four simulations represent weak ($\rho = 0.25$) and strong ($\rho = 0.75$) correlation between adjacent predictors and low and high noise error variances, $\sigma^2 = 2$ and 9,
respectively, with $\sigma = \sigma_{\epsilon Y} = \sigma_{\epsilon T}$. The regularization parameters were 19 values in $[1/n^2, 2/\sqrt{n}]$ and identical for scad, hard, MSAL, log, bridge, and adaptive lasso. The simulation results from 100 Monte Carlo studies are displayed in Table 3.

In Table 3, we find some similar patterns as in Table 2 but identifying the best estimator is less clear. Hard thresholding had the smallest false positive rate under weak correlation and was among the top two methods under stronger correlation with low noise. Bridge and MSAL regression consistently performed moderately better than logarithmic penalty. In fact, MSAL had the smallest FP error rates under strong correlation. As in the uncensored data simulations, AL(H) continued to have much higher FP error rates compared other procedures with the same asymptotic properties. While we cannot declare log-penalized estimation a clear winner in the medical cost simulation, it does not perform significantly worse than its competitors either.

6. High-dimensional Considerations

This paper is concerned with data sets where $n > d$ and a natural question is whether log penalized least squares regression can apply to high-dimensional data also. To illustrate fitting high-dimensional data via Algorithm 1, we simulated one data set with $n = 40$ and $d = 60$. Six coefficients were non-zero and the design matrix was standard normal with identity covariance matrix. We plot the coefficient paths for four estimators — lasso, logarithmic penalty, bridge ($\ell_{0.5}$), and optimal adaptive lasso — all using coordinate-wise optimization. The optimal adaptive lasso uses weights $w_j = 1/(|\beta_0| + \epsilon)$. The coefficient paths are displayed in Figure 3. We note that lasso and optimal adaptive lasso (AL(Opt)) have smoother coefficient paths than either bridge and logarithmic penalty because they are exact solutions to convex optimization problems. However, as far as we can tell, the smoothness of the coefficient path has modest value in practical work.

There is no conceptual difficulty in extending Algorithm 1 to computing coefficient paths on the mark-scale for censored medical cost data after an initial estimate $\varphi^*$ on the time-scale is computed. But choosing the best coefficient estimate on the time-scale that leads to optimal model performance on the mark-scale is challenging. A second challenge is to conceive of a new strategy for parameter tuning that does not rely on an estimate of the coefficients from the full model. We have no direct evidence to suggest a concrete solution to either of these challenges. Johnson, Long and Chung (2011) considered a related problem in unbiased transformations for censored outcomes in high-dimensional data where they first screen the data, estimate coefficients on the smaller subset, then impute the censored outcomes. A similar heuristic could be applied to our problem here on the time-scale, then use these estimates to calibrate the coefficient estimates on the mark-scale. As we have no data to motivate these investigations, we leave these problems for future work.
Fig 3. Regularized coefficient paths for data set with $n < d$. 
7. Discussion

In this paper, we reviewed, adopted, and extended iteratively reweighted $\ell_1$ minimization to achieve simultaneous coefficient estimation and variable selection in regression models. The basic asymptotic theory behind log-penalized least squares estimators was presented elsewhere Zou and Li (2008); Johnson, Lin and Zeng (2008) but inferior computational algorithms were presented in those papers. We present a better numerical algorithm using ideas from the signal processing literature Candes, Wakin and Boyd (2008). In our limited simulation studies for uncensored data, the log-penalized least squares estimator had somewhat better operating characteristics than adaptive lasso and scad. While one numerical study is not enough evidence to draw any final conclusions, our findings generally agree with those reported elsewhere (Gasso, Rakotomamonjy and Canu, 2009) albeit in a different setup.

