Published in final edited form as:
Stat Sin. 2012 ; 22: 27–294. doi:10.5705/ss.2010.107.

ELASTIC NET FOR COX'S PROPORTIONAL HAZARDS MODEL WITH A SOLUTION PATH ALGORITHM

Yichao Wu
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
wu@stat.ncsu.edu

Abstract

For least squares regression, Efron et al. (2004) proposed an efficient solution path algorithm, the least angle regression (LAR). They showed that a slight modification of the LAR leads to the whole LASSO solution path. Both the LAR and LASSO solution paths are piecewise linear. Recently Wu (2011) extended the LAR to generalized linear models and the quasi-likelihood method. In this work we extend the LAR further to handle Cox's proportional hazards model. The goal is to develop a solution path algorithm for the elastic net penalty (Zou and Hastie (2005)) in Cox's proportional hazards model. This goal is achieved in two steps. First we extend the LAR to optimizing the log partial likelihood plus a fixed small ridge term. Then we define a path modification, which leads to the solution path of the elastic net regularized log partial likelihood. Our solution path is *exact* and piecewise determined by ordinary differential equation systems.

Key words and phrases

Cox's proportional hazards model; elastic net; LARS; LASSO; ordinary differential equation; solution path algorithm

1. Introduction

The main goal of survival analysis is to characterize the dependence of the survival time Y on a covariate vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. Cox's proportional hazards model (Cox (1972)) assumes that the hazard function $h(y|\boldsymbol{x})$ of a subject with covariate vector $\boldsymbol{x}$ takes the form

$$h(y|\boldsymbol{x}) = h_0(y)\exp(\boldsymbol{x}^T\beta), \quad (1.1)$$

where $h_0(y)$ is a completely unspecified baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. In practice, it is not necessary that all covariates contribute to predicting survival outcomes. Thus, another goal of survival analysis is to identify important risk factors and quantify their risk contributions. As survival data with many predictors prevail in clinical trial studies, risk factor identification becomes more important than ever for analyzing high-dimensional survival data. The problem is to select a submodel of (1.1) by providing a sparse estimate of $\boldsymbol{\beta}$.

There are many model selection techniques in the literature and most of them have been successfully extended to survival analysis. They include such classical methods as the best-subset selection and stepwise selection. More recently, Tibshirani (1996) proposed to use the L_1 penalty to regularize least squares regression; sparse estimate of the regression parameter is made possible due to the L_1 penalty's singularity at the origin. This technique was named the least absolute shrinkage and selection operator (LASSO), and later extended to the Cox proportional hazards model in Tibshirani (1997). However the LASSO penalty leads to

biased estimates for true non-zero coefficients. To alleviate this bias issue, Fan and Li (2001) proposed the SCAD penalty, which is symmetric and piecewise quadratic. It is linear around the origin and flattens out near the two ends; in between, it is smoothly connected by two quadratic pieces. They showed that asymptotically the SCAD penalized estimate behaves like the oracle estimate were the true sparsity pattern known *a priori*. The oracle property of the SCAD was later extended to survival models in Fan and Li (2002). The adaptive-LASSO was proposed for least squares regression by Zou (2006), and for Cox's proportional hazards model by Zhang and Lu (2007), and its oracle properties were established as well. There are many other techniques available for variable selection, including the elastic net (Zou and Hastie (2005)). See Fan and Lv (2010) and references therein for an overview of variable selection methods.

A novel least angle regression (LAR) solution path algorithm was proposed in Efron et al. (2004). The LAR produces a piecewise linear solution path for the least squares regression. They showed that slight modifications of the LAR lead to the LASSO and Forward Stagewise linear regression solution paths. Together they are called LARS. For data $\{(y_i, \mathbf{z}_i), i = 1, \ldots, n\}$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^T \in \mathbb{R}^p$, ordinary least squares (OLS) regression solves

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^T \boldsymbol{w})^2 \quad (1.2)$$

to estimate $\boldsymbol{w} = (w_1, \ldots, w_p)^T$. Applying location and scale transformations if necessary, we assume without loss of generality that $\sum_{i=1}^{n} z_{ij} = 0$, $\sum_{i=1}^{n} z_{ij}^2 = 1$ for $j = 1, \ldots, p$, and $\sum_{i=1}^{n} y_i = 0$.

For OLS, the LAR provides a solution path $\boldsymbol{w}(t)$ indexed by $t \in [0, \infty)$. It starts at the origin with $\boldsymbol{w}(0) = \mathbf{0}$; for large enough t, $\boldsymbol{w}(t)$ is the same as the full solution to (1.2). The intermediate solution path is piecewise linear; over each piece, it moves along the direction that keeps the correlation between the current residuals and each active predictor equal in absolute value. Denote the jth predictor vector by $\mathbf{z}^{(j)} = (z_{1j}, \ldots, z_{nj})^T$, and define the residual vector at t by $\boldsymbol{e}(\boldsymbol{w}(t)) = (e_1(\boldsymbol{w}(t)), \ldots, e_n(\boldsymbol{w}(t)))^T$ with $e_i(\boldsymbol{w}(t)) = y_i - \boldsymbol{z}_i^T \boldsymbol{w}(t)$ for $i = 1, \ldots, n$. Then along the LAR solution path $\boldsymbol{w}(t)$, the current correlation $\boldsymbol{e}(\boldsymbol{w}(t))^T \mathbf{z}^{(j)}$ has the same absolute value for each active predictor j. Note that

$$\boldsymbol{e}(\boldsymbol{w}(t))^T \boldsymbol{z}^{(j)} = -\frac{1}{2} \left(\frac{\partial}{\partial w_j} \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^T \boldsymbol{w})^2 \right) \Big|_{\boldsymbol{w}(t)}.$$

This implies that the objective function has the same absolute value of the first-order partial derivatives for each active predictor along the LAR solution path. Mathematically,

$$\left| \left(\frac{\partial}{\partial w_j} \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^T \boldsymbol{w})^2 \right) \Big|_{\boldsymbol{w}(t)} \right| = \left| \left(\frac{\partial}{\partial w_{j'}} \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^T \boldsymbol{w})^2 \right) \Big|_{\boldsymbol{w}(t)} \right| \quad (1.3)$$

for any j and j' among the active set at t. For the diabetes data in the R package LARS, we plot the LAR solution path in the top left panel of Figure 1. The first-order partial derivatives along the LAR solution path are shown in the top right panel of Figure 1. The derivatives in absolute value, namely $\left| \left(\frac{\partial}{\partial w_j} \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^T \boldsymbol{w})^2 \right) \Big|_{\boldsymbol{w}(t)} \right|$, are given in the bottom panel of Figure 1. One sees that, at the end of each LAR step, a new predictor joins the

group of active predictors, sharing the honor of having the same largest absolute value of the first-order partial derivatives. The LAR algorithm terminates at the full OLS estimate of (1.2) when all the first-order partial derivatives are exactly zero. Based on this observation, Wu (2011) proposed an extension to handle generalized linear models and more generally the quasi-likelihood method.

In this work, we extend the LAR to the Cox's proportional hazards model. With the elastic net penalty in mind, we add a fixed small ridge term to the log partial likelihood function and call this extension CoxLAR-ridge. When the ridge term is exactly zero, we have the original log partial likelihood and call the corresponding algorithm CoxLAR. As in Efron et al. (2004) and Wu (2011), we show that the CoxLAR-ridge can be slightly modified to get the corresponding whole solution path for the LASSO regularized counterpart; it is called CoxEN as the LASSO penalty with a small ridge term leads to the elastic net penalty. By setting the ridge term to be zero the CoxEN includes the CoxLASSO, the LASSO regularized log partial likelihood, as a special case. Together, we use CoxLARS in the same spirit as LARS in Efron et al. (2004). In addition to considering different models, another difference from Wu (2011) is that we include a ridge penalty term to consider the more general elastic net penalty. The elastic net penalty is highly desirable in that it is capable of selecting more predictors than the sample size, while it is known that the number of predictors selected by the LASSO can be at most equal to the sample size. See more discussion on this issue in Zou and Hastie (2005).

Previously Park and Hastie (2007) provided a solution path algorithm for L_1-regularized generalized linear models and Cox's proportional hazards model. Their algorithm is based on the predictor-corrector method of convex optimization. In their R package "glmpath", one may choose an extreme small bound for arc length (L_1 norm) of each step to obtain an exact solution path. In this case, it essentially uses a warm start each time to compute the exact solution at a fine grid of the tuning parameter and connects these exact solutions by straight lines. They still need to solve many optimization problems, one at each tuning parameter point. They did not address how the solution changes when the tuning parameter changes. Our new algorithms CoxLAR and CoxLASSO answer this question, the solution path propagates according to ordinary differential equation (ODE) systems. Thus the commonly used fourth-order Runge-Kutta method can be used to solve these ODE systems to obtain the whole CoxLARS solution paths. Other papers on solution path algorithms include Hastie et al. (2004), Rosset and Zhu (2007), Zou (2008), Friedman, Hastie, and Tibshirani (2008), Yuan and Zou (2009), and references therein. In particular, Zou (2008) proposed an efficient adaptive shrinkage method for the Cox's proportional hazards model and adapted the LARS to provide a piecewise linear solution path.

The rest of the article is organized as follows. Section 2 presents our new algorithm CoxLARS. Properties of the CoxLARS are given in Section 3. Numerical examples in Section 4 illustrate how our new algorithm works with data sets. A summary is given in Section 5. The appendix gives all technical proofs.

2. Extension of LARS: CoxLARS

Consider a sample of n subjects. Let T_i and C_i be the failure time and censoring time, respectively, for subject $i = 1, \ldots, n$. Write $Y_i = \min(T_i, C_i)$ and let the censoring indicator be $\delta_i = I(T_i \leq C_i)$. Denote the covariate vector of the ith subject by $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Assume that T_i and C_i are conditionally independent given covariate vector $\mathbf{x}_i$, and that the censoring mechanism is noninformative. Our data set is $\{(\mathbf{x}_i, y_i, \delta_i), i = 1, \ldots, n\}$.

Assume the data come from model (1.1). For simplicity, we suppose there are no ties in the observed failure times, otherwise techniques in Breslow (1974) may be used. The log partial likelihood is given by

$$L_{pl}(\beta)=\sum_{i=1}^{n}\delta_i \boldsymbol{x}_i^T\beta-\sum_{i=1}^{n}\delta_i\log\left(\sum_{j\in R_i}\exp(\boldsymbol{x}_j^T\beta)\right), \quad (2.1)$$

where $R_i = \{j = 1, \ldots, n : y_j \geq y_i\}$ denotes the risk set just before the time y_i.

Note that when the elastic net penalty (Zou and Hastie, 2005) is considered, we are solving

$$\min_{\beta}\sum_{i=1}^{n}\delta_i \boldsymbol{x}_i^T\beta-\sum_{i=1}^{n}\delta_i\log\left(\sum_{j\in R_i}\exp(\boldsymbol{x}_j^T\beta)\right)+\frac{\gamma}{2}\sum_{j=1}^{p}\beta_j^2+\lambda\sum_{j=1}^{p}|\beta_j|, \quad (2.2)$$

where $\gamma \geq 0$ and $\lambda \geq 0$ are two regularization parameters. In order to incorporate the elastic net into our consideration, we include a small ridge penalty term and set

$$L(\beta)\equiv L_\gamma(\beta)\triangleq\sum_{i=1}^{n}\delta_i \boldsymbol{x}_i^T\beta-\sum_{i=1}^{n}\delta_i\log\left(\sum_{j\in R_i}\exp(\boldsymbol{x}_j^T\beta)\right)+\frac{\gamma}{2}\sum_{j=1}^{p}\beta_j^2 \quad (2.3)$$

for some fixed small $\gamma \geq 0$. Note that this reduces to the LASSO penalized counterpart when $\gamma = 0$. It is known that the LASSO penalty can select at most n predictors for the $p > n$ case. However as long as $\gamma > 0$, we can select more than n predictors by solving (2.2) when $p > n$. Our consideration is similar to the LARS-EN algorithm proposed in Zou and Hastie (2005) in that the LARS-EN adapted the LARS algorithm to obtain elastic net solution path for each fixed ridge term.

