

Model selection in nonparametric hazard regression

CHENLEI LENG*† and HAO HELEN ZHANG‡

†Department of Statistics and Probability, National University of Singapore,
SG 117546, Singapore

‡Department of Statistics, North Carolina State University, USA

(Received 8 June 2006; revised 12 July 2006; in final form 10 September 2006)

We propose a novel model selection method for a nonparametric extension of the Cox proportional hazard model, in the framework of smoothing splines ANOVA models. The method automates the model building and model selection processes simultaneously by penalizing the reproducing kernel Hilbert space norms. On the basis of a reformulation of the penalized partial likelihood, we propose an efficient algorithm to compute the estimate. The solution demonstrates great flexibility and easy interpretability in modeling relative risk functions for censored data. Adaptive choice of the smoothing parameter is discussed. Both simulations and a real example suggest that our proposal is a useful tool for multivariate function estimation and model selection in survival analysis.

Keywords: COSSO; Cox proportional hazard model; Model selection; Penalized likelihood

1. Introduction

One main issue in time to event data analysis is to study the dependence of the survival time T on covariates $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$. This task is often simplified by using the Cox's proportional hazard model [1], where the log hazard function is the sum of a totally unspecified log baseline hazard function and a parameterized form of the covariates. More precisely, the Cox model can be conveniently written as

$$\log h(T|\mathbf{X}) = \log h_0(T) + \eta(\mathbf{X}) \quad \text{with } \eta(\mathbf{X}) = \mathbf{X}^T \beta,$$

where β is an unknown vector to be estimated. The parametric form of $\eta(\mathbf{X})$ is useful when the linear assumption is appropriate but may be too rigid for complicated problems. More flexible models allow $\eta(\mathbf{X})$ to vary in some infinite dimensional space [2, 3].

In many practical situations, the number of covariates d is large and not all the covariates contribute to the prediction of survival outcomes. Many variable selection techniques in linear regression models have been extended to the context of survival models such as the best subset selection and stepwise selection procedures. More recently, a number of regularization

*Corresponding author. Tel.: 65 6516-4462; Fax: 65 6872-3919; Email: stalc@nus.edu.sg

methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) [4, 5] and the SCAD [6] have been proposed. It has been shown that these regularization methods improve both prediction accuracy and stability of models. Note that all these methods are based on linear or parametric hazard models. In this article, we consider the problem of variable selection in nonparametric hazard models.

The problem of variable selection in nonparametric regression is quite challenging. Hastie and Tibshirani [7, Chapter 9.4] considered several nonlinear model selection procedures in the spirit of stepwise selection, where the familiar additive models were entertained. Gray [8] used splines with fixed degrees of freedom as an exploratory tool to assess the effect of covariates, and applied hypothesis testing procedures for model selection. Kooperberg *et al.* [9] employed a heuristic search algorithm with polynomial splines to model the hazard function. Recently, Zhang *et al.* [10] investigated a possible nonparametric extension of the LASSO via a basis pursuit method.

Smoothing spline ANOVA (SS-ANOVA) models are widely applied to estimate multivariate functions. See Wahba [11] and Gu [12] and references therein. A recent progress on nonparametric variable selection was made by Lin and Zhang [13]. They proposed the (Component Selection and Smoothing Operator COSSO) method in the SS-ANOVA models, which renders automatic model selection with a novel form of penalty. Instead of constraining squared norms as usually seen in the SS-ANOVA, a penalty on the sum of the component norms is imposed in the COSSO. As shown in Lin and Zhang [13], the COSSO penalty is a functional analogue of the L_1 constraint used in the LASSO and it is this shrinkage-type penalty that brings sparse estimated components.

Due to the nature of censored data, it is a challenging problem to conduct joint nonparametric estimation and variable selection in survival analysis, and there are very few methods for it in the literature. In this article, we show how the COSSO-type regularization can be extended to hazard regression, and establish a unified framework for smoothing and shrinkage in nonparametric proportional hazard models. The rest of the article is organized as follows. Section 2 derives the partial likelihood and reviews the SS-ANOVA models. In section 3, the new penalized partial likelihood is formulated and an efficient algorithm for computation is presented. We demonstrate the usefulness of the new method via simulations in section 4. The method is then applied to the primary biliary cirrhosis (PBC) data. Some concluding remarks are given in section 6.

2. Hazard regression

2.1 The partial likelihood

In survival data analysis, it is typical to observe censored survival times $Z_i = \min\{T_i, C_i\}$, $i = 1, \dots, n$ and the corresponding censoring indicators $\delta_i = I(T_i \leq C_i)$. Here T is the survival time and C is the censoring time. Assume that T and C are conditionally independent given $\mathbf{X} = \mathbf{x}$, and the censoring mechanism is noninformative. The data then consists of the triplets $(Z_i, \delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$. We assume that each continuous covariate is in the range of $[0, 1]$, otherwise each continuous covariate is scaled to $[0, 1]$.

For simplicity, assume that there are no ties in the observed failure times. When ties are present, we may use the technique in Breslow [14]. Let $t_1^0 < \dots < t_N^0$ be the ordered observed failure times. Using the subscript (j) to label the item failing at time t_j^0 , the covariates associated with N failures are $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(N)}$. Let R_j be the risk set right before t_j^0 :

$$R_j = \{i : Z_i \geq t_j^0\}.$$

For the family of proportional hazard models, the conditional hazard rate of an individual with covariate \mathbf{x} is

$$h(t|\mathbf{x}) = h_0(t) \exp\{\eta(\mathbf{x})\},$$

where $h_0(t)$ is an arbitrary baseline hazard function and $\eta(\mathbf{x})$ is the logarithm of the relative risk function. The log partial likelihood is then written as

$$\sum_{j=1}^N \left\{ \eta(\mathbf{x}_{(j)}) - \log \left[\sum_{i \in R_j} \exp(\eta(\mathbf{x}_i)) \right] \right\}. \tag{1}$$

See Kalbfleisch and Prentice [15] and Fan and Li [6] for more details on proportional hazard models. Parametric hazard models often assume the linear form $\eta(\mathbf{x}) = \mathbf{x}^T \beta$, which is simple and useful in practice but may be too rigid for complicated problems. In this article, we will not make any restriction on the function form of $\eta(\mathbf{x})$, *i.e.*, $\eta(\mathbf{x})$ can be any arbitrary multivariate function of \mathbf{x} .