We presented a new family of $\ell_1$-regularized variable selection procedures for censored medical cost through an extension of Huang’s (2002) calibration estimator. The problem is more challenging than variable selection in survival data because of the induced dependent censorship and joint modeling. By carefully choosing weight functions, we can write the solution to the penalized estimating equations into the minimizer of a penalized least squares objective function, with a surrogate least squares loss function. This simplification leads to a substantial improvement in the constrained, convex optimization and allows us to perform simultaneous coefficient estimation and variable selection through optimization transfer. We propose information-based model selection criteria for our problem through a new dispersion statistic. The asymptotic theory says that the log-penalized calibration estimator for lifetime medical cost possesses an oracle property. Simulation studies confirms that our log-penalized estimator performs well in practice.

Appendix A: Large Sample Theory

In this section we outline the oracle property for the proposed log-penalized estimator of lifetime medical cost. This theory makes use of existing asymptotic theories for Huang’s (2002) calibration estimator and a theory of penalized estimating functions by Johnson, Lin and Zeng (2008). To begin, we review the asymptotic properties of Huang’s (2002) estimator.

Define $\theta = (\beta^T, \vartheta^T)^T$ and the “stacked” estimating function, $S_\varepsilon(\theta) = \{S_{\varepsilon Y}(\theta), S_{\varepsilon T}(\vartheta)\}^T$. Define $\hat{\theta}_{\text{Full}}$ as the zero-crossing of the estimating equations, $0 = S_\varepsilon(\theta)$. Under Conditions A-E in Huang (2002, Appendix), $\hat{\theta}_{\text{Full}}$ is strongly consistent for the true coefficient vector $\theta_0$ and $n^{1/2}(\hat{\theta}_{\text{Full}} - \theta_0)$ converges in distribution to a mean-zero normal random vector with covariance $V = \Lambda^{-1}\Omega\Lambda^{-1}$, where $\Lambda \equiv \Lambda(\theta_0)$ is the asymptotic slope matrix of $S_\varepsilon(\theta)$ evaluated at $\theta = \theta_0$ (see Huang, 2002, expression (A.2) in the Appendix) and $\Omega \equiv \Omega(\theta_0)$ is the asymptotic covariance of $n^{-1/2}S_\varepsilon(\theta_0)$. 
To state the oracle property, we need some additional notation. Define the true active set $\mathcal{A} = \{ j | \beta_{0j} \neq 0, j = 1, \ldots, d \}$, its complement set $\mathcal{A}^c = \{ j | \beta_{0j} = 0, j = 1, \ldots, d \}$, as well as the sample quantities $\mathcal{A}_n = \{ j | \hat{\beta}_j \neq 0, j = 1, \ldots, d \}$ and similarly for $\mathcal{A}^c_n$. Define the stacked penalized estimating function,

$$S^P(\theta) = \begin{pmatrix} S_{\epsilon Y}(\theta) \\ S_{\epsilon T}(\theta) \end{pmatrix} - \begin{pmatrix} n\lambda_n s(\beta)|\beta|^{-1} \\ 0 \end{pmatrix},$$

and the partitioned asymptotic covariance $V$:

$$V = \begin{pmatrix} V_{YY} & V_{YT} \\ V_{TY} & V_{TT} \end{pmatrix} = \begin{pmatrix} V_{YAY_A} & V_{YAY_{Ac}} & V_{YAT} \\ V_{YAY_A} & V_{YAY_{Ac}} & V_{YAT} \\ V_{TYA} & V_{TY_{Ac}} & V_{TT} \end{pmatrix}.$$

If the true active set were known a priori, then the asymptotic covariance of the oracle estimator would be

$$V_O = \begin{pmatrix} V_{YAY_A} & V_{YAT} \\ V_{TYA} & V_{TT} \end{pmatrix} = \Lambda_O^{-1}\Omega_O\Lambda_O^{-1},$$

Theorem A.1. Under Conditions A-E in Huang (2002, Appendix), $n^{-1/2}\lambda_n \to 0$ and $\lambda_n \to \infty$, the following results hold:

(i) Sparsity: $\lim_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$;
(ii) Asymptotic normality:

$$n^{1/2}(\Lambda_O + \Gamma) \left\{ \hat{\theta}_{\mathcal{A}} - \theta_{\mathcal{A}} + (\Lambda_O + \Gamma)^{-1}b_n \right\} \to_d N(0, \Omega_O),$$

where $\Gamma = \text{diag}(0_d, \lambda_n s(\beta_{\mathcal{A}})\beta_{\mathcal{A}}^{-2})$ and $b_n = (0_d^T, (\lambda_n s(\beta_{\mathcal{A}})|\beta_{\mathcal{A}}|^{-1})^T)^T$.