We use t to index our solution path. As motivated by (1.3), our extension CoxLAR-ridge seeks a solution path $\boldsymbol{\beta}(t)$ of (2.3) that satisfies

$$\left|\frac{\partial}{\partial\beta_j}L(\beta)\Big|_{\beta(t)}\right|=\left|\frac{\partial}{\partial\beta_{j'}}L(\beta)\Big|_{\beta(t)}\right| \quad (2.4)$$

for any two predictors j and j' that are active at t.

For $L(\boldsymbol{\beta})$, denote its vector of first-order partial derivatives by $\boldsymbol{b}(\boldsymbol{\beta}) = (b_1(\boldsymbol{\beta}), \ldots, b_p(\boldsymbol{\beta}))^T$ and its matrix of second-order partial derivatives by $\boldsymbol{M}(\boldsymbol{\beta}) = (m_{jk}(\boldsymbol{\beta}))_{1\leq j,k\leq p}$, where

$$b_j(\beta)=\frac{\partial}{\partial\beta_j}L(\beta)=\sum_{i=1}^{n}\delta_i x_{ij}-\sum_{i=1}^{n}\delta_i\frac{\sum_{l\in R_i}\exp(\boldsymbol{x}_l^T\beta)\boldsymbol{x}_{lj}}{\sum_{l\in R_i}\exp(\boldsymbol{x}_l^T\beta)}+\gamma\beta_j$$

and $m_{jk}(\beta)=\dfrac{\partial^2}{\partial\beta_j\partial\beta_k}L(\beta)$ is given by

$$-\sum_{i=1}^{n}\delta_i\left[\frac{\sum_{l\in R_i}\exp(x_l^T\beta)\boldsymbol{x}_{lj}x_{lk}}{\sum_{l\in R_i}\exp(x_l^T\beta)}-\frac{(\sum_{l\in R_i}\exp(\boldsymbol{x}_l^T\beta)\boldsymbol{x}_{lj})(\sum_{l\in R_i}\exp(\boldsymbol{x}_l^T\beta)\boldsymbol{x}_{lk})}{(\sum_{l\in R_i}\exp(\boldsymbol{x}_l^T\beta))^2}\right]+\gamma I_{\{j=k\}}$$

for $1 \le j, k \le p$, where $I_{\{j=k\}} = 1$ if $j = k$ and 0 otherwise.

At t with solution $\boldsymbol{\beta}(t)$, denote the corresponding active index set by $\mathcal{A}\boldsymbol{\beta}(t))$ and, interchangeably, by $\mathcal{A}_t$. For any two index sets $\mathcal{A}$ and $\mathcal{B}$, vector $\boldsymbol{b}$, and matrix $\boldsymbol{M}$, let $\boldsymbol{b}_{\mathcal{A}}$ be the sub-vector of $\boldsymbol{b}$ consisting of those elements with index in $\mathcal{A}$ and $\boldsymbol{M}_{\mathcal{A},\mathcal{B}}$ be the sub-matrix of $\boldsymbol{M}$ consisting of those elements with row index in $\mathcal{A}$ and column index in $\mathcal{B}$. When $\mathcal{A} = \{j\}$ is a singleton, we write $\boldsymbol{M}_{j,\mathcal{B}}$ similarly $\boldsymbol{M}_{\mathcal{A},k}$ when $\mathcal{B} = \{k\}$. Denote the complement of $\mathcal{A}$ by $\mathcal{A}^c = \{1, \ldots, p\} \setminus \mathcal{A}$

Note that, at any t with active predictor set $\mathcal{A}$ the corresponding solution component is set to zero for any inactive predictor, namely $\beta_j(t) = 0$ for any $j \notin \mathcal{A}$ Thus it is enough to find how the solution coefficient components, corresponding to active predictors $\beta_j(t)$ with $j \in \mathcal{A}$ are updated. Recall that our desired solution path should be such that the active predictors have the same absolute value of the first-order partial derivatives as at (2.4). Thus, as t grows, $|b_j(\boldsymbol{\beta}(t))|$ decreases at the same speed for $j \in \mathcal{A}$ Assume that in a small neighborhood of t, the active set $\mathcal{A}_t$ remains the same as $\mathcal{A}$ say. Note that

$$\frac{d}{dt}\boldsymbol{b}_{\mathcal{A}}(\beta(t)) = \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\beta(t))\frac{d}{dt}\beta_{\mathcal{A}}(t), \quad (2.5)$$

since the active set $\mathcal{A}_t$ remains the same as $\mathcal{A}$ and thus $\boldsymbol{\beta}_{\mathcal{A}^c}(t) = \mathbf{0}$ in a small neighborhood of t.

According to (2.4),

$$\frac{d}{dt}\boldsymbol{b}_{\mathcal{A}}(\beta(t)) = c(t)\ \{-\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t)))\} \quad (2.6)$$

for some $c(t) > 0$. Here the negative sign on the right hand side ensures that $|b_j(\boldsymbol{\beta}(t))|$ is decreasing in t for each $j \in \mathcal{A}$ Furthermore (2.6) guarantees that the $|b_j(\boldsymbol{\beta}(t))|, j \in \mathcal{A}$ decrease at the same speed. Note that we can think of t as a function of τ with the solution path indexed by τ. With an appropriate choice of $t(\tau)$, (2.6) holds with t replaced by τ and the $c(t)$ term replaced a constant. Different $c(t)$ lead to different indexing systems of the solution path. Thus, without loss of generality, we set $c(t) \equiv 1$ in (2.6) and write

$$\frac{d}{dt}\boldsymbol{b}_{\mathcal{A}}(\beta(t)) = -\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t))). \quad (2.7)$$

In fact this turns out to be a good choice in that t here is simply related to the maximum absolute value of the first-order derivatives at t, as we shall see later.

Based on (2.5) and (2.7), the solution path should satisfy

$$\frac{d}{dt}\beta_{\mathcal{A}}(t) = -(\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\beta(t)))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t))). \quad (2.8)$$

Recall that $\boldsymbol{\beta}_{\mathcal{A}^c}(t) = \mathbf{0}$. These completely define the path updating direction $\frac{d}{dt}\beta(t)$. Thus for any $t^* > t$, we may take a tentative solution path piece

$$\tilde{\beta}_{\mathcal{A}}(t^*) = \beta_{\mathcal{A}}(t) - \int_t^{t^*}(\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\tilde{\beta}(\tau)))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\tilde{\beta}(\tau)))d\tau \text{ and } \tilde{\beta}_{\mathcal{A}^c}(t^*) = \mathbf{0}. \quad (2.9)$$

With this tentative solution path piece, we implicitly assume that the active set remains the same between t and t^*. Assume that, at the start with t, $|b_j(\boldsymbol{\beta}(t))| = |b_{j'}(\boldsymbol{\beta}(t))|$ for any $j, j' \in$

$\mathscr{A}$ Then (2.9) guarantees that $|b_j(\tilde{\boldsymbol{\beta}}(t^*))| = |b_{j'}(\tilde{\boldsymbol{\beta}}(t^*))|$ for any $j, j' \in \mathscr{A}$ along the tentative solution path piece $\tilde{\boldsymbol{\beta}}(t^*)$ for $t^* > t$ and, further, that $|b_j(\tilde{\boldsymbol{\beta}}(t^*))|$ is decreasing in t^* for $j \in \mathscr{A}$ Thus, as t^* increases, some inactive predictor $m \notin \mathscr{A}$ may have $|b_m(\tilde{\boldsymbol{\beta}}(t^*))| \geq |b_j(\tilde{\boldsymbol{\beta}}(t^*))|$ for $j \in \mathscr{A}$ Whenever this happens, the active predictor set has changed and we cannot use (2.9) any more. For any $j \notin \mathscr{A}$ define

$$T_j = \min\left\{t^* > t : \left|b_j(\tilde{\beta}(t^*))\right| \geq \left|b_m(\tilde{\beta}(t^*))\right|\right\}, \quad (2.10)$$

where m is any member of the active predictor set $\mathscr{A}$ Then the active set changes at $T = \min_{j \notin \mathscr{A}} T_j$ from $\mathscr{A}$ to $\mathscr{A} \cup \{j^*\}$, where $j^* = \operatorname{argmin}_{j \notin \mathscr{A}} T_j$.

2.1. Algorithm CoxLAR(-ridge)

The previous discussion leads us to our extension CoxLAR(-ridge) algorithm that is systematically presented next.

We initialize our solution path by identifying the predictor j so that the objective function $L(\boldsymbol{\beta})$ changes fastest with respect to β_j beginning at $\boldsymbol{\beta} = \mathbf{0}$; set

$$t_0 = -\max_{j=1,\ldots,p} |b_j(\mathbf{0})|. \quad (2.11)$$

This specially defined t_0 together with (2.7), leads to $t = -\max_j |b_j(\boldsymbol{\beta}(t))|$ along our solution path. Our solution path begins with $\boldsymbol{\beta}(t_0) = \mathbf{0}$; the corresponding initial active predictor set is $\mathscr{A}_{t_0} = \left\{\operatorname{argmax}_{1 \leq j \leq p} |b_j(\mathbf{0})|\right\}$.

Given t_0, $\boldsymbol{\beta}(t_0)$, and $\mathscr{A}_{t_0}$, we update our solution path using (2.9) until a new variable joins the active set at some $t_1 (> t_0)$ to be determined. We may temporarily update the solution using

$$\tilde{\beta}_{\mathscr{A}_{t_0}}(t) = \beta_{\mathscr{A}_{t_0}}(t_0) - \int_{t_0}^{t} \left(\boldsymbol{M}_{\mathscr{A}_{t_0},\mathscr{A}_{t_0}}(\tilde{\beta}(\tau))\right)^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathscr{A}_{t_0}}(\tilde{\beta}(\tau)))d\tau \text{ and } \tilde{\beta}_{\mathscr{A}_{t_0}^c}(t) = \mathbf{0} \quad (2.12)$$

for $t > t_0$. Here $\tilde{\boldsymbol{\beta}}(t)$ is a temporary solution path defined for any $t > t_0$. For any $j \notin \mathscr{A}_{t_0}$, let

$$T_j = \min\left\{t > t_0 : \left|b_j(\tilde{\beta}(t))\right| \geq \left|b_m(\tilde{\beta}(t))\right|\right\}, \quad (2.13)$$

where $m \in \mathscr{A}_{t_0}$. Then

$$t_1 = \min_{j \notin \mathscr{A}_{t_0}} T_j \quad (2.14)$$

is a *transition point* because the set of active predictor variables changes there.

The CoxLAR-ridge algorithm updates by setting

$$\beta_{\mathscr{A}_{t_0}}(t) = \beta_{\mathscr{A}_{t_0}}(t_0) - \int_{t_0}^{t} \left(\boldsymbol{M}_{\mathscr{A}_{t_0},\mathscr{A}_{t_0}}(\beta(\tau))\right)^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathscr{A}_{t_0}}(\beta(\tau)))d\tau \text{ and } \beta_{\mathscr{A}_{t_0}^c}(t) = \mathbf{0} \quad (2.15)$$

for all $t \in [t_0, t_1]$. The active predictor set stays the same for $t \in [t_0, t_1)$, namely $\mathscr{A}_t = \mathscr{A}_{t_0}$. At t_1, we update the active predictor set by setting $\mathscr{A}_{t_1} = \mathscr{A}_{t_0} \cup \{j \notin \mathscr{A}_{t_0} : T_j = t_1\}$.

