Nonparametric Model Selection for Cox's Proportional Hazard Model

Chenlei Leng

August 27, 2003

Abstract

We consider Cox proportional hazard models using smoothing splines. By imposing a novel penalty instead of the usual square penalty in smoothing spline ANOVA cases, our method shrinks functional components and produces some components that are exactly zeros. It is a variant of the 'COSSO' proposal and our method extends it to nonparametric Cox proportional hazard models. An efficient algorithm based on a reformulation of the penalized likelihood is given. To choose the smoothing parameter, an approximation to the leave-out-one likelihood cross validation score criterion is derived. Both simulations and real examples suggest our proposal is very powerful for model selection and component estimation in survival analysis.

Keyword: Cox proportional hazard model, penalized likelihood, Smoothing spline ANOVA models, model selection, COSSO, Kullback-Leibler distance, cross validation.

1 Introduction

An objective of time to event data analysis is to identify possible risk factors and quantify their contributions. At the beginning of the analysis, a large number of covariates are introduced to reduce modeling bias. A critical task then is to select a subset of significant variables which contribute to the survival time.

Cox proportional hazard models are among the most popular techniques in modeling survival data using partial likelihood since its introduction by Cox (1972). For a thorough exposure to this topic, see for example Kalbfleisch and Prentice (1980), Klein and Moeschberger (1997), among others. The multiplicative hazards model assumes a semiparametric model where the baseline hazard is totally unspecified and a parametric form is hypothesized only for the covariates effects. The baseline hazard is treated nonparametrically. To explore the functional form of the covariates, a number of nonparametric techniques arise as a powerful alternative to the parametric fitting. General treatments of the nonlinear forms can be found in Wahba (1990), Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijbels (1996), among others. Penalized partial likelihood was studied by O'Sullivan (1988),Gu (1996). An upper bound of rate of convergence is given by O'Sullivan (1993). Zucker and Karr (1990) considered a generalization of the proportional hazard model which allows time varying coefficients. Gray (1992) and Gray (1994) used splines with fixed degree of freedom as the exploratory tool to assess the effect of covariates. The covariate effects are quantified in a hypothesis testing framework.

Subset selection such as forward, backward and best subset selections can be employed to select significant variables. However, the procedure is unstable due to the work by Breiman (1995). Tibshirani (1996) proposes a L^1 type of penalty in the linear model selection cases called LASSO, which stands for 'Least Absolute Shrinkage and Selection Operator'. Its application in survival analysis is studied in Tibshirani (1997). Fan and Li (2001) and Fan and Li (2002) propose a variant of LASSO called SCAD, which stands for 'Smoothly Clipped Absolute Deviation'. SCAD has oracle property, namely, the true regression coefficients are estimated as if we know the correct model. In both methods, a penalty is added to the likelihood to suppress some of the variables. But the methods are limited only in linear variable selection cases. If some of the covariate effects are nonlinear, a possibly large number of parametric terms for each covariate are needed to correctly recover the underlying effects. Furthermore, Faraggi and Simon (1998) and Ibrahim, Chen, and MacEachern (1999) investigated Bayesian variable selection methods for censored survival data.

Smoothing spline ANOVA models provide a unified framework for Gaussian regression, generalized regression, density estimation and hazard rate regression, see Wahba (1990), Wahba, Wang, Gu, Klein, and Klein (1995), Lin, Wahba, Xiang, Gao, Klein, and Klein (2000), Gao, Wahba, Klein, and Klein (2001), Gu (2002) for details. Zhang, Wahba, Lin, Voelker, Ferris, Klein, and Klein (2002) proposes a likelihood basis pursuit method. The actual model selection is conducted using a sequential bootstrap procedure. Recently, Lin and Zhang (2002) extends the model selection idea in linear cases for Gaussian data to nonlinear cases using smoothing spline ANOVA models, which is called COSSO. COSSO stands for 'COmponent Selection and Smoothing Operator'. By imposing a penalty on the sum of the component norms instead of the usual squared norms, the model has very good properties in terms of selecting the right components and estimating. The COSSO methodology can model reasonably high dimensional data. The global modeling using penalized likelihood and the smoothing spline ANOVA is ideal for nonparametric model selection. Wood, Kohn, Shively, and Jiang (2002) used Bayesian formulation and Gibbs sampler to select a model for nonparametric spline regression. However, the authors commented the Bayesian method can only efficiently implemented for up to 5 to 6 components.