2.2 Smoothing spline ANOVA (SS-ANOVA)

Similar to the classical ANOVA in designed experiments, a functional ANOVA decomposition of any d dimensional function $\eta(\mathbf{x})$ is

$$\eta(\mathbf{x}) = \eta_0 + \sum_{k=1}^d \eta_k(x^{(k)}) + \sum_{k < l} \eta_{k,l}(x^{(k)}, x^{(l)}) + \dots + \eta_{1,\dots,d}(x^{(1)}, \dots, x^{(d)}), \tag{2}$$

where η_0 is constant, η_k 's are main effects, and $\eta_{k,l}$'s are two-way interactions and so on. The identifiability of terms is assured by certain side conditions. We estimate the functional components of $\eta(\mathbf{x})$ in a reproducing kernel Hilbert space (RKHS) corresponding to the decomposition (2). For a thorough exposure to RKHS, see Wahba [11]. In particular, if $x^{(k)}$ is continuous, we estimate the main effect $\eta_k(x^{(k)})$ in the second-order Sobolev space $W^{(k)}[0, 1] = \{f : f(t), f'(t) \text{ are absolutely continuous, } f''(t) \in L_2[0, 1]\}$. When endowed with the following inner product:

$$\langle f, g \rangle = \int_0^1 f(t) dt \int_0^1 g(t) dt + \int_0^1 f'(t) dt \int_0^1 g'(t) dt + \int_0^1 f''(t)g''(t) dt, \tag{3}$$

$W^{(k)}[0, 1]$ is an RKHS with a reproducing kernel

$$K(s, t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s - t|).$$

Here $k_1(s) = s - 0.5$, $k_2(s) = [k_1^2(s) - 1/12]/2$, $k_4(s) = [k_1^4(s) - k_1^2(s)/2 + 7/240]/24$. This is a special case of equation (10.2.4) in Wahba [11] with $m = 2$. And this inner product introduces a set of side conditions such that each component in the decomposition (2) integrates to zero. Note the space $W^{(k)}$ can be decomposed into the direct sum of two orthogonal subspaces as $W^{(k)} = 1^{(k)} \oplus W_1^{(k)}$, where $1^{(k)}$ is the ‘mean’ space and $W_1^{(k)}$ is the ‘contrast’ space generated by the kernel $K_1(s, t) = K(s, t) - 1$. If $x^{(k)}$ is a categorical variable taking finite values $\{1, \dots, L\}$, the function $\eta_k(x^{(k)})$ is then a vector of length L and the evaluation is simply the coordinate extraction. We decompose $W^{(k)}$ as $1^{(k)} \oplus W_1^{(k)}$, where $1^{(k)} = \{f : f(1) = \dots = f(L)\}$ and $W_1^{(k)} = \{f : f(1) + \dots + f(L) = 0\}$ associated with the reproducing kernel $K_1(s, t) = LI_{(s=t)} - 1$, $s, t \in \{1, \dots, L\}$. This kernel defines a shrinkage estimate which is shrunk toward the mean, as discussed in Gu [12, Chapter 2.2].

We estimate the interaction terms in equation (2) in the tensor product spaces of the corresponding univariate function spaces. The reproducing kernel of a tensor product space is simply the product of the reproducing kernels of individual spaces. For example, the reproducing kernel of $W_1^{(k)} \otimes W_1^{(l)}$ is $K_1(s^{(k)}, t^{(k)})K_1(s^{(l)}, t^{(l)})$. This structure greatly facilitates the use of SS-type methods for high-dimensional data. Corresponding to equation (2), the full metric space for estimating $\eta(\mathbf{x})$ is the tensor product space

$$\bigotimes_{k=1}^d W_1^{(k)} = \{1\} \bigoplus_{k=1}^d W_1^{(k)} \bigoplus_{k < l} \{W_1^{(k)} \otimes W_1^{(l)}\} \bigoplus \dots$$

High-order terms in the decomposition (2) are often excluded to control the model complexity. For example, excluding all the interactions yields the additive models [7], $\eta(\mathbf{x}) = \eta_0 + \sum_{k=1}^d \eta_k(x^{(k)})$. Including all the two-way interactions and main effect terms leads to the two-way interaction models

$$\eta(\mathbf{x}) = \eta_0 + \sum_{k=1}^d \eta_k(x^{(k)}) + \sum_{k < l} \eta_{k,l}(x^{(k)}, x^{(l)}).$$

The family of low-dimensional ANOVA decompositions represents a nonparametric compromise in an attempt to overcome the ‘curse of dimensionality’, since estimating a general multivariate function $\eta(x^{(1)}, \dots, x^{(d)})$ requires a large sample size even for a moderate d . See relevant discussions in Huang *et al.* [16]. In general, the truncated series of equation (2) is written as

$$\eta(\mathbf{x}) = \eta_0 + \sum_{\alpha=1}^p \eta_\alpha(\mathbf{x}), \quad (4)$$

and lies in a direct sum of p orthogonal subspaces

$$\mathcal{H} = \{1\} \bigoplus_{\alpha=1}^p \mathcal{H}_\alpha.$$

With some abuse of notation, we use $K_\alpha(s, t)$ to denote the reproducing kernel for \mathcal{H}_α . Consequently, the reproducing kernel of \mathcal{H} is given by $1 + \sum_{\alpha=1}^p K_\alpha$.