The proof of Theorem A.1 follows as a direct consequence of Johnson, Lin and Zeng (2008) and is omitted. For general weight functiones $W_{\epsilon Y}(t, \theta)$ in $S_{\epsilon Y}(\theta)$, Huang’s estimator results in an approximate zero-crossing due to potential multiple zero-crossings of the original unregularized estimating function $S_{\epsilon Y}(\theta)$. However, by restricting the class of weight functions $W_{\epsilon Y}(t, \theta)$ to depend on $\varphi$ only (i.e. not on $\beta$) and choosing the specific marked process with $\psi(t) = t$, the solution to the penalized estimating function $S^P(\theta)$ is exact (Johnson, Lin and Zeng, 2008).

References


Table 1
Coefficient estimates for SWOG 9509 data from the full model, lasso, and logarithmic penalty. Table entries are multiplied by 100.

<table>
<thead>
<tr>
<th>Var.</th>
<th>Full AIC</th>
<th>BIC</th>
<th>Lasso AIC</th>
<th>BIC</th>
<th>Log AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gehan weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. trt</td>
<td>16.74</td>
<td>13.56</td>
<td>9.44</td>
<td>15.37</td>
<td>9.99</td>
<td></td>
</tr>
<tr>
<td>2. sex</td>
<td>15.33</td>
<td>5.49</td>
<td>1.96</td>
<td>8.09</td>
<td>3.62</td>
<td></td>
</tr>
<tr>
<td>3. progstat</td>
<td>35.75</td>
<td>30.55</td>
<td>26.25</td>
<td>32.68</td>
<td>31.35</td>
<td></td>
</tr>
<tr>
<td>4. ps</td>
<td>-10.34</td>
<td>-8.51</td>
<td>-6.48</td>
<td>-8.12</td>
<td>-3.29</td>
<td></td>
</tr>
<tr>
<td>5. stage</td>
<td>8.11</td>
<td>0.56</td>
<td>0</td>
<td>2.07</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6. pleural</td>
<td>3.31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7. weight</td>
<td>5.32</td>
<td>1.70</td>
<td>0</td>
<td>0.73</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8. height</td>
<td>-11.64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9. creat</td>
<td>4.75</td>
<td>1.35</td>
<td>0</td>
<td>0.29</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10. albumin</td>
<td>13.65</td>
<td>8.52</td>
<td>2.79</td>
<td>11.12</td>
<td>3.35</td>
<td></td>
</tr>
<tr>
<td>11. calcium</td>
<td>-10.36</td>
<td>-5.38</td>
<td>0</td>
<td>-7.81</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>12. ldh</td>
<td>-4.21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13. alkptase</td>
<td>-3.74</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>14. bilirubin</td>
<td>8.31</td>
<td>3.41</td>
<td>0</td>
<td>3.72</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15. wbc</td>
<td>-15.24</td>
<td>-11.66</td>
<td>-8.22</td>
<td>-14.17</td>
<td>-10.34</td>
<td></td>
</tr>
<tr>
<td>16. platelet</td>
<td>-2.80</td>
<td>-2.59</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17. hgb</td>
<td>6.70</td>
<td>5.55</td>
<td>4.87</td>
<td>4.94</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>18. age</td>
<td>-1.14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Logrank weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. trt</td>
<td>12.56</td>
<td>7.43</td>
<td>3.65</td>
<td>7.56</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>2. sex</td>
<td>-0.72</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3. progstat</td>
<td>35.00</td>
<td>26.95</td>
<td>22.66</td>
<td>31.57</td>
<td>28.87</td>
<td></td>
</tr>
<tr>
<td>5. stage</td>
<td>-5.60</td>
<td>-4.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6. pleural</td>
<td>12.84</td>
<td>2.03</td>
<td>0</td>
<td>4.63</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7. weight</td>
<td>2.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8. height</td>
<td>-13.29</td>
<td>0</td>
<td>0</td>
<td>-0.35</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9. creat</td>
<td>15.78</td>
<td>1.69</td>
<td>0</td>
<td>4.57</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10. albumin</td>
<td>16.90</td>
<td>9.31</td>
<td>3.70</td>
<td>10.83</td>
<td>3.31</td>
<td></td>
</tr>
<tr>
<td>11. calcium</td>
<td>-16.22</td>
<td>-6.24</td>
<td>0</td>
<td>-8.23</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13. alkptase</td>
<td>-6.86</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>14. bilirubin</td>
<td>12.67</td>
<td>4.38</td>
<td>0</td>
<td>5.06</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15. wbc</td>
<td>-17.83</td>
<td>-12.46</td>
<td>-8.74</td>
<td>-12.69</td>
<td>-8.10</td>
<td></td>
</tr>
<tr>
<td>16. platelet</td>
<td>0.38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17. hgb</td>
<td>15.05</td>
<td>11.86</td>
<td>9.80</td>
<td>11.64</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>18. age</td>
<td>-5.44</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Table 2
Simulation results comparing regularizations for uncensored data in a least squares framework. Table entries include median $\ell_2$ error, median model error (ME), average number of false positives (FP) and average number of false negatives (FN).