At $t = t_1$, the number of active predictors is two. Due to (2.5), (2.8), (2.12), (2.13) and (2.14), solution $\boldsymbol{\beta}(t_1)$ satisfies $|b_j(\boldsymbol{\beta}(t_1))| = |b_{j'}(\boldsymbol{\beta}(t_1))| > |b_k(\boldsymbol{\beta}(t_1))|$ for any $k \notin \mathcal{A}_{t_1}$ and any $j, j' \in \mathcal{A}_{t_1}$.

The CoxLAR-ridge algorithm continues with the updated t_1, $\boldsymbol{\beta}(t_1)$, and $\mathcal{A}_{t_1}$, proceeding according to Algorithm 1. Note that at the end of the mth CoxLARridge step, the transition point t_m, solution $\boldsymbol{\beta}(t_m)$, and active predictor set $\mathcal{A}_{t_m}$ satisfy $t_m = -|b_j(\boldsymbol{\beta}(t_m))|$ for any $j \in \mathcal{A}_{t_m}$, and $|b_j(\boldsymbol{\beta}(t_m))| = |b_{j'}(\boldsymbol{\beta}(t_m))| > |b_k(\boldsymbol{\beta}(t_m))|$ for any $k \notin \mathcal{A}_{t_m}$ and any $j, j' \in \mathcal{A}_{t_m}$.

At the end of the $(p-1)$th CoxLAR-ridge step in Step 2 of Algorithm 1, all predictors are active. Then, in Step 3, the CoxLAR solution path moves along a direction such that the absolute values of the first-order partial derivatives decrease at the same speed until all the first-order partial derivatives are exactly zero, which happens at $t = 0$. The solution at $t = 0$ exactly corresponds to the full solution $\operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$, just as the LAR solution ends at the full OLS estimate. This completes our CoxLAR-ridge solution path. When the ridge term in $L(\boldsymbol{\beta})$ is exactly zero by setting $\gamma = 0$, we are essentially working directly with the original log partial likelihood function and the CoxLAR-ridge is also the CoxLAR in this case.

Remark 1. Note that the instantaneous path updating direction is given by $-(\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}(\boldsymbol{\beta}(t)))^{-1}\operatorname{sign}(\boldsymbol{b}_{\mathcal{A}_t}(\boldsymbol{\beta}(t)))$. For least squares regression, the objective function is exactly quadratic and thus $\boldsymbol{M}_{\mathcal{A}_t,\mathcal{A}_t}$ depends only on the active set $\mathcal{A}_t$, but not on the current solution $\boldsymbol{\beta}_{\mathcal{A}_t}(t)$. Note that $\operatorname{sign}(\boldsymbol{b}_{\mathcal{A}_t}(\boldsymbol{\beta}(t)))$ does not change in a small neighborhood of t. This implies that, within a small neighborhood of t, the instantaneous path updating direction is the same for least squares regression. This leads to the piecewise linearity of the LAR path (Efron et al. (2004)) and in a more general setting (Rosset and Zhu (2007)).

Algorithm 1. CoxLAR(-ridge) for the Cox's proportional hazards model.

1. Initialize by setting $t_0 = -\max_{j=1,\ldots,p}|b_j(\mathbf{0})|, \beta(t_0) = \mathbf{0}$, and $\mathcal{A}_{t_0} = \left\{\operatorname{argmax}_{1\le j\le p}|b_j(\mathbf{0})|\right\}$.
2. For $m = 0, 1, \ldots, p-2$, take the tentative solution path using

$$\tilde{\beta}_{\mathcal{A}_{t_m}}(t) = \beta_{\mathcal{A}_{t_m}}(t_m) - \int_{t_m}^{t}(\boldsymbol{M}_{\mathcal{A}_{t_m},\mathcal{A}_{t_m}}(\tilde{\beta}(\tau)))^{-1}\operatorname{sign}(\boldsymbol{b}_{\mathcal{A}_{t_m}}(\tilde{\beta}(\tau)))d\tau \text{ and } \tilde{\beta}_{\mathcal{A}_{t_m}^c}(t) = \mathbf{0}$$

for $t \ge t_m$. Let $t_{m+1} = \min_{j\notin\mathcal{A}_{t_m}} T_j$, where

$$T_j = \min\left\{t > t_m : \left|b_j(\tilde{\beta}(t))\right| \ge \left|b_k(\tilde{\beta}(t))\right| \text{ for some } k \in \mathcal{A}_{t_m}\right\} \text{ for } j \notin \mathcal{A}_{t_m}.$$

Update the solution path with

$$\beta_{\mathcal{A}_{t_m}}(t) = \beta_{\mathcal{A}_{t_m}}(t_m) - \int_{t_m}^{t}(\boldsymbol{M}_{\mathcal{A}_{t_m},\mathcal{A}_{t_m}}(\beta(\tau)))^{-1}\operatorname{sign}(\boldsymbol{b}_{\mathcal{A}_{t_m}}(\beta(\tau)))d\tau \text{ and } \beta_{\mathcal{A}_{t_m}^c}(t) = \mathbf{0}$$

for $t \in [t_m, t_{m+1}]$. Set $\mathcal{A}_t = \mathcal{A}_{t_m}$ for $t \in [t_m, t_{m+1})$ and $\mathcal{A}_{t_{m+1}} = \mathcal{A}_{t_m} \cup [j \notin \mathcal{A}_{t_m} : T_j = t_{m+1}\}$.
3. At the end of Step 2, $\mathcal{A}_{t_{p-1}}$ is $\{1, 2, \ldots, p\}$. Next take

$$\beta(t) = \beta(t_{p-1}) - \int_{t_{p-1}}^{t}(\boldsymbol{M}_{\mathcal{A}_{t_{p-1}},\mathcal{A}_{t_{p-1}}}(\beta(\tau)))^{-1}\operatorname{sign}(\boldsymbol{b}_{\mathcal{A}_{t_{p-1}}}(\beta(\tau)))d\tau,$$

and $\mathscr{A}_t = \{1, 2, \ldots, p\}$ for t between t_{p-1} and $t_p = 0$.

2.2. Cox-LASSO modification

Efron et al. (2004) showed that the whole LASSO regularized least squares regression solution path can be obtained by a slight modification of the LAR. This is confirmed by Wu (2011). Next we define our Cox-LASSO modification, and prove that the CoxLAR-ridge with the Cox-LASSO modification produces the whole elastic net regularized solution path for the Cox's proportional hazards model by noting that adding another LASSO penalty into $L(\boldsymbol{\beta})$ leads to the elastic net penalized log partial likelihood function in (2.2).

Consider the LASSO regularized counterpart of (2.3),

$$\min_{\beta_0,\beta} L(\beta)+\lambda\sum_{j=1}^{p}|\beta_j| \quad (2.16)$$

which is exactly the same as (2.2), and is equivalent

$$\min_{\beta_0,\beta} L(\beta) \text{ subject to} \sum_{j=1}^{p}|\beta_j| \le s, \quad (2.17)$$

where two regularization parameters $\lambda \ge 0$ and $s \ge 0$ are in some one-to-one correspondence.

Let $\hat{\boldsymbol{\beta}}$ be a LASSO solution to (2.16). We can show that the sign of any nonzero component $\hat{\beta}_j$ must disagree with the sign of the current derivative $b_j(\hat{\boldsymbol{\beta}})$, see Lemma 2 in Section 3.

Suppose $t = t^*$ at the end of a CoxLAR-ridge step and that we have a new active set $\mathscr{A}$. At the next CoxLAR-ridge step with $t \in [t^*, T]$ for some T to be determined, our solution path moves along the tentative solution path

$$\tilde{\beta}_{\mathscr{A}^*}(t)=\beta_{\mathscr{A}^*}(t^*) - \int_{t^*}^{t}(\boldsymbol{M}_{\mathscr{A}^*,\mathscr{A}^*}(\tilde{\beta}(\tau)))^{-1}\operatorname{sign}(\boldsymbol{b}_{\mathscr{A}^*}(\tilde{\beta}(\tau)))d\tau \text{ and } \tilde{\beta}_{(\mathscr{A}^*)^c}(t)=\mathbf{0} \quad (2.18)$$

for $t \ge t^*$. The end point T is given by $T = \min_{j\notin\mathscr{A}} T_j$, where

$$T_j=\min\left\{t>t^*: \left|b_j(\tilde{\beta}(t))\right| \ge \left|b_k(\tilde{\beta}(t))\right| \text{ for some } k \in \mathscr{A}^*\right\} \text{ for } j \notin \mathscr{A}^*.$$

.

For some $j \in \mathscr{A}$, $\tilde{\beta}_j(t)$ may have changed sign at some point between t^* and T, in which case the sign restriction given in Lemma 2 must have been violated. We set $S_j = \min\{t \in (t^*, \infty) : \tilde{\beta}_j(t) = 0\}$ for $j \in \mathscr{A}$, where $\tilde{\beta}_j(t)$ is the jth component of $\tilde{\boldsymbol{\beta}}(t)$ defined by (2.18). If $S = \min_{j\in\mathscr{A}} S_j < T$, $\tilde{\boldsymbol{\beta}}(T)$ defined by (2.18) cannot be a LASSO regularized solution to (2.16) since the sign restriction in Lemma 2 has already been violated. The Cox-LASSO modification can be applied to ensure that we can get the LASSO regularized solution to (2.16).

Cox-LASSO modification: If $S < T$, stop the ongoing CoxLAR-ridge step at S and remove j from the active set $\mathscr{A}$ by setting $\mathscr{A}_S = \mathscr{A}_{t^*} \setminus \{j\}$, where j is chosen such that $S_j = S$. At the new transition point S, the new path updating direction $\frac{d}{dt}\beta(t)$ is calculated using (2.8) based on the new active predictor set $\mathscr{A} \setminus \{j\}$.

Theorem 1 guarantees that the Cox-LASSO modification leads to the LASSO regularized solution path to (2.16), which is the LASSO regularized log partial likelihood (CoxLASSO) when $\gamma = 0$, and the elastic net regularized log partial likelihood (CoxEN) when $\gamma > 0$. We use CoxLARS to refer to CoxLAR, CoxLARridge, CoxLASSO, and CoxEN.

Note that at each transition point of our CoxLARS solution path, two kinds of event can happen: either an inactive predictor joins the active predictor set or an active predictor is removed from the active predictor set. As in Efron et al. (2004), we assume a "one at a time" condition holds. With the "one at a time" condition, at each transition point t^* only a single event can happen, namely, either one inactive predictor variable becomes active or one currently active predictor variable becomes inactive.

Theorem 1. *Under the Cox-LASSO modification, and assuming the "one at a time" condition, the CoxLAR-ridge algorithm yields the LASSO regularized solution path to* (2.16).

Remark 2. For simplicity we make the "one at a time" assumption. But, even when the "one at a time" condition does not hold, a CoxLASSO/CoxEN solution path is still available. The same discussion in Efron et al. (2004) applies. In applications, some slight jittering may be applied, if necessary, to ensure the "one at a time" condition holds.

2.3. Updating via ODE

Our CoxLARS algorithm involves an essential piecewise updating step

$$\tilde{\beta}_{\mathcal{A}_{t^*}}(t)=\beta_{\mathcal{A}_{t^*}}(t^*) - \int_{t^*}^{t}(\boldsymbol{M}_{\mathcal{A}_{t^*},\mathcal{A}_{t^*}}(\tilde{\beta}(\tau)))^{-1}\mathrm{sign}(\boldsymbol{b}_{\mathcal{A}_{t^*}}(\tilde{\beta}(\tau)))d\tau \quad \text{and } \tilde{\beta}_{\mathcal{A}^c_{t^*}}(t)=\mathbf{0} \qquad (2.19)$$

beginning at a transition point t^* with solution $\boldsymbol{\beta}(t^*)$ and active predictor set $\mathcal{A}_{t^*}$.