3. Model formulation

3.1 Penalized partial likelihood

One popular approach to the nonparametric estimation of $\eta(\mathbf{x})$ is via the minimization of a penalized partial likelihood [3,12]

$$\min_{\eta \in \mathcal{H}} -\frac{1}{n} \sum_{j=1}^N \left\{ \eta(\mathbf{x}_{(j)}) - \log \left[\sum_{i \in R_j} \exp(\eta(\mathbf{x}_i)) \right] \right\} + \tau J(\eta), \quad (5)$$

where $J(\eta) = \sum_{\alpha=1}^p \theta_\alpha^{-1} \|P^\alpha \eta\|^2$ is a roughness penalty and $P^\alpha \eta$ is the projection of η onto \mathcal{H}_α . The θ 's are smoothing parameters which control the goodness of fit and the solution roughness. For high-dimensional data, fitting a model with multiple tuning parameters can be computationally intensive. Gu and Wahba [17] proposed an algorithm to choose optimal

parameters via the multiple dimensional minimization. However, their algorithm operates on τ and $\log(\theta_\alpha)$'s, therefore none of the component estimates of η is exactly zero. Some ad hoc variable selection techniques, say, geometric diagnostics techniques [18], need to be applied after model fitting.

In order to combine model fitting and automatic model selection in one unified framework, we propose a new penalized partial likelihood score

$$-\frac{1}{n} \sum_{j=1}^N \left\{ \eta(\mathbf{x}_{(j)}) - \log \left[\sum_{i \in R_j} \exp(\eta(\mathbf{x}_i)) \right] \right\} + \tau \sum_{\alpha=1}^p \|P^\alpha \eta\|. \tag{6}$$

Different from the SSs, the penalty functional $\sum_{\alpha=1}^p \|P^\alpha \eta\|$ is a sum of RKHS component norms instead of the squared RKHS norm. This penalty was first suggested by Lin and Zhang [13] in the ordinary regression models for Gaussian data and named as ‘COSSO’. Note there is a single tuning parameter τ in equation (6), which is advantageous compared with multiple tuning parameters in the SS. The fundamental difference between the COSSO and the SS mirrors that between the LASSO and the ridge regression. The LASSO tends to shrink coefficients to be exactly zeros, and the ridge regression shrinks them but hardly produces zeros. Similarly, the COSSO penalty can produce sparse components in the solution while the standard SS does not in general.

In the special case of linear models, $\eta(\mathbf{x}) = \beta_0 + \sum_{k=1}^d \beta_k x^{(k)}$ and the model space \mathcal{H} is $\{1\} \oplus \{x^{(1)} - 1/2\} \oplus \dots \oplus \{x^{(d)} - 1/2\}$ equipped with the L^2 inner product. The COSSO penalty $\sum_{\alpha=1}^p \|P^\alpha \eta\|$ becomes $(12)^{-1/2} \sum_{k=1}^d |\beta_k|$, equivalent to the L_1 penalty on linear coefficients. Therefore, the LASSO method proposed by Tibshirani [4] for variable selection in the Cox model can be seen as a special case of our penalized likelihood estimate (6).

3.2 Equivalent formulation

Though the minimizer of equation (6) is searched over the infinite dimensional RKHS \mathcal{H} , in the next lemma, we show that the solution $\hat{\eta}$ always lies in a finite dimensional subspace of \mathcal{H} .

LEMMA 1 Denote $\hat{\eta} = \hat{b} + \sum_{\alpha=1}^p \hat{\eta}_\alpha$ as the minimizer of equation (6) with $\hat{\eta}_\alpha \in \mathcal{H}_\alpha$. Then $\hat{\eta}_\alpha \in \text{span}\{K_\alpha(\mathbf{x}_i, \cdot), i = 1, \dots, n\}$, where $K_\alpha(\cdot, \cdot)$ is the reproducing kernel of \mathcal{H}_α .

A proof can be found in Lin and Zhang [13].

To solve equation (6), we consider its equivalent formulation following Lin and Zhang [13]. It is easy to show that minimizing equation (6) is equivalent to solving

$$\begin{aligned} \min_{\eta, \boldsymbol{\theta}} & -\frac{1}{n} \sum_{j=1}^N \left\{ \eta(\mathbf{x}_{(j)}) - \log \left[\sum_{i \in R_j} \exp(\eta(\mathbf{x}_i)) \right] \right\} + \lambda_0 \sum_{\alpha=1}^p \theta_\alpha^{-1} \|P^\alpha \eta\|^2 \\ & \text{subject to } \sum_{\alpha=1}^p \theta_\alpha \leq M, \quad \theta_\alpha \geq 0, \quad \alpha = 1, \dots, p, \end{aligned} \tag{7}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ are introduced as non-negative slack variables. In equation (7), λ_0 is a fixed parameter and M is the smoothing parameter. There is one-to-one corresponding relationship between M in equation (7) and τ in equation (6). When $\boldsymbol{\theta}$ is fixed, this formulation has the same form as the usual SS-ANOVA except that the sum of θ_α 's is penalized. We remark that the additional penalty on $\boldsymbol{\theta}$ makes it possible to shrink some θ_α 's to zeros, leading to zero components in the function estimate.