<table>
<thead>
<tr>
<th>Method</th>
<th>$|\hat{\beta} - \beta_0|_2$</th>
<th>ME</th>
<th>FP</th>
<th>FN</th>
<th>$|\hat{\beta} - \beta_0|_2$</th>
<th>ME</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2 = 2$</td>
<td></td>
<td></td>
<td></td>
<td>$\sigma^2 = 9$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>2.02</td>
<td>1.80</td>
<td>22.00</td>
<td>0</td>
<td>8.76</td>
<td>7.84</td>
<td>22.00</td>
<td>0</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.32</td>
<td>1.25</td>
<td>9.33</td>
<td>0</td>
<td>5.06</td>
<td>5.13</td>
<td>8.39</td>
<td>0.43</td>
</tr>
<tr>
<td>Scad</td>
<td>0.81</td>
<td>0.72</td>
<td>3.33</td>
<td>0</td>
<td>6.39</td>
<td>5.74</td>
<td>6.30</td>
<td>0.94</td>
</tr>
<tr>
<td>Hard</td>
<td>1.63</td>
<td>1.40</td>
<td>8.03</td>
<td>0</td>
<td>7.27</td>
<td>6.57</td>
<td>4.78</td>
<td>1.29</td>
</tr>
<tr>
<td>MSAL</td>
<td>1.32</td>
<td>1.20</td>
<td>2.48</td>
<td>0</td>
<td>6.68</td>
<td>5.93</td>
<td>3.52</td>
<td>1.40</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.88</td>
<td>0.79</td>
<td>2.10</td>
<td>0</td>
<td>5.45</td>
<td>4.99</td>
<td>2.44</td>
<td>1.70</td>
</tr>
<tr>
<td>Log</td>
<td>0.74</td>
<td>0.70</td>
<td>1.35</td>
<td>0.01</td>
<td>6.25</td>
<td>5.67</td>
<td>2.14</td>
<td>2.15</td>
</tr>
<tr>
<td>AL(Z)</td>
<td>0.88</td>
<td>0.82</td>
<td>2.88</td>
<td>0</td>
<td>5.64</td>
<td>5.45</td>
<td>4.42</td>
<td>1.25</td>
</tr>
<tr>
<td>AL(H)</td>
<td>2.08</td>
<td>1.88</td>
<td>11.54</td>
<td>0.20</td>
<td>7.81</td>
<td>7.37</td>
<td>6.61</td>
<td>1.51</td>
</tr>
<tr>
<td>AL(Opt)</td>
<td>0.54</td>
<td>0.51</td>
<td>0.11</td>
<td>0</td>
<td>2.38</td>
<td>2.23</td>
<td>0.08</td>
<td>0</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.85</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>2.65</td>
<td>2.53</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2 = 9$</th>
<th></th>
<th></th>
<th></th>
<th>$\sigma^2 = 2$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>5.97</td>
<td>1.70</td>
<td>22.00</td>
<td>0</td>
<td>28.44</td>
<td>8.14</td>
<td>22.00</td>
<td>0</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.58</td>
<td>0.97</td>
<td>9.25</td>
<td>0.03</td>
<td>10.89</td>
<td>4.61</td>
<td>8.69</td>
<td>1.21</td>
</tr>
<tr>
<td>Scad</td>
<td>3.60</td>
<td>1.13</td>
<td>3.15</td>
<td>0.87</td>
<td>25.76</td>
<td>7.52</td>
<td>4.55</td>
<td>4.96</td>
</tr>
<tr>
<td>Hard</td>
<td>5.27</td>
<td>1.55</td>
<td>10.25</td>
<td>0.23</td>
<td>28.39</td>
<td>8.07</td>
<td>10.51</td>
<td>2.86</td>
</tr>
<tr>
<td>MSAL</td>
<td>3.61</td>
<td>1.82</td>
<td>2.62</td>
<td>0.60</td>
<td>20.92</td>
<td>7.01</td>
<td>3.42</td>
<td>4.61</td>
</tr>
<tr>
<td>Bridge</td>
<td>2.64</td>
<td>0.95</td>
<td>1.82</td>
<td>0.53</td>
<td>18.53</td>
<td>6.13</td>
<td>2.19</td>
<td>5.12</td>
</tr>
<tr>
<td>Log</td>
<td>3.04</td>
<td>0.97</td>
<td>1.48</td>
<td>0.78</td>
<td>19.18</td>
<td>6.22</td>
<td>2.26</td>
<td>4.99</td>
</tr>
<tr>
<td>AL(Z)</td>
<td>3.17</td>
<td>1.11</td>
<td>3.15</td>
<td>0.55</td>
<td>18.17</td>
<td>6.30</td>
<td>5.14</td>
<td>3.83</td>
</tr>
<tr>
<td>AL(H)</td>
<td>3.28</td>
<td>1.13</td>
<td>11.53</td>
<td>0.06</td>
<td>13.10</td>
<td>4.76</td>
<td>11.23</td>
<td>1.42</td>
</tr>
<tr>
<td>AL(Opt)</td>
<td>1.23</td>
<td>0.50</td>
<td>0.04</td>
<td>0</td>
<td>6.48</td>
<td>2.52</td>
<td>0.09</td>
<td>0</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.51</td>
<td>1.06</td>
<td>0</td>
<td>0</td>
<td>6.60</td>
<td>2.97</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## Table 3