Note that the piecewise solution path (2.19) can be easily obtained by setting $\tilde{\beta}_j(t) = 0$ for $j \notin \mathcal{A}_{t^*}$ and $t > t^*$, and solving the following ordinary differential equation (ODE) system

$$\frac{d}{dt}\tilde{\beta}_{\mathcal{A}_{t^*}}(t)= -(\boldsymbol{M}_{\mathcal{A}_{t^*},\mathcal{A}_{t^*}}(\tilde{\beta}(t)))^{-1}\mathrm{sign}(\boldsymbol{b}_{\mathcal{A}_{t^*}}(\tilde{\beta}(t)))$$

with initial value condition $\tilde{\boldsymbol{\beta}}_{\mathcal{A}_{t^*}}(t)|_{t=t^*} = \boldsymbol{\beta}_{\mathcal{A}_{t^*}}(t^*)$. This is a standard initial-value ODE system and there are many efficient methods to solve it, for example Euler method, backward Euler method, midpoint method, and the family of Runge-Kutta methods, among many others. The commonly used member of the Runge-Kutta method family is the fourth-order Runge-Kutta method. See Atkinson, Han, and Stewart (2009) for a comprehensive introduction to the methods for solving ordinary differential equations. Our numerical examples employ the Matlab ODE solver "ODE45", which exactly implements the fourth-order Runge-Kutta method.

3. Properties of CoxLARS

In this section, we establish some properties of our CoxLARS path, and prove Theorem 1.

With the "one at a time" condition, at each transition point t^* either one inactive predictor becomes active or one active predictor becomes inactive. For the first type, the active set changes from $\mathcal{A}$ to $\mathcal{A} = \mathcal{A} \cup \{j^*\}$ for some $j^* \notin \mathcal{A}$ We show in Lemma 1 that this new active predictor joins in a "correct" manner. Lemma 1 applies to CoxLARS.

Lemma 1. *For any transition point t* during the CoxLARS solution path, if predictor j* is the only addition to the active set at t* with* $\boldsymbol{\beta}(t^*)$ *and active set changing from* $\mathcal{A}$ *to* $\mathcal{A} = \mathcal{A} \cup \{j^*\}$*, then the path updating direction* $\frac{d}{dt}\beta_{\mathcal{A}^*}(t)$ *at t* has its j*th component disagreeing in sign with the current derivative* $b_{j^*}(\boldsymbol{\beta}(t^*))$.

Lemma 1 is a key property for showing that the Cox-LASSO modification leads to the LASSO or elastic net regularized log partial likelihood solution path in that Lemma 1 ensures that, at any transition point, the new predictor variable enters in a "correct" manner. This "correct" manner is required by the LASSO penalty as is seen in Lemma 2.

Next we extend Lemmas 7–10 of Efron et al. (2004) to the Cox's proportional hazards model. Our Lemmas 2–5 concern properties of the LASSO regularized solution path for (2.16) or equivalently (2.17), and as a result they lead to the proof of Theorem 1. For any $s \geq 0$, we denote the unique solution of (2.17) by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(s)$, which is continuous in s; uniqueness is due to the convexity of $\sum_{j=1}^{p} |\beta_j|$ and the strict convexity of $L(\boldsymbol{\beta})$. Throughout, we use the hat notation to designate a solution of (2.16), equivalently (2.17). For any $s \geq 0$, let $\mathcal{N}_s \equiv \mathcal{N}(\hat{\boldsymbol{\beta}}(s)) \triangleq \{j : \hat{\beta}_j(s) \neq 0\}$ denote the index set of nonzero components of $\hat{\boldsymbol{\beta}}(s)$. Our goal is to show that the nonzero set $\mathcal{N}_s$ is also the active predictor set that determines the CoxLARS path updating direction.

Let $\hat{\boldsymbol{\beta}}$ be a solution of (2.16). We show that any non-zero component $\hat{\beta}_j$ must disagree in sign with the current first-order derivative.

Lemma 2. *A LASSO regularized solution* $\hat{\boldsymbol{\beta}}$ *to (2.16) satisfies* $\text{sign}(\hat{\boldsymbol{\beta}}_j) = -\text{sign}\,(b_j(\hat{\boldsymbol{\beta}}))$ *for any* $j \in \mathcal{N}(\hat{\boldsymbol{\beta}})$.

Let $\mathcal{S}$ be an open interval of the s axis, with infimum $\underline{s}$, within which the nonzero set $\mathcal{N}_s$ of $\hat{\boldsymbol{\beta}}(s)$ remains constant, $\mathcal{N}_s = \mathcal{N}$ for $s \in \mathcal{S}$ and some $\mathcal{N}$

Lemma 3. *For* $s \in \{\underline{s}\} \cup \mathcal{S}$ *the LASSO regularized estimate* $\hat{\boldsymbol{\beta}}(s)$ *of (2.17) updates along the CoxLARS path updating direction.*

Lemma 4. *For an open interval* $\mathcal{S}$ *with a constant nonzero set* $\mathcal{N}$ *during the LASSO regularized path* $\hat{\boldsymbol{\beta}}(s)$ *of (2.17), let* $\underline{s} = \inf(\mathcal{S})$*. Then for* $s \in \mathcal{S} \cup \{\underline{s}\}$*, the first-order derivatives of* $L(\boldsymbol{\beta})$ *at* $\hat{\boldsymbol{\beta}}(s)$ *satisfy* $|b_j(\hat{\boldsymbol{\beta}}(s))| = \max_{l=1,2,\ldots,p} |b_l(\hat{\boldsymbol{\beta}}(s))|$ *for* $j \in \mathcal{N}$ *and* $|b_j(\hat{\boldsymbol{\beta}}(s))| \leq \max_{l=1,2,\ldots,p} |b_l(\hat{\boldsymbol{\beta}}(s))|$ for $j \notin \mathcal{N}$

Let $\underline{s}$ denote such a point, $\underline{s} = \inf(\mathcal{S}$ as in Lemma 4, with the LASSO regularized solution $\hat{\boldsymbol{\beta}}$ to (2.17), current derivatives $b_j(\hat{\boldsymbol{\beta}})$ for $j = 1, 2, \ldots, p$, and maximum absolute derivative $\hat{D}(\hat{\boldsymbol{\beta}}) = \max_{j=1,2,\ldots,p} |b_j(\hat{\boldsymbol{\beta}})|$. Let $\mathcal{A}_1 = \{j : \hat{\boldsymbol{\beta}}_j \neq 0\}$, $\mathcal{A}_0 = \{j : \hat{\boldsymbol{\beta}}_j = 0 \text{ and } |b_j(\hat{\boldsymbol{\beta}})| = \hat{D}(\hat{\boldsymbol{\beta}})\}$, and $\mathcal{A}_{10} = \mathcal{A}_1 \cup \mathcal{A}_0$. Take $\boldsymbol{\beta}^{(\theta)} = \hat{\boldsymbol{\beta}} + \theta \boldsymbol{d}$ for some vector $\boldsymbol{d} \in \mathbb{R}^p$, $T(\theta) = L(\boldsymbol{\beta}^{(\theta)})$, and $S(\theta) = \sum_{j=1}^{p} |\beta_j^{(\theta)}|$. Let $\dot{S}(\theta) = \frac{d}{d\theta} S(\theta)$, $\dot{T}(\theta) = \frac{d}{d\theta} T(\theta)$, and $\ddot{T}(\theta) = \frac{d^2}{d\theta^2} T(\theta)$.

Lemma 5. *At* $\underline{s}$*, we have*

$$R(\boldsymbol{d}) = -\frac{\dot{T}(0)}{\dot{S}(0)} \leq \widehat{D}(\widehat{\beta}), \quad (3.1)$$

with equality only if $d_j = 0$ *for* $j \in \mathcal{A}_{10}^c$ *and* $\text{sign}(d_j) = -\text{sign}(b_j(\hat{\boldsymbol{\beta}}))$ *for* $j \in \mathcal{A}_0$*. If so,*

$$\ddot{T}(0)=\boldsymbol{d}_{\mathscr{A}_{10}}^{T}\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathscr{A}_{10}}. \quad (3.2)$$

Lemma 5 implies that, at any transition point, the active predictor set of the LASSO regularized solution to (2.17) is a subset of $\mathscr{A}_{10}$. With the LASSO regularization, we are minimizing $L(\boldsymbol{\beta})$ subject to a constraint on the one norm of $\boldsymbol{\beta}$. In a small neighborhood $\hat{\boldsymbol{\beta}}+\theta\boldsymbol{d}$ around $\hat{\boldsymbol{\beta}}$, we are minimizing $T(\theta)$ subject to an upper bound on $S(\theta)$. The first part of Lemma 5 implies that the instantaneous relative changing rate of $T(\theta)$ and $S(\theta)$ is $\geq -\hat{D}(\hat{\boldsymbol{\beta}})$. For $\boldsymbol{\beta}^{(\theta)}$, its one-norm $S(\theta)$ is increasing in θ as long as $-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|+\sum_{j\in\mathscr{A}_{10}^c}|d_j|>0$ and the best instantaneous relative changing rate is achieved for moving along $\hat{\boldsymbol{\beta}}+\theta\boldsymbol{d}$ as long as $d_j=0$ for $j\in\mathscr{A}_{10}^c$ and $\text{sign}(d_j)=-\text{sign}(b_j(\hat{\boldsymbol{\beta}}))$ for $j\in\mathscr{A}_0$. In particular, $\text{sign}(d_j)=-\text{sign}(b_j(\hat{\boldsymbol{\beta}})$ for $j\in\mathscr{A}_0$ requires that the coefficient of any new active predictor variable should disagree in sign with the corresponding current first-order partial derivative. This is ensured by Lemma 1 and the "one at a time" condition.

The second part of Lemma 5 provides second-order information on the relative change of $T(\theta)$ with respect to $S(\theta)$. As we only care about direction, assume $\dot{S}(0)=-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|=\Delta$ for some $\Delta>0$. Note that $T(\theta)\approx T(0)+\dot{T}(0)\theta+\frac{1}{2}\ddot{T}(0)\theta^2+o(\theta^2)$. Then we need to find the most efficient direction $\boldsymbol{d}$ to decrease $T(\theta)$ among all possible direction $\boldsymbol{d}$ satisfying $\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|=\Delta$ and $\text{sign}(d_j)=-\text{sign}(b_j(\hat{\boldsymbol{\beta}}))$ for $j\in\mathscr{A}_0$. In terms of the second-order information, we need to solve

$$\min \quad \boldsymbol{d}_{\mathscr{A}_{10}}^{T}\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathscr{A}_{10}} \quad (3.3)$$

$$\text{subject to } -\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|=\Delta,$$

$$\text{sign}(d_j)=-\text{sign}(b_j(\widehat{\beta})) \text{ for } j\in\mathscr{A}_0$$

with a fixed $\Delta>0$ to select the optimal solution updating direction $\boldsymbol{d}$. It turns out that the optimal solution to (3.3) is exactly given by our CoxLARS path updating direction as proved next.

Lemma 6. *Our CoxLARS path updating direction* (2.8) *solves* (3.3).