3.3 Form of solutions

For any fixed θ , the problem (7) is equivalent to the SS. By the representer theorem, the solution has the form $\eta(\mathbf{x}) = b + \sum_{i=1}^n K_\theta(\mathbf{x}, \mathbf{x}_i)c_i$, where $K_\theta = \sum_{\alpha=1}^p \theta_\alpha K_\alpha$. For the identifiability of η , we absorb b into the baseline hazard function, or equivalently, set $b = 0$ in the following discussion. Therefore the exact solution to equation (7) has the form

$$\eta(\mathbf{x}) = \sum_{i=1}^n \sum_{\alpha=1}^p \theta_\alpha K_\alpha(\mathbf{x}, \mathbf{x}_i)c_i.$$

For large data sets, we can reduce the computational load of optimizing equation (7) via parsimonious approaches [19]. The idea is to minimize the objective function in a subspace of \mathcal{H} spanned by a subset $\{\mathbf{x}_1^*, \dots, \mathbf{x}_m^*\}$ of $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ ($m < n$). In the standard SS setting, Kim and Gu [20] showed that, there is little sacrifice in the solution accuracy even when m is small. The approximate solution in the subspace is then $\eta(\mathbf{x}) = \sum_{i=1}^m \sum_{\alpha=1}^p \theta_\alpha K_\alpha(\mathbf{x}, \mathbf{x}_i^*)c_i$. In our numerical examples, we use a random sampling scheme to choose the subset.

3.4 Alternating optimization algorithm

It is possible to minimize equation (7) with respect to θ 's and c 's simultaneously. However, we propose a simpler algorithm which alternatively minimizes the objective function with respect to one set of θ 's or c 's while keeping the other set fixed. We refer to this algorithm as the alternating optimization algorithm. Here M is fixed and we will discuss its selection in the next section.

Denote the objective function in equation (7) as $A(\mathbf{c}, \theta)$, where $\mathbf{c} = (c_1, \dots, c_m)^T$ and $m \leq n$. When $m = n$, all the samples are used to generate basis functions. Let Q be an $m \times m$ matrix with (k, l) entry being $K_\theta(\mathbf{x}_k^*, \mathbf{x}_l^*)$ and Q_α an $m \times m$ matrix with (k, l) entry $K_\alpha(\mathbf{x}_k^*, \mathbf{x}_l^*)$. Let U be an $n \times m$ matrix with (k, l) entry being $K_\theta(\mathbf{x}_k, \mathbf{x}_l^*)$ and U_α an $n \times m$ matrix with (k, l) entry $K_\alpha(\mathbf{x}_k, \mathbf{x}_l^*)$. Straightforward calculations show that $(\eta(\mathbf{x}_1), \dots, \eta(\mathbf{x}_n))^T = U\mathbf{c}$ and $\|P^\alpha \eta\|^2 = \theta_\alpha^2 \mathbf{c}' Q_\alpha \mathbf{c}$. Denoting $\delta = (\delta_1, \dots, \delta_n)^T$ as the vector of censoring indicators, we can write equation (7) in the following matrix form

$$A(\mathbf{c}, \theta) = -\frac{1}{n} \delta^T U \mathbf{c} + \frac{1}{n} \sum_{j=1}^N \log \left(\sum_{i \in R_j} e^{U_i \mathbf{c}} \right) + \lambda_0 \mathbf{c}' Q \mathbf{c}, \quad \text{s.t. } \sum_{\alpha=1}^p \theta_\alpha \leq M, \quad \theta_\alpha \geq 0, \quad (8)$$

where U_i is the i th row of U . The alternative optimization algorithm consists of two parts.

(1) When θ is fixed, the gradient vector and Hessian matrix of A with respect to \mathbf{c} are

$$\begin{aligned} \frac{\partial A}{\partial \mathbf{c}} &= -\frac{1}{n} U^T \delta + \frac{1}{n} \sum_{j=1}^N \frac{\sum_{i \in R_j} U_i^T e^{U_i \mathbf{c}}}{\sum_{i \in R_j} e^{U_i \mathbf{c}}} + 2\lambda_0 Q \mathbf{c}, \\ \frac{\partial^2 A}{\partial \mathbf{c} \partial \mathbf{c}^T} &= \frac{1}{n} \sum_{j=1}^N \left[\frac{\sum_{i \in R_j} U_i^T U_i e^{U_i \mathbf{c}}}{\sum_{i \in R_j} e^{U_i \mathbf{c}}} - \frac{\sum_{i \in R_j} U_i^T e^{U_i \mathbf{c}}}{\sum_{i \in R_j} e^{U_i \mathbf{c}}} \frac{\sum_{i \in R_j} U_i e^{U_i \mathbf{c}}}{\sum_{i \in R_j} e^{U_i \mathbf{c}}} \right] + 2\lambda_0 Q. \end{aligned} \quad (9)$$

The Newton–Raphson iteration is used to update \mathbf{c} as

$$\mathbf{c} = \mathbf{c}_0 - \left(\frac{\partial^2 A}{\partial \mathbf{c} \partial \mathbf{c}^T} \right)_{\mathbf{c}_0}^{-1} \left(\frac{\partial A}{\partial \mathbf{c}} \right)_{\mathbf{c}_0}, \quad (10)$$

where \mathbf{c}_0 is the current estimate of the coefficient vector and the Hessian and gradient are evaluated at \mathbf{c}_0 .

- (2) When \mathbf{c} is fixed, we denote G as an $m \times p$ matrix with the α th column being $Q_\alpha \mathbf{c}$ and S as an $n \times p$ matrix with the α th column being $U_\alpha \mathbf{c}$. The objective function in equation (7) can be written as a function of $\boldsymbol{\theta}$

$$A(\mathbf{c}, \boldsymbol{\theta}) = -\frac{1}{n} \boldsymbol{\delta}^T S \boldsymbol{\theta} + \frac{1}{n} \sum_{j=1}^N \log \left(\sum_{i \in R_j} e^{\mathbf{S}_i \boldsymbol{\theta}} \right) + \lambda_0 \mathbf{c}^T G \boldsymbol{\theta}, \quad \text{s.t. } \sum_{\alpha=1}^p \theta_\alpha \leq M, \quad \theta_\alpha \geq 0, \tag{11}$$

where \mathbf{S}_i is the i th row of S . We further expand $A(\mathbf{c}, \boldsymbol{\theta})$ around the current estimate $\boldsymbol{\theta}_0$ via the second-order Taylor expansion