Simulation results comparing regularizations for marked outcomes. Table entries include median $\ell_2$ error, median model error (ME), average number of false positives (FP) and average number of false negatives (FN).

<table>
<thead>
<tr>
<th>Method</th>
<th>$|\hat{\beta} - \beta_0|_2$</th>
<th>ME</th>
<th>FP</th>
<th>FN</th>
<th>$|\hat{\beta} - \beta_0|_2$</th>
<th>ME</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2 = 2$</td>
<td></td>
<td></td>
<td></td>
<td>$\sigma^2 = 9$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.25$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>1.25</td>
<td>1.12</td>
<td>22.00</td>
<td>0</td>
<td>6.32</td>
<td>5.68</td>
<td>22.00</td>
<td>0</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.01</td>
<td>1.03</td>
<td>10.43</td>
<td>0</td>
<td>4.53</td>
<td>4.5</td>
<td>11.65</td>
<td>0.14</td>
</tr>
<tr>
<td>Scad</td>
<td>0.59</td>
<td>0.53</td>
<td>2.16</td>
<td>0</td>
<td>4.99</td>
<td>4.45</td>
<td>5.93</td>
<td>0.75</td>
</tr>
<tr>
<td>Hard</td>
<td>0.65</td>
<td>0.59</td>
<td>1.14</td>
<td>0</td>
<td>5.03</td>
<td>4.37</td>
<td>3.14</td>
<td>1.24</td>
</tr>
<tr>
<td>MSAL</td>
<td>0.85</td>
<td>0.84</td>
<td>1.72</td>
<td>0</td>
<td>4.78</td>
<td>4.43</td>
<td>4.13</td>
<td>0.88</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.67</td>
<td>0.63</td>
<td>2.38</td>
<td>0</td>
<td>4.14</td>
<td>3.84</td>
<td>3.89</td>
<td>0.79</td>
</tr>
<tr>
<td>Log</td>
<td>0.81</td>
<td>0.74</td>
<td>2.17</td>
<td>0</td>
<td>5.45</td>
<td>5.36</td>
<td>5.15</td>
<td>1.57</td>
</tr>
<tr>
<td>AL(Z)</td>
<td>0.68</td>
<td>0.64</td>
<td>2.18</td>
<td>0</td>
<td>4.33</td>
<td>3.94</td>
<td>5.65</td>
<td>0.53</td>
</tr>
<tr>
<td>AL(H)</td>
<td>1.10</td>
<td>1.02</td>
<td>11.86</td>
<td>0.07</td>
<td>5.28</td>
<td>4.84</td>
<td>9.25</td>
<td>0.94</td>
</tr>
<tr>
<td>AL(Opt)</td>
<td>0.48</td>
<td>0.48</td>