This paper considers model selection in Cox proportional hazard model. We formulate the problem using penalized partial likelihood with the sum of the absolute norms, which is different from the usual squared norm in the usual smoothing spline ANOVA case. An equivalent reformulation of the penalized partial likelihood leads to an efficient working algorithm. We derive a leave-one-out cross validation criterion based upon partial likelihood cross validation. Very good model selection performance is obtained via our suggested method. We sketch an extension to incorporate non proportionality in the end.

The article is organized as following. Section 2 formulates the problem using penalized likelihood. An efficient algorithm is given in section 3. The cross validation criterion to choose the smoothing parameter is investigated in section 4. We present some simulation results in section 5. The method is applied to real data sets in section 6. An extension to incorporate non proportionality models is discussed in section 7.

2 The problem

2.1 Nonparametric proportional hazard model

Let $T, C, \mathbf{x}$ be respectively survival time, censoring time and their associated covariates. Observing $(Z_i, \delta_i, \mathbf{x}_i), i = 1, ...n$, where $Z_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I_{[T_i \leq C_i]}$, one is to estimate the hazard rate $\eta(\mathbf{x})$. It is assumed that T and C are conditionally independent given x and that the censoring mechanism is uninformative.

Denote the survival function as $S(t|x) = \text{Prob}(T > t|x)$ and the corresponding hazard function as $h(t|x) = -\partial S(t|x)/\partial t = \exp(\eta(t, x))$. It is easy to see the full likelihood is

$$\prod_{i=1}^{n} h(Z_i|x_i)^{\delta_i} \prod_{i=1}^{n} S(Z_i|x_i). \tag{2.1}$$

The penalized log likelihood is then

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i \eta(Z_i, x_i) - \int_0^{Z_i} e^{\eta(t,x_i)} dt\} + \lambda J(\eta), \tag{2.2}$$

where $J(\eta)$ is a penalty of the roughness on η.

Without loss of generality, we assume there are no ties in the observed failure time. We can use Breslow (1974) to handle ties. Let $t_1^0 < ... < t_N^0$ be the ordered observed failure time. Let (j) be the label for the item falling at t_j^0 so the covariates associated with the N failures are $x_{(1)}, ..., x_{(N)}$. Denote R_j as the risk set right before the time t_j^0:

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

We consider proportional hazard models

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

where $\eta(x)$ is nonparametrically modeled. In this setup, $\eta(t,x) = \eta_0(t) + \eta_1(x)$. Consider the Breslow's estimate for baseline hazard in which the cumulative baseline hazard $H_0(t) = \sum_{i=1}^{N} h_j I_{[t_j^0 \le t]}$. Then the log likelihood is

$$\sum_{i=1}^{N} \log h_i + \sum_{i=1}^{n} \delta_i \eta(x_i) - \sum_{i=1}^{n} [\exp(\eta(x_i)) \sum_{j=1}^{N} h_j I_{j \in R_i}]. \tag{2.3}$$

Taking derivative with respect to h_i and solving for h_j, one gets

$$\hat{h}_j = \{\sum_{i \in R_j} \exp(\eta(x_i))\}^{-1}.$$

The penalized partial log likelihood is then

$$-\frac{1}{n} \sum_{i=1}^{N} \{\eta(x_{(i)}) - \log[\sum_{j \in R_i} \exp(\eta(x_j))]\} + \lambda J(\eta). \tag{2.4}$$

2.2 Smoothing spline ANOVA

The functional decomposition of η of d variables is

$$\eta = \eta_0 + \sum_{j=1}^{d} \eta_j(x^{(j)}) + \sum_{j<k} \eta_{j,k}(x^{(j)}, x^{(k)}) + ... + \eta_{1,...,d}(x^{(1)}, ..., x^{(d)}). \tag{2.5}$$

It can be written as $\eta = \eta_0 + \sum_{\alpha=1}^{p} \eta_\alpha$, where p is the number of components.

In the smoothing spline analysis of variance models over $[0,1]^d$, it is assumed η lies in a reproducing kernel Hilbert space corresponding to the decomposition (2.5). See (Aronszajn 1950) and Wahba (1990). Let $W_2^{(j)}[0,1] = \{f : f, f'$ are absolutely continuous and $f'' \in L_2[0,1]\}, j = 1, ..., d$ be the second order Sobolev Hilbert space corresponding to the function space over $X^{(j)}$ on $[0,1]$. When endowed with the inner product

$$(f,g)_{W_2^{(j)}[0,1]} = \{\int_0^1 f(t)dt\}\{\int_0^1 g(t)dt\} + \{\int_0^1 f'(t)dt\}\{\int_0^1 g'(t)dt\} + \int_0^1 f''(t)g''(t)dt,$$