$$A(\mathbf{c}, \boldsymbol{\theta}) \approx A(\mathbf{c}, \boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\frac{\partial A}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}_0} + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\frac{\partial^2 A}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)_{\boldsymbol{\theta}_0} (\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where

$$\begin{aligned} \frac{\partial A}{\partial \boldsymbol{\theta}} &= -\frac{1}{n} S^T \boldsymbol{\delta} + \frac{1}{n} \sum_{j=1}^N \frac{\sum_{i \in R_j} \mathbf{S}_i^T e^{\mathbf{S}_i \boldsymbol{\theta}}}{\sum_{i \in R_j} e^{\mathbf{S}_i \boldsymbol{\theta}}} + \lambda_0 G^T \mathbf{c}, \\ \frac{\partial^2 A}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \frac{1}{n} \sum_{j=1}^N \left[\frac{\sum_{i \in R_j} \mathbf{S}_i^T \mathbf{S}_i e^{\mathbf{S}_i \boldsymbol{\theta}}}{\sum_{i \in R_j} e^{\mathbf{S}_i \boldsymbol{\theta}}} - \frac{\sum_{i \in R_j} \mathbf{S}_i^T e^{\mathbf{S}_i \boldsymbol{\theta}}}{\sum_{i \in R_j} e^{\mathbf{S}_i \boldsymbol{\theta}}} \frac{\sum_{i \in R_j} \mathbf{S}_i e^{\mathbf{S}_i \boldsymbol{\theta}}}{\sum_{i \in R_j} e^{\mathbf{S}_i \boldsymbol{\theta}}} \right]. \end{aligned} \tag{12}$$

The iteration for updating $\boldsymbol{\theta}$ is via the minimization of the following linearly constrained quadratic objective function:

$$\frac{1}{2} \boldsymbol{\theta}^T \left(\frac{\partial^2 A}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)_{\boldsymbol{\theta}_0} \boldsymbol{\theta} + \left[\left(\frac{\partial A}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}_0} - \left(\frac{\partial^2 A}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)_{\boldsymbol{\theta}_0} \boldsymbol{\theta}_0 \right]^T \boldsymbol{\theta}, \quad \text{s.t. } \sum_{\alpha=1}^p \theta_\alpha \leq M, \quad \theta_\alpha \geq 0. \tag{13}$$

The linear constraint on $\sum_{\alpha=1}^p \theta_\alpha$ makes it possible to have sparse solutions in $\boldsymbol{\theta}$.

For any fixed M , the optimization algorithm iterates between updating \mathbf{c} and $\boldsymbol{\theta}$. Similar to Fan and Li [6], we found that the one-step iteration can produce good solutions very close to the one given by the fully iterative algorithm, provided a good initial estimate $\hat{\eta}_0$. We suggest using the SS estimate as a starting point and the one-step procedure.

3.5 Smoothing parameter selection

The problem of smoothing parameter selection for nonparametric hazard regression is important. On the basis of a Kullback–Leibler distance for hazard estimation, Gu [12, Chapter 7.2] derived a cross-validation score to tune smoothing parameters:

$$PL(M) + \left\{ \frac{\text{tr}(\Delta U^T H^{-1} U \Delta)}{n(n-1)} - \frac{\boldsymbol{\delta}^T U^T H^{-1} U \boldsymbol{\delta}}{n^2(n-1)} \right\},$$

where $\Delta = \text{diag}(\delta_1, \dots, \delta_n)$ and $PL(M)$ stands for the fitted minus log partial likelihood. We propose a simple modification of Gu’s cross-validation criterion, called the approximate

cross-validation (ACV), to choose M ,

$$\text{ACV}(M) = \text{PL}(M) + \frac{N}{n} \left\{ \frac{\text{tr}(U^T H^{-1} U)}{n(n-1)} - \frac{\mathbf{1}^T U^T H^{-1} U \mathbf{1}}{n^2(n-1)} \right\}.$$

The proposed ACV score naturally takes into account the censoring factor; it is convenient to compute since no extra effort is needed once the minimizer of equation (7) is obtained. Combining the one-step update fitting procedure and parameter tuning, we have the following complete algorithm:

- (1) Fix $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (1, \dots, 1)^T$, tune λ_0 according to ACV and fix it from now on.
- (2) For each M in a reasonable range, solve $\hat{\boldsymbol{\eta}}$ with the alternating optimization scheme:
 - (a) with $\boldsymbol{\theta}$ fixed at current values, use Newton–Raphson iteration (10) to update \mathbf{c} ;
 - (b) with \mathbf{c} fixed at current values, solve equation (13) for $\boldsymbol{\theta}$. Denote the solution as $\boldsymbol{\theta}_M$;
 - (c) with $\boldsymbol{\theta}_M$ fixed, solve equation (10) again for \mathbf{c} and denote the solution as \mathbf{c}_M .
- (3) Choose the optimal \hat{M} which minimizes the ACV score. The corresponding solution is our final solution.
- (4) Compute the function estimate as $\hat{\boldsymbol{\eta}} = K_{\boldsymbol{\theta}_{\hat{M}}} \mathbf{c}_{\hat{M}}$.

Extensive simulations show that the number of nonzero components appearing in the final model is close to M . This correspondence greatly facilitates the specification of a reasonable range for M .

4. Simulation examples

We simulate the following proportional hazard models.