4. Numerical Examples

In this section, we use numerical examples to demonstrate how the extension CoxLARS works. In our implementation we first calculate t_0, then set $\delta_t=-t_0/K$, where K is some large positive number. In our examples we use $K=2{,}000$. In addition to the transition points t_ks, we evaluate the solution over our solution path at a grid of size δ_t. More specifically, for each piece of our solution path over $[t_k, t_{k+1}]$, we calculate our solution $\boldsymbol{\beta}(t)$ at $t=t_k+m\delta_t$

for $m = 1, 2, \ldots, \lfloor (t_{k+1} - t_k)/\delta_t \rfloor$, where $\lfloor a \rfloor$ denotes the integer part of a, even though the CoxLARS solution paths are defined for any $t \in [t_0, 0]$.

Example 1. We use a simulated dataset to demonstrate that the true LASSO regularized solution path is *not* piecewise linear. We set $p = 3$ and $n = 40$. The predictor covariates were generated as $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is the variance-covariance matrix with (i, j)th element 1 if $i = j$, and 0.9 otherwise. Conditional on $\boldsymbol{X} = (x_1, x_2, x_3)^T$, the lifetime was generated from model (1.1) with a constant baseline hazard function $h_0(y) = 1$ and true regression coefficient vector $\boldsymbol{\beta} = (2, -2, 2.5)^T$. The censoring time was uniformly distributed over [0, 8] and the corresponding censoring rate is 32.3%. We applied the CoxLASSO (with a ridge term $\gamma = 0$).

The CoxLASSO solution path is shown by the solid lines in the top left panel of Figure 2. The dashed straight lines are obtained by connecting the solutions at the transition points. The true LASSO regularized solution path is clearly not piecewise linear. The first-order partial derivatives along the CoxLASSO solution path are shown in the top right panel of Figure 2. The absolute value of the first-order partial derivatives along the CoxLASSO solution path are shown in the bottom two panels of Figure 2. with different horizontal axis scales. The bottom left panel is plotted with respect to the one-norm of $\boldsymbol{\beta}(t)$ while the right panel uses t. A straight diagonal line is observed in the bottom right panel since our CoxLARS ensures that $t = -\max_{j=1,\ldots,p} |b_j(\boldsymbol{\beta}(t))|$.

Example 2. Here we demonstrate how the Cox-LASSO modification leads to the CoxLASSO path when $\gamma = 0$ and the CoxEN path when $\gamma > 0$. We chose $n = 200$ and $p = 12$. The predictor covariates $\boldsymbol{X}$ were generated from $N(\mathbf{0}, \Sigma)$, with (i, j) element of Σ being 1 when $i = j$, 0.3 when $1 \le i, j \le 11$ and $|i-j| = 1$, $(-0.18)^{i+1}$ when $j = 12$ and $1 \le i \le 11$, and $(-0.18)^{j+1}$ when $i = 12$ and $1 \le j \le 11$. Conditional on covariates, the lifetime was generated from model (1.1) with $h_0(y) = 1$ and true coefficient vector given by $(-0.8, 1.6, -0.8, 1, 0, 1.5, -1.2, 3, 0, 0, 0, 0.5)^T$. The censoring time was generated from Uniform[0, 10] leading to a censoring rate of 30.5%. In general, the Cox-LASSO modification may not have any effect and conseqently the CoxLAR and CoxLASSO paths are exactly the same. We designed Example 2 to show the effect of the Cox-LASSO modification.

CoxLARS solution paths are shown in Figure 3. When $\gamma = 0$, solution paths of the CoxLAR and CoxLASSO are given in the top left and bottom left panels, respectively. The CoxLAR path shows that coefficient of variable X_{12} switches sign between the 9th and 10th transition points. Thus a new transition point is added to the CoxLASSO solution path, in which the coefficient corresponding to X_{12} is kept at zero between the 10th and 13th transition points. When we add a small ridge term by setting $\gamma = 0.2$, the corresponding paths are shown in the right panels of Figure 3. A similar phenomenon is observed.

Example 3. The primary biliary cirrhosis data were collected in the Mayo Clinic trial on primary biliary cirrhosis of liver conducted between 1974 and 1984, see Therneau and Grambsch (2001). This study included a total of 424 patients. Clinical, biochemical, serological, and histological parameters were collected for each patient. Before the end of the follow-up, 125 patients died. We study the dependency of the survival time on seventeen covariates: continuous variables are age (in years), albumin (albumin in g/dl), alk (alkaline phosphatase in units/litre), bili (serum bilirubin in mg/dl), chol (serum cholesterol in mg/dl), copper (urine copper in g/day), platelets (platelets per cubic ml/1,000), prothrombin (prothrombin time in seconds), sgot (liver enzyme in units/ml), and trig (triglycerides in mg/dl); categorical variables are ascites (0 denotes absence of ascites and 1 denotes presence of ascites), edema (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema, and 1 denotes unsuccessfully treated oedema), hepatom (0 denotes absence of hepatomegaly

and 1 denotes presence of hepatomegaly), sex (0 denotes male and 1 denotes female), spiders (0 denotes absence of spiders and 1 denotes presence of spiders), stage (histological stage of disease, graded 1, 2, 3, or 4), and treatment (1 for control and 2 for treatment). See Dickson et al. (1989) for more detailed information.

After excluding patients with any missing value, there are 276 patients. Out of these 276 patients, 111 died before the end of the follow-up. We standardized each predictor variable to have mean zero and variance one. CoxLARS was applied to the standardized data with all seventeen variables included. With ridge parameter $\gamma = 0$, the CoxLAR and CoxLASSO gave the same solution path, see Figure 4.

5. Discussion

In this work, we have proposed the extension CoxLAR(-ridge) of the LAR to handle Cox's proportional hazards model. Our CoxLAR(-ridge) solution paths are based on ODE systems. Results show that a Cox-LASSO modification on CoxLAR(-ridge) leads to the exact solution of the corresponding LASSO regularized solution path. As the solution path propagates according to ODE systems, it allows us to develop a solution path package using efficient ODE solvers.

LARS is very attractive due to its speed that is possible because the corresponding path is piecewise linear. However when it comes to the Cox's proportional hazards model, the solution path is not piecewise linear due to the nature of the log partial likelihood, as demonstrated by Example 1. This makes the implementation of the CoxLARS more difficult. Currently we have implemented the primitive version of our algorithm using the fourth-order Runge-Kutta method, which works fairly well. In addition, it is commonly assumed that the regression coefficients are sparse in the high dimension variable selection literature. Consequently there is not much need for us to compute the whole solution path. A BIC criterion may be combined as we progress along the solution path to identify an optimal solution and terminate our solution path algorithm thereafter, as done in Wu (2011).

Acknowledgments

The author thanks Jianqing Fan and Chuanshu Ji for mentoring and longtime encouragement. The author also thanks Dennis Boos, Jingfang Huang, Yufeng Liu, John Monahan, and Leonard Stefanski for helpful comments and encouragement. This work is supported in part by NSF grant DMS-0905561, DMS-1055210, NIH/NCI grant R01-CA149569, and NCSU Faculty Research and Professional Development Award grant. The content is solely the author's responsibility and does not necessarily represent the official views of the NSF, NIH, or NCI.

Appendix

Proof of Lemma 1. The new path updating direction defined using the new active predictor set $\mathcal{A}$ is given by $\frac{d}{dt}\beta_{\mathcal{A}^*}(t) = -\boldsymbol{M}_{\mathcal{A}^*,\mathcal{A}^*}(\beta(t^*))^{1}\text{sign}(\boldsymbol{b}_{\mathcal{A}^*}(\beta(t^*)))$. Using the formula for inverting a block matrix, the j^*th component of our path updating direction $\left(\frac{d}{dt}\beta(t)\right)|_{t^*}$ is given by

$$\frac{1}{\eta}[\boldsymbol{M}_{j^*,\mathcal{A}}(\beta(t^*)) - \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\beta(t^*))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t^*))) - \text{sign}(b_{j^*}(\beta(t^*)))], \quad \text{(A.1)}$$

where $\eta = \boldsymbol{M}_{j^*,j^*}(\boldsymbol{\beta}(t^*)) - \boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*))\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1}\boldsymbol{M}_{\mathcal{A},j^*}(\boldsymbol{\beta}(t^*)) > 0$ in that $\boldsymbol{M}(\boldsymbol{\beta})$ is positive definite when $n > p$, and $x^{(j)}$, $j = 1, 2, \ldots, p$ are linearly independent. The first term in (A.1) involves $\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*)))$, which is exactly the opposite of the path

updating direction calculated at t^* using the old active set $\mathcal{A}$ by ignoring the addition of predictor variable j^*.

Consider ignoring the new active variable j^* and updating path along the path updating direction evaluated by the old active predictor set $\mathcal{A}$ This leads to another solution path piece $\bar{\boldsymbol{\beta}}(t)$ defined by

$$\bar{\beta}_{\mathcal{A}}(t)=\beta_{\mathcal{A}}(t^*) - \int_{t^*}^{t} \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\beta}(\tau))^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\beta}(\tau)))d\tau \text{ and } \bar{\beta}_{\mathcal{A}^c}(t)=\mathbf{0}$$

when t is inside a small neighborhood $[t^* - \Delta_b, t^* + \Delta_t]$. The neighborhood is chosen such that both solution component $\bar{\beta}_j(t)$ and the first-order partial derivative $b_j(\bar{\boldsymbol{\beta}}(t))$ do not change sign for $t \in [t^* - \Delta_b, t^* + \Delta_t]$ and $j \in \mathcal{A}$ Consequently when $t \in [t^* - \Delta_b, t^* + \Delta_t]$,

$$\frac{d}{dt} b_j(\bar{\beta}(t))= - \operatorname{sign}(b_j(\bar{\beta}(t)))= - \operatorname{sign}(b_j(\bar{\beta}(t^*)))$$

for $j \in \mathcal{A}$ due to (2.5) and (2.8). Note that, for $t \in [t^* - \Delta_b, t^* + \Delta_t]$,

$$\frac{d}{dt} b_{j^*}(\bar{\beta}(t))=\sum_{j=1}^{p} m_{j^*j}(\bar{\beta}(t)) \frac{d}{dt}\bar{\beta}_j(t)= - \boldsymbol{M}_{j^*,\mathcal{A}}(\bar{\beta}(t)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\beta}(t))^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\beta}(t))) \quad \text{(A.2)}$$

due to the definition of $\bar{\boldsymbol{\beta}}(t)$ (because $\bar{\beta}_j(t) = 0$ for $j \notin \mathcal{A}$ and $t \in [t^* - \Delta_b, t^* + \Delta_t]$).

Recall that for $t \in [t^* - \Delta_b, t^*]$, $\boldsymbol{\beta}(t) = \bar{\boldsymbol{\beta}}(t)$, and our CoxLARS solution matches $\bar{\boldsymbol{\beta}}(t)$ exactly. Note that our CoxLARS definition implies that

$$|b_{j^*}(\bar{\beta}(t))|<|b_j(\bar{\beta}(t))| \quad \text{(A.3)}$$

for any $j \in \mathcal{A}$ and $t \in [t^* - \Delta_b, t^*)$. This means that the predictor variable j^* has a smaller absolute value of the first-order partial derivative than active predictors in $\mathcal{A}$ for $t \in [t^* - \Delta_b, t^*)$ and that it catches up with active predictors in $\mathcal{A}$ at t^*, noting the definition of j^*.