$W_2^{(j)}[0,1]$ can be decomposed as $W_2^{(j)}[0,1] = \{1\} \oplus \{W^{(j)}\}$, where $W^{(j)}$ is a reproducing kernel Hilbert space with the norm

$$\|f\|^2_{W^{(j)}} = \{\int_0^1 f'(t)dt\}^2 + \int_0^1 (f''(t))^2 dt,$$

The reproducing kernel corresponds to $W^{(j)}$ is

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

with $k_1(s) = s - 0.5, k_2(s) = (k_1^2(s) - 1/12)/2, k_4(s) = (k_1^4(s) - k_1^2(s)/2 + 7/240)/24$. Then the tensor product of d variables is

$$\bigotimes_{j=1}^{d} W_2^{(j)}[0,1] = \bigotimes[\{1\} \bigoplus \{W^{(j)}\}] = \{1\} \bigoplus \sum_{j=1}^{d} [\{W^{(j)}\}] \bigoplus \sum_{j<k} [\{W^{(j)}\} \bigotimes \{W^{(k)}\}] + ... \tag{2.6}$$

The tensor sum in (2.6) can be truncated using the same functional ANOVA argument. We will denote the truncated series as $\mathcal{H} = \{1\} \bigoplus \mathcal{W} = \{1\} \bigoplus \sum_{\alpha=1}^{p}[\{W^\alpha\}]$. The reproducing kernel of the main effect on $X^{(j)}$ is $K^{(j)}(s,t)$. The reproducing kernel for a product term is the product of individual reproducing kernels. For example, the reproducing kernel corresponding to two way interaction $X^{(j)}$ and $X^{(k)}$ is $K^{(j)}(s,t)K^{(k)}(s,t)$. With some abuse of the notation, we use $K_\alpha(x,x')$ for the reproducing kernel of the term η_α in the functional decomposition of η.

To model a categorical variable on the domain $\{1,...,k\}$, we use $K(x,y) = kI_{(x=y)} - 1$ as the reproducing kernel. In this sense, we can handle continuous and categorical variables in a unified fashion.

2.3 The COSSO estimate

The usual smoothing spline ANOVA penalizes the squared norm of each component, namely

$$J(\eta) = \sum_{\alpha=1}^{p} \theta_\alpha^{-1} ||P^\alpha \eta||^2,$$

where θ_α can be adapted to different smoothness of each component.

The COSSO estimate minimizes

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j \in R_i} \exp(\eta(x_j))]\} + \lambda J(\eta) \text{ with } J(\eta) = \sum_{\alpha=1}^{p} \|P^\alpha \eta\|, \tag{2.7}$$

with the penalty being the sum of the absolute norms instead of sum of squared norms. The LASSO in linear cases can be seen as a special case of the COSSO. For the input space $[0,1]^d$, consider the linear function space $\{1\} \oplus \{x_1 - 1/2\} \oplus ... \oplus \{x_d - 1/2\}$, with the usual L^2 inner product $(f,g) = \int fg$. The penalty in the COSSO becomes $J(\eta) = (12)^{-1/2} \sum_{j=1}^{d} |\beta_j|$ for $\eta(x) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j$. This is equivalent to the L^1 norm on the linear coefficients, which leads to the LASSO estimator.

The difference between the form of COSSO and the usual smoothing spline mirrors the difference between the LASSO and the ridge regression.

2.4 Existence and the form of the solution

The existence of the smoothing type estimate in various settings has been studied by Tapia and Thompson (1978), Silverman (1982), Gu and Qiu (1993). For COSSO estimate, $J(\eta)$ is clearly convex, and the existence of a solution in the null space guarantees the existence of the solution as a penalized likelihood estimator.

Lemma 2.1. *There exists a solution to (2.7).*

Theorem 2.2 (Representer theorem, Kimeldorf and Wahba (1971)). *The solution to (2.7) has the form* $\eta = \sum_{i=1}^{n} \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_i) c_i$.

For the proofs, see appendix.

3 Algorithm

The COSSO problem (2.7) is very hard to optimize due to the sum of the absolute norms. Lin and Zhang (2002) gives an equivalent formulation as follows

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \lambda_0\sum_{\alpha=1}^{p}\theta_\alpha^{-1}\|P^\alpha\eta\|^2\} \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M, \theta_\alpha \ge 0, \alpha = 1, ..., p. \tag{3.1}$$

This formulation has the same form as the usual smoothing spline ANOVA setup except we constraint the sum of θ's. It is possible to simultaneously minimize the objective function in (3.1) with respect to both θ and η, however, the optimization is nonlinear and complex. It is noticed that it is much easier to optimize the objective function with respect to one set of variables (θ or η) when the other set (η or θ) is fixed. In our implementation, we do it iteratively between θ and η.