Example 1 (Nonlinear additive models): In this example, we first generate eight-dimensional covariates $\mathbf{X} = (X^{(1)}, \dots, X^{(8)})$ such that $X^{(j)} \sim N(0, 1)$ marginally, and the pairwise correlation between $X^{(i)}$ and $X^{(j)}$ is $\rho^{|i-j|}$ with $\rho = 0.5$. Each covariate is then truncated into $[-2, 2]$ and scaled to $[0, 1]$. The baseline hazard is set as $h_0(t) = 1$. The true relative risk function depends on only three covariates $X^{(1)}, X^{(4)}, X^{(7)}$ and takes the form

$$\eta(\mathbf{x}) = \eta_1(x^{(1)}) + \eta_4(x^{(4)}) + \eta_7(x^{(7)}),$$

where

$$\eta_1(t) = 3(3t - 2)^2, \quad \eta_4(t) = 4 \cos((3t - 1.5)\pi/5), \quad \eta_7(t) = I(t < 0.5).$$

The functions η_1 and η_2 were also used in Fan *et al.* [21], although in a slightly different context. To check the performance of variable selection on categorical variables, we further transform $X^{(8)}$ to $I(X^{(8)} > 0.6)$. The censoring time C is generated to follow an exponential distribution with mean $V \exp(-\eta(\mathbf{x}))$, where V is randomly generated from the $\text{Unif}[a, a + 2]$. The parameter a is chosen such that about 45%, 30%, 15% of the observations are censored. They are respectively referred to as ‘high’, ‘moderate’ and ‘low’ censoring rates. The censoring scheme is noninformative since $\eta(\mathbf{x})$ is a known function.

We consider two scenarios: mutually independent covariates ($\rho = 0$) and moderately correlated covariates ($\rho = 0.5$), and two sample sizes: $n = 100$ and $n = 200$. In each setting, 50 observations are randomly chosen to generate the basis functions. In order to measure the model selection performance of our method, we report the number of variables selected correctly by the model (denoted as ‘Correct’) and the number of variables selected incorrectly

Table 1. Simulation results for the nonparametric additive model of Example 1.

n	Censoring	$\rho = 0$			$\rho = 0.5$		
		Correct	Incorrect	ISE	Correct	Incorrect	ISE
100	High	4.22	0.41	1.087 (0.06)	4.07	0.56	1.246 (0.07)
	Moderate	4.13	0.27	0.868 (0.04)	4.10	0.34	0.902 (0.05)
	Low	4.26	0.18	0.616 (0.03)	4.38	0.30	0.692 (0.03)
200	High	4.26	0.02	0.406 (0.02)	4.26	0.06	0.433 (0.02)
	Moderate	4.32	0.00	0.332 (0.02)	4.21	0.01	0.353 (0.01)
	Low	4.33	0.01	0.269 (0.01)	4.30	0.00	0.291 (0.01)

by the model (denoted as ‘Incorrect’). To measure the model estimation performance of our method, we compute the integrated square error $ISE(\hat{\eta}) = E_{\mathbf{X}}\{\sum_{\alpha=1}^p(\eta_{\alpha} - \hat{\eta}_{\alpha})^2\}$, which is estimated by a Monte Carlo integration using 2000 test point.

Each example is repeated 100 times and table 1 summarizes the average ‘Correct’, ‘Incorrect’, ISE and its associated standard error (in parentheses). When all the variables are independent, the true ‘Correct’ number of zeros is 5 and ‘Incorrect’ number of zeros is 3. Our method gives overall good performance in both variable selection and model estimation. As the sample size increases or the censoring rate decreases, the ISE gets much better and the model identifies important variables more correctly (with smaller ‘Incorrect’ scores).

We measure the magnitude of each function component by its empirical L_1 norm defined as $1/n \sum_{i=1}^n |\eta_{\alpha}(x_i^{(\alpha)})|$ for $\alpha = 1, \dots, d$. Figure 1 shows the ACV curves and how the empirical L_1 norms of the estimated components change with the tuning parameter M in two typical runs: the top row for $n = 100, \rho = 0.5$, and high censoring rate; the bottom row for $n = 200, \rho = 0$,

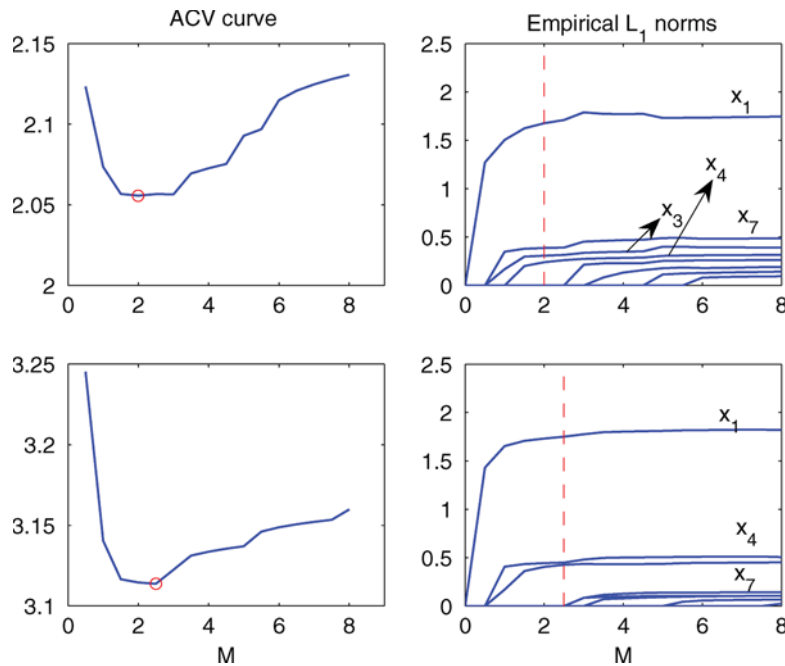


Figure 1. The ACV curve and the empirical L_1 norms of the estimated components against the tuning parameter M . They correspond to the 50th best in terms of ISE. Top row: $n = 100, \rho = 0.5$, and censoring is high. Bottom row: $n = 200, \rho = 0$, and censoring is low. Red circles and red dashed lines (grey) indicate the chosen tuning parameter.