Lemma 1 can be proved by contradiction. If our claim is wrong, then

$$[\boldsymbol{M}_{j*,\mathcal{A}}(\beta(t^*)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\beta(t^*))^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t^*))) - \operatorname{sign}(b_{j^*}(\beta(t^*)))] \operatorname{sign}(b_{j^*}(\beta(t^*)))>0$$

due to (A.1) and the fact that $\eta > 0$. This means that

$$\boldsymbol{M}_{j*,\mathcal{A}}(\beta(t^*)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\beta(t^*))^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathcal{A}}(\beta(t^*))) \operatorname{sign}(b_{j^*}(\beta(t^*)))>1.$$

. The fact that $\bar{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t)$ for $t \in [t^* - \Delta_b, t^*]$ implies that there exists some $0 < \varepsilon < \Delta_t$ such that

$$\boldsymbol{M}_{j*,\mathcal{A}}(\bar{\beta}(t)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\beta}(t))^{-1} \operatorname{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\beta}(t))) \operatorname{sign}(b_{j^*}(\bar{\beta}(t)))>1 \text{ for } t \in (t^* - \varepsilon, t^*) \quad \text{(A.4)}$$

due to continuity. By noting (A.2) and $\frac{d}{dt} b_j(\bar{\beta}(t))= - \operatorname{sign}(b_j(\bar{\beta}(t)))$ for $j \in \mathcal{A}$ and $t \in (t^* - \varepsilon, t^*)$, (A.4) contradicts the conclusion that the predictor variable j^* has a smaller absolute

value of the first-order partial derivative than active predictors in $\mathscr{A}$ for $t \in [t^* - \Delta_t, t^*)$ and that it catches up with active predictors in $\mathscr{A}$ at t^*. This completes our proof.

Proof of Lemma 2. For any $j \in \mathscr{N}(\hat{\boldsymbol{\beta}})$, taking differentiation of the objective function in (2.16) with respect β_j, we get

$$\frac{\partial}{\partial \beta_j} L(\beta) + \lambda \operatorname{sign}(\beta_j) \quad \text{(A.5)}$$

which has to be equal to zero at $\hat{\boldsymbol{\beta}}$ because $\hat{\boldsymbol{\beta}}$ is the corresponding optimal solution. This completes the proof by noting that $\lambda \geq 0$ and, when $\lambda = 0$, $\frac{\partial}{\partial \beta_j} L(\beta) = 0$ for all j.

Proof of Lemma 3. Note that $\hat{\boldsymbol{\beta}}(s)$ is the optimal solution to (2.17) and has a nonzero set $\mathscr{N}_s$ that is constant for $s \in \mathscr{S}$ say $\mathscr{N}$ Then $\bar{\boldsymbol{\beta}}_{\mathscr{N}}(s)$ also minimizes

$$L(\widehat{\beta}_{\mathscr{N}}) \triangleq \sum_{i=1}^{n} \delta_i \boldsymbol{x}_{i\mathscr{N}}^T \widehat{\beta}_{\mathscr{N}} - \sum_{i=1}^{n} \delta_i \log \left(\sum_{m \in R_i} \exp(\boldsymbol{x}_{m\mathscr{N}}^T \widehat{\beta}_{\mathscr{N}}) \right) + \frac{\gamma}{2} \widehat{\beta}_{\mathscr{N}}^T \widehat{\beta}_{\mathscr{N}} \quad \text{(A.6)}$$

subject to

$$-s_{\mathscr{N}}^T \widehat{\beta}_{\mathscr{N}} = s \text{ and } \operatorname{sign}(\widehat{\beta}_j) = -s_j \text{ for } j \in \mathscr{N}, \quad \text{(A.7)}$$

where $s_j = \operatorname{sign}(b_j(\hat{\boldsymbol{\beta}}(s)))$, $j = 1, 2, \ldots, p$, denotes the sign of the current first-order partial derivatives, $\boldsymbol{s} = (s_1, s_2, \ldots, s_p)^T$, and the second sign constraint is due to Lemma 2. Here $\boldsymbol{x}_{i\mathscr{N}}$ is the sub-vector of $\boldsymbol{x}_i$ with index in $\mathscr{N}$ Note that the inequality constraint in (2.17) can be replaced by the constraint $\sum_{j=1}^{p} |\beta_j| = s$ as long as s is less than the one-norm of the full solution $\operatorname{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$. This justifies (A.7). Note further that the optimal solution $\hat{\boldsymbol{\beta}}_{\mathscr{N}}(s)$ is strictly inside the simplex (A.7) since $\hat{\beta}_j(s) \neq 0$ for $j \in \mathscr{N}$ and $s \in \mathscr{S}$ This, in combination with the strict convexity of the object function $L(\hat{\boldsymbol{\beta}}_{\mathscr{N}})$, implies that the condition $\operatorname{sign}(\hat{\beta}_j) = -s_j$ for $j \in \mathscr{N}$ can be dropped. Consequently $\hat{\boldsymbol{\beta}}_{\mathscr{N}}(s)$ solves

$$\min L(\widehat{\beta}_{\mathscr{N}}) \text{ subject to } -\boldsymbol{s}_{\mathscr{N}}^T \widehat{\beta}_{\mathscr{N}} = s.$$

By introducing a Lagrange multiplier λ, we get

$$\frac{\partial}{\partial \widehat{\beta}_{\mathscr{N}}} L(\widehat{\beta}_{\mathscr{N}}) - \lambda \mathbf{s}_{\mathscr{N}}, \quad \text{(A.8)}$$

which is equal to $\mathbf{0}$ at $\hat{\boldsymbol{\beta}}_{\mathscr{N}} = \hat{\boldsymbol{\beta}}_{\mathscr{N}}(s)$ because $\hat{\boldsymbol{\beta}}_{\mathscr{N}}(s)$ is the corresponding optimal solution.

Now consider two different values $s^{(1)}$ and $s^{(2)}$ in $\mathscr{S}$ with $\underline{s} < s^{(1)} < s^{(2)}$. The corresponding Lagrange multiplier are denoted by $\lambda^{(1)}$ and $\lambda^{(2)}$, and they satisfy $\lambda^{(1)} > \lambda^{(2)}$. Putting them into (A.8) and differencing, we get

$$\frac{\partial}{\partial \widehat{\beta}_{\mathscr{N}}} L(\widehat{\beta}_{\mathscr{N}})\Big|_{\widehat{\beta}_{\mathscr{N}}(s^{(2)})} - \frac{\partial}{\partial \widehat{\beta}_{\mathscr{N}}} L(\widehat{\beta}_{\mathscr{N}})\Big|_{\widehat{\beta}_{\mathscr{N}}(s^{(1)})} = (\lambda^{(2)} - \lambda^{(1)}) \boldsymbol{s}_{\mathscr{N}}. \quad \text{(A.9)}$$

Note that $\hat{\boldsymbol{\beta}}_{\mathscr{N}^c}(s) = \mathbf{0}$ for any $s \in \mathscr{S}$ Thus (A.9) is the same as

$$\boldsymbol{b}_{\mathcal{N}}(\widehat{\beta}(s^{(2)})) - \boldsymbol{b}_{\mathcal{N}}(\widehat{\beta}(s^{(1)}))=(\lambda^{(2)} - \lambda^{(1)})\boldsymbol{s}_{\mathcal{N}}. \quad \text{(A.10)}$$

Dividing both sides of (A.10) by $s^{(2)} - s^{(1)}$ and letting $s^{(2)} \to s^{(1)}$, we get

$$\frac{d}{ds}\boldsymbol{b}_{\mathcal{N}}(\widehat{\beta}(s))\Big|_{s^{(1)}}=\lambda'(s^{(1)})\boldsymbol{s}_{\mathcal{N}}, \quad \text{(A.11)}$$

where $\lambda'(s)=\frac{d}{ds}\lambda(s)<0$. Noting that $\frac{d}{ds}\boldsymbol{b}(\widehat{\beta}(s))=\boldsymbol{M}(\widehat{\beta}(s))\frac{d}{ds}\widehat{\beta}(s), \widehat{\beta}_{\mathcal{N}^c}(s)=\mathbf{0}$ for $s \in \mathcal{S}$ (A.11) becomes $\boldsymbol{M}_{\mathcal{N},\mathcal{N}}(\widehat{\beta}(s^{(1)}))\frac{d}{ds}\widehat{\beta}_{\mathcal{N}}(s)\Big|_{s^{(1)}}=\lambda'(s^{(1)})s_{\mathcal{N}}$, which leads to $\frac{d}{ds}\widehat{\beta}_{\mathcal{N}}(s)\Big|_{s^{(1)}}=\lambda'(s^{(1)})(\boldsymbol{M}_{\mathcal{N},\mathcal{N}}(\widehat{\beta}(s^{(1)})))^{-1}\boldsymbol{s}_{\mathcal{N}}$. Noting that $\lambda'(s) < 0$, this shows that for any $s \in \mathcal{S}$ the solution of (2.17) progresses along the CoxLARS path updating direction. It also holds for $\underline{s}$ due to continuity.

Proof of Lemma 4. Due to (A.5), $|b_j(\hat{\boldsymbol{\beta}}(s))| = |b_{j'}(\hat{\boldsymbol{\beta}}(s))|$ for any $j, j' \in \mathcal{N}$ Thus it is enough to prove that $|b_l(\hat{\boldsymbol{\beta}}(s))| \le |b_j(\hat{\boldsymbol{\beta}}(s))|$ for any $l \notin \mathcal{N}$ $j \in \mathcal{N}$ $s \in \mathcal{S} \cup \{\underline{s}\}$. We this first for $s \in \mathcal{S}$ by contradiction. Suppose there is some $j^* \notin \mathcal{N}$ and some $s^* \in \mathcal{S}$ such that

$$|b_{j^*}(\widehat{\beta}(s^*))|>|b_j(\widehat{\beta}(s^*))|. \quad \text{(A.12)}$$

Let $\boldsymbol{d} = (d_1, d_2, \ldots, d_p)^T$ with $d_j = -\text{sign}(\hat{\boldsymbol{\beta}}_j(s^*))(= \text{sign}(b_j(\hat{\boldsymbol{\beta}}(s^*))))$, due to Lemma 2) for $j \in \mathcal{N}$ $d_{j^*} = -n_{\mathcal{N}}\text{sign}(b_{j^*}(\hat{\boldsymbol{\beta}}(s^*)))$, and $d_{j'} = 0$ for $j \in (\mathcal{N} \cup \{j^*\})^c$, where $n_{\mathcal{N}}$ denote the size of $\mathcal{N}$

Consider $L(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d})$ as a function of u. Its derivative is given by

$$\frac{d}{du}L(\widehat{\beta}(s^*)+u\boldsymbol{d})=\sum_{j=1}^{p} b_j(\widehat{\beta}(s^*)+u\boldsymbol{d})d_j+O(u). \quad \text{(A.13)}$$

When $u = 0$, the right side of (A.13) becomes

$$n_{\mathcal{N}}|b_j(\widehat{\beta}(s^*))| - n_{\mathcal{N}}|b_{j^*}(\widehat{\beta}(s^*))|<0, \quad \text{(A.14)}$$

where $j \in \mathcal{N}$ and negativity is due to (A.12). Note that $\min_{j \in \mathcal{N}} |\hat{\boldsymbol{\beta}}_j(s^*)| > 0$ since $s^* \in \mathcal{S}$ When $0 < u < \min_{j \in \mathcal{N}} \left|\widehat{\beta}_j(s^*)\right|, \sum_{j=1}^{p} \left|\widehat{\beta}_j(s^*)\right| = \sum_{j=1}^{p} \left|\widehat{\beta}_j(s^*)+ud_j\right|$, noting the above definition of $\boldsymbol{d}$. However, (A.14) contradicts the fact that $\hat{\boldsymbol{\beta}}(s^*)$ is an optimal solution to (2.17). This proves our claim for $s \in \mathcal{S}$ Our claim holds at $\underline{s}$ simply due to continuity.