<td>0.14</td>
<td>0</td>
<td>2.01</td>
<td>1.94</td>
<td>0.48</td>
<td>0</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.43</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
<td>1.94</td>
<td>1.86</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\rho = 0.75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>5.51</td>
<td>1.59</td>
<td>22.00</td>
<td>0</td>
<td>22.45</td>
<td>6.59</td>
<td>22.00</td>
<td>0</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.86</td>
<td>1.25</td>
<td>9.56</td>
<td>0.08</td>
<td>9.49</td>
<td>4.10</td>
<td>9.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Scad</td>
<td>4.16</td>
<td>1.47</td>
<td>3.69</td>
<td>0.73</td>
<td>19.83</td>
<td>6.37</td>
<td>5.92</td>
<td>4.06</td>
</tr>
<tr>
<td>Hard</td>
<td>3.51</td>
<td>1.31</td>
<td>3.12</td>
<td>0.71</td>
<td>20.41</td>
<td>6.37</td>
<td>4.82</td>
<td>4.00</td>
</tr>
<tr>
<td>MSAL</td>
<td>3.63</td>
<td>1.64</td>
<td>2.89</td>
<td>0.55</td>
<td>16.34</td>
<td>5.71</td>
<td>3.64</td>
<td>3.74</td>
</tr>
<tr>
<td>Bridge</td>
<td>2.90</td>
<td>1.23</td>
<td>3.18</td>
<td>0.57</td>
<td>13.76</td>
<td>4.76</td>
<td>3.74</td>
<td>3.37</td>
</tr>
<tr>
<td>Log</td>
<td>4.16</td>
<td>1.56</td>
<td>4.03</td>
<td>1.03</td>
<td>18.38</td>
<td>5.95</td>
<td>4.91</td>
<td>4.23</td>
</tr>
<tr>
<td>AL(Z)</td>
<td>3.13</td>
<td>1.19</td>
<td>4.68</td>
<td>0.34</td>
<td>14.76</td>
<td>5.08</td>
<td>6.07</td>
<td>2.80</td>
</tr>
<tr>
<td>AL(H)</td>
<td>4.63</td>
<td>1.68</td>
<td>13.6</td>
<td>0.28</td>
<td>12.73</td>
<td>5.03</td>
<td>12.11</td>
<td>1.41</td>
</tr>
<tr>
<td>AL(Opt)</td>
<td>1.60</td>
<td>0.71</td>
<td>1.15</td>
<td>0</td>
<td>5.94</td>
<td>2.39</td>
<td>0.78</td>
<td>0.13</td>
</tr>
<tr>
<td>Oracle</td>
<td>1.16</td>
<td>0.51</td>
<td>0</td>
<td>0</td>
<td>5.43</td>
<td>2.28</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>