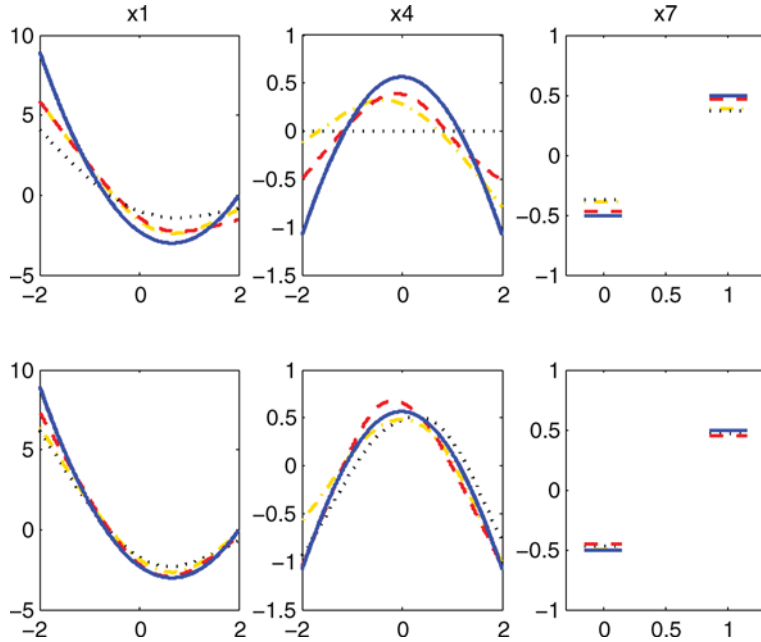


Figure 2. Top row: the estimated function components when $n = 100$, $\rho = 0.5$, and censoring is high. Bottom row: The estimated function components when $n = 200$, $\rho = 0$, and censoring is low. Red dashed lines (dark grey) indicate the fifth best; yellow dashed-dotted lines (light grey) indicate the 50th best. Blue solid lines are the true functional components.

and low censoring rate. The ACV criterion chooses $\hat{M} = 2$ and $\hat{M} = 2.5$, respectively, resulting in a model with four nonzero components (in the top row) and a model with three correctly identified components (in the bottom row).

Figure 2 plots the true functional components and their estimates for two scenarios. The 5th, 50th, 95th best estimates over 100 runs are ranked according to their ISE values. We can see that the proposed method provides very good estimates for those important functional components. Overall, our method performs very well in identifying the true model and estimating the true functional components.

Example 2 (Two-way interaction model): We generate four-dimensional covariates $X^{(j)} \sim N(0, 3^2)$ truncated at $[-3, 3]$. The pairwise correlation is ρ and the baseline hazard is $h_0(t) = 3t^2$. The censoring random variable C is uniformly distributed on $[0, 3]$ such that about 30–40% data are censored. The true hazard function is

$$\eta(\mathbf{x}) = \eta_1(x^{(1)}) + \eta_2(x^{(2)}) + \eta_{1,2}(x^{(1)}, x^{(2)}),$$

where

$$\eta_1(t) = \frac{e^{3.5t}}{(1 + e^{3.5t})}, \quad \eta_2(t) = \frac{t}{3}, \quad \eta_{1,2}(t, s) = -\frac{ts}{3}.$$

Table 2. Model selection results for the interaction model in Example 2.

n	$\rho = 0$			$\rho = 0.5$		
	Correct	Incorrect	ISE	Correct	Incorrect	ISE
100	5.41	0.04	0.233 (0.01)	5.47	0.07	0.233 (0.01)
200	5.59	0.01	0.112 (0.01)	5.59	0.01	0.115 (0.01)