Proof of Lemma 5. Note that $\dot{S}(0)= -\sum_{j \in \mathcal{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j \in \mathcal{A}_0}|d_j|+\sum_{j \in \mathcal{A}_{10}^c}|d_j|$ due to Lemma 2 and $\dot{T}(0)= -\sum_{j \in \mathcal{A}_1} b_j(\widehat{\beta})d_j+\sum_{j \in \mathcal{A}_0} b_j(\widehat{\beta})d_j+\sum_{j \in \mathcal{A}_{10}^c} b_j(\widehat{\beta})d_j$. Thus, due to Lemma 4 and the above definition of $\mathcal{A}_0$, we have

$$R(\boldsymbol{d}) \triangleq -\frac{\dot{T}(0)}{\dot{S}(0)}=\widehat{D}(\widehat{\beta})\frac{-\sum_{j \in \mathcal{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j - \sum_{j \in \mathcal{A}_0}\text{sign}(b_j(\widehat{\beta}))d_j - \sum_{j \in \mathcal{A}_{10}^c} d_j b_j(\widehat{\beta})/\widehat{D}(\widehat{\beta})}{-\sum_{j \in \mathcal{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j \in \mathcal{A}_0}|d_j|+\sum_{j \in \mathcal{A}_{10}^c}|d_j|},$$

which is analogous to Equation (5.40) of Efron et al. (2004). It is enough to consider all $\boldsymbol{d}$ satisfying $-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|+\sum_{j\in\mathscr{A}_{10}^c}|d_j|>0$, which corresponds to $\dot{S}(0) > 0$. Thus we need $d_j\text{sign}(b_j(\widehat{\beta}))\le 0$ for $j\in\mathscr{A}_0\cup(\mathscr{A}_{10}^c)$ in order to maximize $R(\boldsymbol{d})$. In this case we have

$$R(\boldsymbol{d})=\widehat{D}(\widehat{\beta})\frac{-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|+\sum_{j\in\mathscr{A}_{10}^c}|d_j|\frac{|b_j(\widehat{\beta})|}{\widehat{D}(\widehat{\beta})}}{-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|+\sum_{j\in\mathscr{A}_{10}^c}|d_j|}, \quad \text{(A.15)}$$

which is $<\hat{D}(\hat{\boldsymbol{\beta}})$ unless $d_j=0$ for $j\in\mathscr{A}_{10}^c$, since $|b_j(\hat{\boldsymbol{\beta}})|<\hat{D}(\hat{\boldsymbol{\beta}})$ for $j\in\mathscr{A}_{10}^c$. This proves (3.1). In this case a second order Taylor expansion leads to (3.2).

Proof of Lemma 6. The positive definiteness of $\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}$ implies that (3.3) is equivalent to

$$\min \quad \boldsymbol{d}_{\mathscr{A}_{10}}^T\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathscr{A}_{10}} \quad \text{(A.16)}$$

$$\text{subject to } -\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j+\sum_{j\in\mathscr{A}_0}|d_j|\ge\Delta$$

$$\text{sign}(d_j)=-\text{sign}(b_j(\widehat{\beta}))\text{ for } j\in\mathscr{A}_0.$$

For (A.16), we combine the two constraints and solve the simpler version

$$\min \boldsymbol{d}_{\mathscr{A}_{10}}^T\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathscr{A}_{10}} \text{ subject to} -\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j-\sum_{j\in\mathscr{A}_0}\text{sign}(b_j(\widehat{\beta}))d_j\ge\Delta. \quad \text{(A.17)}$$

Afterward, we show that the solution to (A.17) satisfies the sign constraint in (A.16). By introducing a Lagrange multiplier for (A.17), we solve

$$\min \boldsymbol{d}_{\mathscr{A}_{10}}^T\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\widehat{\beta})\boldsymbol{d}_{\mathscr{A}_{10}}+\lambda\left(-\sum_{j\in\mathscr{A}_1}\text{sign}(b_j(\widehat{\beta}))d_j-\sum_{j\in\mathscr{A}_0}\text{sign}(b_j(\widehat{\beta}))d_j-\Delta\right). \quad \text{(A.18)}$$

Differentiating the objective function in (A.18) with respect to $\boldsymbol{d}_{\mathscr{A}_{10}}$ and solving for $\boldsymbol{d}_{\mathscr{A}_{10}}$, we get the optimal solution $(\lambda/2)(\boldsymbol{M}_{\mathscr{A}_{10},\mathscr{A}_{10}}(\hat{\boldsymbol{\beta}}))^{-1}\text{sign}(\boldsymbol{b}_{\mathscr{A}_{10}}(\hat{\boldsymbol{\beta}}))$, which is exactly the same as our CoxLARS path updating direction, noting that the Lagrange multiplier $\lambda<0$. Note that the "one at a time" condition implies that $\mathscr{A}_0$ is a singleton. Consequently, this optimal solution satisfies the sign constraint in (A.16) due to Lemma 1.

Proof of Theorem 1. Theorem 1 can be proved by induction as in Efron et al. (2004), by noting that Lemmas 2–5 are extensions of Lemmas 7–10 of Efron et al. (2004), which are the key results for establishing that the LASSO modification leads to the LASSO solutions, and parallel extensions of their Constraints 1–4 on page 437 are straightforward. We skip these details.

References

Atkinson, K.; Han, W.; Stewart, DE. Numerical Solution of Ordinary Differential Equations. Hoboken, New Jersey: John Wiley, Inc; 2009.

Breslow N. Covariance analysis of censored survival data. Biometrics. 1974; 30:89–99. [PubMed: 4813387]

Cox DR. Regression models and life-tables (with discussion). J. Roy. Statist. Soc. Ser. B. 1972; 34:187–220.

Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. Hepatology. 1989; 10:1–7. [PubMed: 2737595]

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with discussions). Ann. Statist. 2004; 32:409–499.

Fan J, Li R. Variable selection via penalized likelihood. J. Amer. Statist. Assoc. 2001; 96:1348–1360.

Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Ann. Statist. 2002; 30:74–99.

Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statist. Sinica. 2010; 20:101–148.

Friedman, J.; Hastie, T.; Tibshirani, R. Regularized paths for generalized linear models via coordinate descent Technical Report. 2008.

Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. J. Machine Learning Research. 2004; 5:1391–1415.

Park MY, Hastie T. l1-regularization path algorithm for generalized linear models. J. Roy. Statist. Soc. Ser. B. 2007; 69:659–677.

Rosset S, Zhu J. Piecewise linear regularized solution paths. Ann. Statist. 2007; 35

Therneau, TM.; Grambsch, PM. Modeling survival data: Extending the Cox model. Springer; 2001.

Tibshirani R. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B. 1996; 58:267–288.

Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16:385–95. [PubMed: 9044528]

Wu Y. An ordinary differential equation based solution path algorith. Journal of Nonparametric Statistics. 2011; 23:185–199. [PubMed: 21532936]

Yuan M, Zou H. Efficient global approximation of generalized nonlinear l1 regularized solution paths and its applications. JASA. 2009; 104:1562–1574.

Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.

Zou H. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 2006; 101:1418–1429.

Zou H. A note on path-based variable selection in the penalized proportional hazards model. Biometrika. 2008; 95:241–247.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J. Roy. Statist. Soc. Ser. B. 2005; 67:301–320.

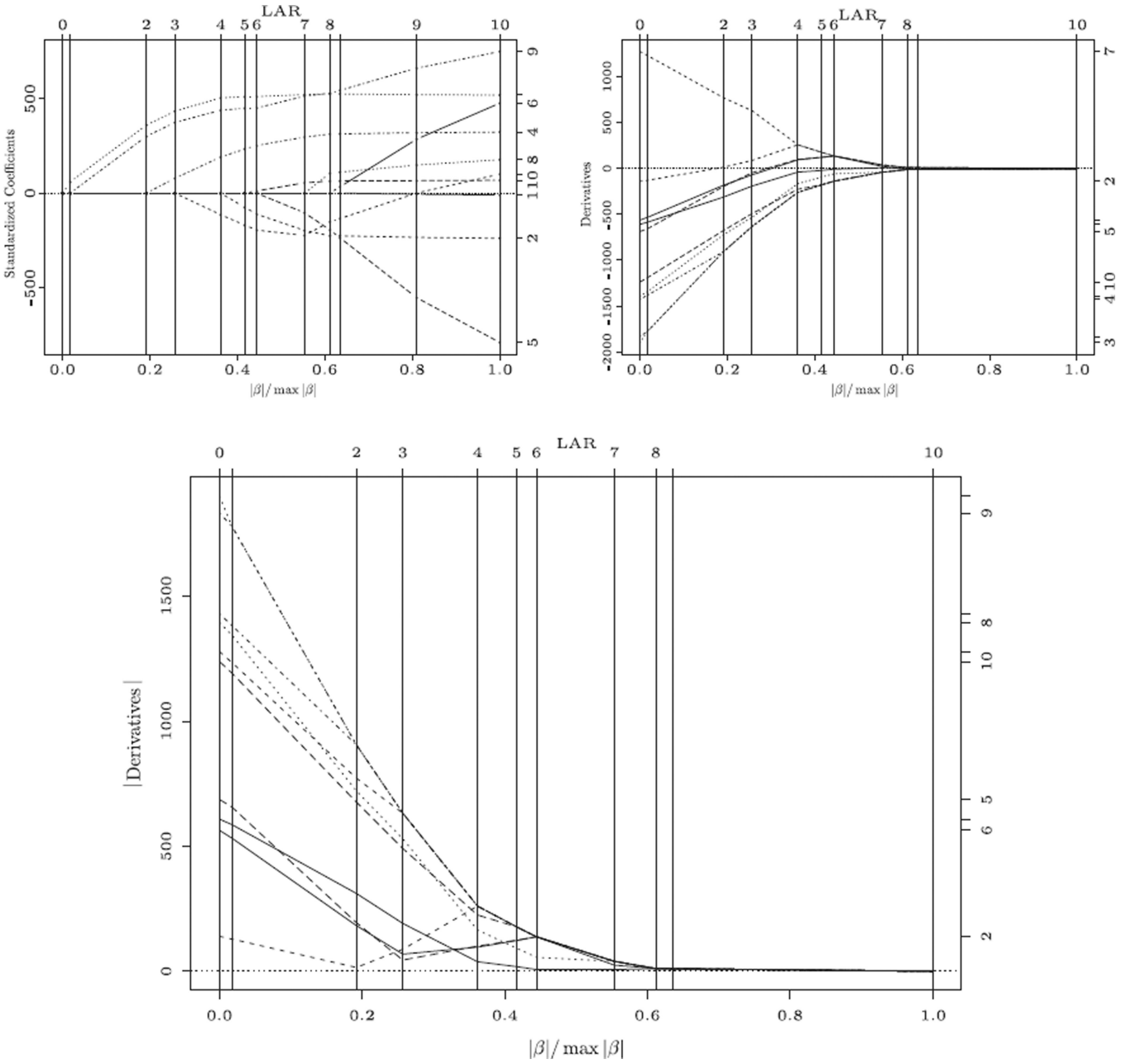

Figure 1.

LAR path of diabetes data: the top left panel plots the LAR path $w_j(t)$ against the relative one-norm $|\boldsymbol{w}(t)|/|\boldsymbol{w}(\infty)|$ for each predictor $j = 1, 2, \ldots, 10$; the top right panel and the bottom panel plot the derivative $\left(\frac{\partial}{\partial w_j}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^T\boldsymbol{w})^2\right)\Big|_{\boldsymbol{w}(t)}$ and its absolute value $\left|\left(\frac{\partial}{\partial w_j}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^T\boldsymbol{w})^2\right)\Big|_{\boldsymbol{w}(t)}\right|$, respectively, along the LAR path, for different predictors.

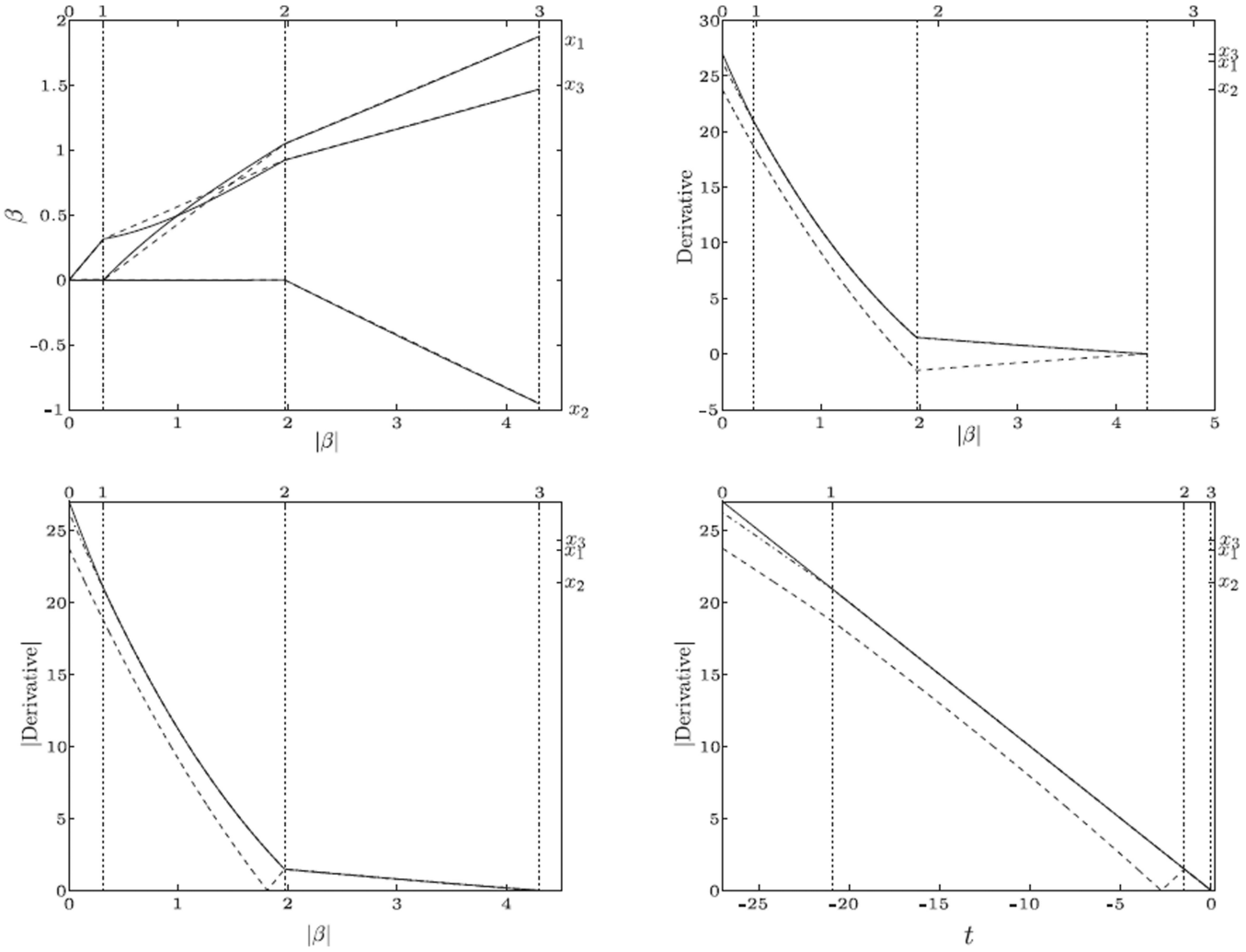

Figure 2.

CoxLASSO path of Example 1: the top left panel plots the CoxLASSO path $\boldsymbol{\beta}(t)$ with respect to the one-norm $|\boldsymbol{\beta}(t)|$; the top right panel plots the first-order derivatives $b_j(\boldsymbol{\beta}(t))$ with respect to $|\boldsymbol{\beta}(t)|$; the bottom left and right panels plot $|b_j(\boldsymbol{\beta}(t))|$ along the CoxLASSO path with respect to $|\boldsymbol{\beta}(t)|$ and t, respectively.

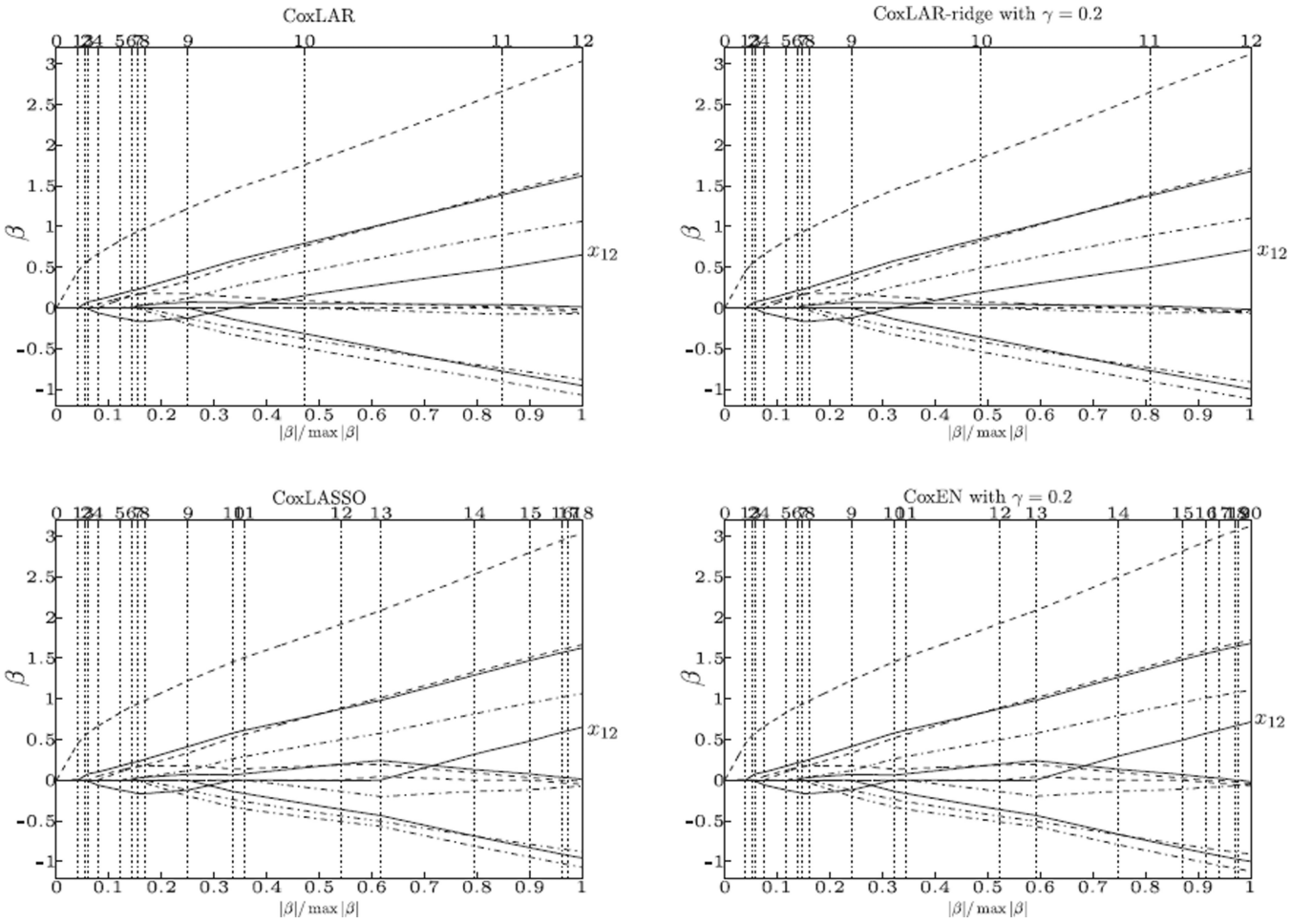

Figure 3.
CoxLARS paths of Example 2.

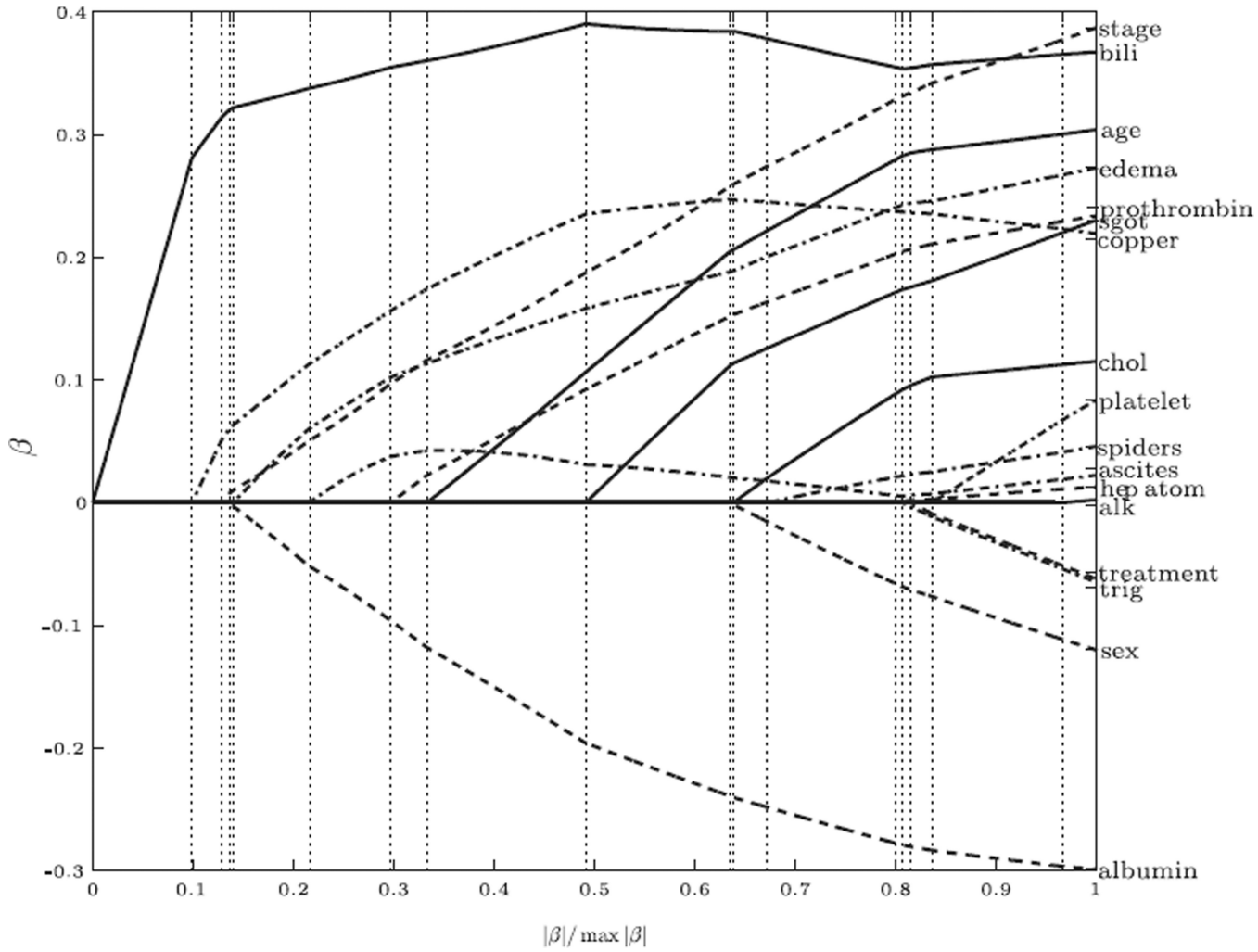

Figure 4.
CoxLARS path of the PBC data with $\gamma = 0$.