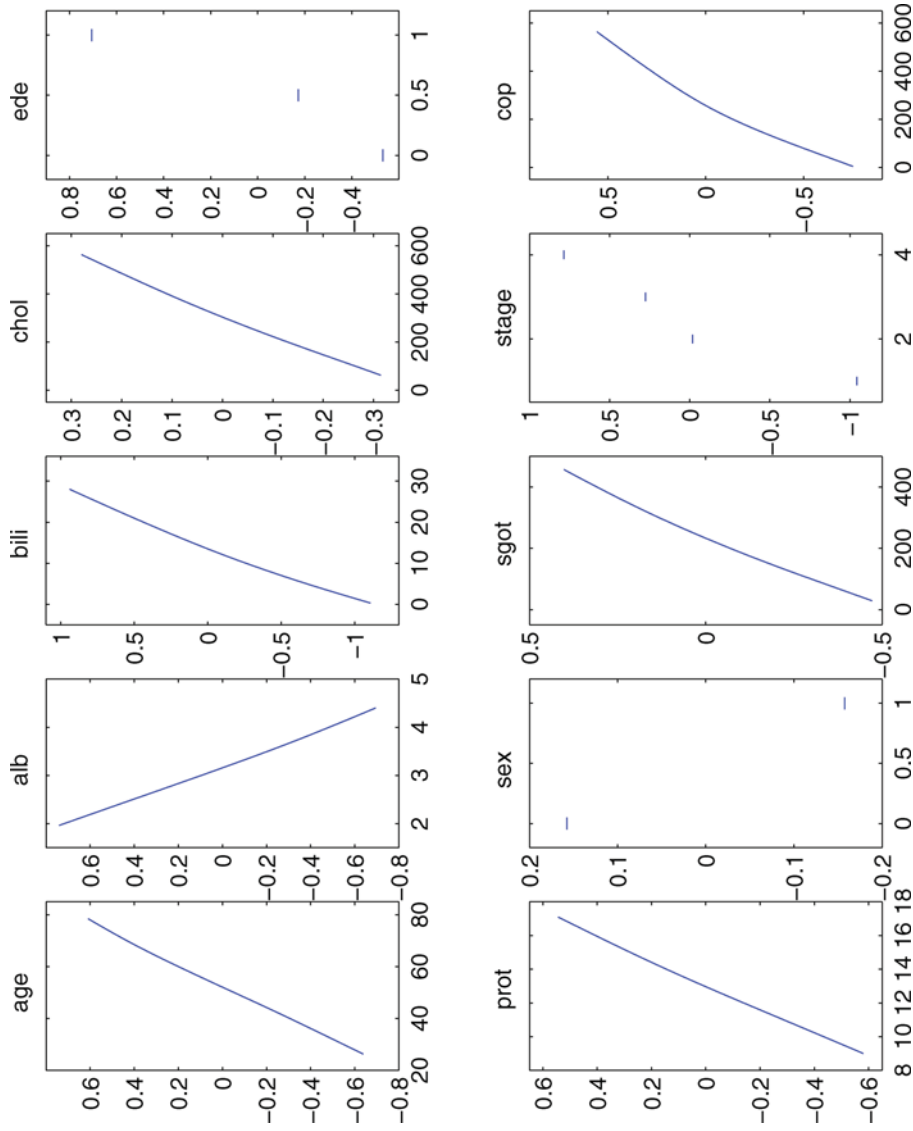


Figure 3. Fitted main effects for PBC data.

We fit the two-way interaction SS-ANOVA model with the COSSO penalty, which contains three important components and seven zero components. The experiment is repeated 100 times for $n = 100$ and $n = 200$, and the results are summarized in table 2. We can see that the proposed method gives reasonable results in terms of both variable selection and model error.

5. Real data example

The PBC data was gathered from the Mayo Clinic trial in PBC of liver conducted between 1974 and 1984. This data is provided in Therneau and Grambsch [22]. In this study, 312 patients from a total of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. For each patient, clinical, biochemical, serologic, and histological parameters are collected. Of those, 125 patients died before the end of follow-up. We study the dependence of the survival time on 17 covariates as detailed in Fleming and Harrington [23].

We restrict our attention to the 276 observations without missing values in the covariates. As reported in Tibshirani [5], the stepwise selection chooses eight variables: age, ede, bili, alb, cop, sgot, prot, and stage. The LASSO procedure selects three more variables, sex, asc, and spid. Compared with the stepwise selection, our procedure selects two more variables sex and chol. Quite interestingly, the stepwise model selects only those covariates with absolute Z -scores larger than 2.00, and our model selects only those covariates with absolute Z -scores larger than 1.00, where Z -scores refer to the scores obtained in the full parametric Cox proportional hazard model. The LASSO, instead, selects two covariates asc (Z -score 0.23) and spid (Z -score 0.42) with Z -scores less than 1 while leaving chol (Z -score 1.11) out of the model. The fitted effects of our model are shown in figure 3. The model fit suggests a nonlinear trend in cop, which is interesting and worth further investigation.

6. Discussion

We generalized the regularization with the COSSO penalty to the nonparametric Cox's proportional hazard models. An efficient criterion is proposed to select the smoothing parameter. The new procedure conducts model selection and function estimation simultaneously for the time-to-event data. Our simulations and the real example suggest the great potential of this method for identifying important risk factors and estimating the components in nonparametric hazard regression. How to extend our work to nonproportional hazard models is a future direction of research.

Acknowledgement

Leng's research was supported in part by NSF Grant DMS 0072292 and NIH Grant EY09946. Zhang's research is supported partially by NSF Grant 0405913. The authors would like to thank Grace Wahba and Yi Lin for helpful discussions.

References

- [1] Cox, D.R., 1972, Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.
- [2] O'Sullivan, F., 1993, Nonparametric estimation in the Cox model. *The Annals of Statistics*, **21**, 124–145.
- [3] Gu, C., 1996, Penalized likelihood hazard estimation: a general procedure. *Statistica Sinica*, **6**, 861–876.
- [4] Tibshirani, R., 1996, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

- [5] Tibshirani, R., 1997, The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.
- [6] Fan, J. and Li, R., 2002, Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74–99.
- [7] Hastie, T. and Tibshirani, R., 1990, *Generalized Additive Models* (London: Chapman & Hall Ltd.).
- [8] Gray, R.J., 1992, Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–951.
- [9] Kooperberg, C., Stone, C.J. and Truong, Y.K., 1995, Hazard regression. *Journal of the American Statistical Association*, **90**, 78–94.
- [10] Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B., 2004, Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, **99**, 659–672.
- [11] Wahba, G., 1990, Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59 (Philadelphia: SIAM).
- [12] Gu, C., 2002, *Smoothing Spline ANOVA Models* (New York: Springer-Verlag).
- [13] Lin, Y. and Zhang, H., 2006, Component selection and smoothing in smoothing spline analysis of variance model. *The Annals of Statistics*, **34**, in press.
- [14] Breslow, N., 1974, Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.
- [15] Kalbfleisch, J.D. and Prentice, R.L., 2002, *The Statistical Analysis of Failure Time Data* (New York: John Wiley and Sons), p. 321.
- [16] Huang, J.Z., Kooperberg, C., Stone, C.J. and Truong, Y.K., 2000, Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics*, **28**, 961–999.
- [17] Gu, C. and Wahba, G., 1991, Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, **12**, 383–398.
- [18] Gu, C., 1992, Diagnostics for nonparametric regression models with additive term. *Journal of the American Statistical Association*, **87**, 1051–1058.
- [19] Xiang, D. and Wahba, G., 1996, A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, **6**, 675–692.
- [20] Kim, Y.-J. and Chong, G., 2004, Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B*, **66**(2), 337–356.
- [21] Fan, J., Lin, H. and Zhou, Y., 2006, Local partial likelihood estimation for life time data. *The Annals of Statistics*, **34**, 290–325.
- [22] Therneau, T. M. and Grambsch, P. M., 2000, *Modeling Survival Data: Extending the Cox Model* (New York: Springer-Verlag Inc.).
- [23] Fleming, T.R. and Harrington, D.P., 1991, *Counting Processes and Survival Analysis* (New York: John Wiley and Sons).