The exact solution to (2.7) has a form $\eta(x) = \sum_{i=1}^{N}\sum_{\alpha=1}^{p}\theta_\alpha K_\alpha(x, x_i)c_i$. In the case when n is large, one way to reduce the computational load is to use the parsimonious approach as suggested by Xiang and Wahba (1996), Ruppert and Carroll (2000), Lin, Wahba, Xiang, Gao, Klein, and Klein (2000). For large data set, we can use a subset $x_{1*}, ..., x_{m*}$ of $x_1, ..., x_n$ as the basis functions and denote the corresponding solution as

$$\eta(x) = \sum_{i=1}^{m}\sum_{\alpha=1}^{p}\theta_\alpha K_\alpha(x, x_{i*})c_i.$$

Denote equation (3.1) as $A(c, \theta)$. Using the notation $K_\theta = \sum_\alpha K_\alpha$, $\{Q_{m\times m}\}_{(k,l)} = K_\theta(x_{k*}, x_{l*})$, $\{U_{n\times m}\}_{(k,l)} = K_\theta(x_k, x_{l*})$, $\delta = (\delta_1, ..., \delta_n)^T$, we can write (3.1) as

$$A(c, \theta) = -\frac{1}{n}\delta^T Uc + \frac{1}{n}\sum_{i=1}^{N}\log(\sum_{j\in R_i} e^{U_j c}) + \lambda c^T Qc, \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M,\ \theta_\alpha \ge 0, \tag{3.2}$$

where U_j is the jth row of U. When θ is fixed, it is easy to see

$$\begin{aligned}\frac{\partial A}{\partial c} &= -\frac{1}{n}U^T\delta + \frac{1}{n}\sum_{i=1}^{N}\frac{\sum_{j\in R_i}U_j^T e^{U_j c}}{\sum_{j\in R_i}e^{U_j c}} + 2\lambda Qc;\\ \frac{\partial^2 A}{\partial c\partial c^t} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i}U_j^T U_j e^{U_j c}}{\sum_{j\in R_i}e^{U_j c}} - \frac{\sum_{j\in R_i}U_j^T e^{U_j c}}{\sum_{j\in R_i}e^{U_j c}}\frac{\sum_{j\in R_i}U_j e^{U_j c}}{\sum_{j\in R_i}e^{U_j c}}\} + 2\lambda Q.\end{aligned} \tag{3.3}$$

The updating equation for c is then

$$c = c_0 - (\frac{\partial^2 A}{\partial c\partial c^t})_{c_o}^{-1}(\frac{\partial A}{\partial c})_{c_0}.$$

Denote $G_{m\times p} = (\{K_1(x_{k*}, x_{l*})\}c, ..., \{K_p(x_{k*}, x_{l*})\}c)$ and $S_{n\times p} = (\{K_1(x_k, x_{l*})\}c, ..., \{K_p(x_k, x_{l*})\}c)$, equation (3.1) can be written as

$$-\frac{1}{n}\delta^T S\theta + \frac{1}{n}\sum_{i=1}^{N}\log(\sum_{j\in R_i} e^{S_j\theta}) + \lambda c^T G\theta \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M, \theta_\alpha \ge 0, \tag{3.4}$$

where S_j is the jth row of S. When c is fixed, it is easy to see

$$\begin{aligned}\frac{\partial A}{\partial\theta} &= -\frac{1}{n}S^T\delta + \frac{1}{n}\sum_{i=1}^{N}\frac{\sum_{j\in R_i} S_j^T e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}} + \lambda G^T c;\\ \frac{\partial^2 A}{\partial\theta\partial\theta^T} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i} S_j^T S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}} - \frac{\sum_{j\in R_i} S_j^T e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\frac{\sum_{j\in R_i} S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\}.\end{aligned} \tag{3.5}$$

Using Taylor expansion,

$$A(c,\theta) \approx A(c,\theta_0) + (\theta-\theta_0)^T(\frac{\partial A}{\partial\theta})_{\theta_0} + \frac{1}{2}(\theta-\theta_0)^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}(\theta-\theta_0).$$

The updating equation for θ is then

$$\frac{1}{2}\theta^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta + [(\frac{\partial A}{\partial\theta})_{\theta_0} - (\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta_0]^T\theta \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M, \theta_\alpha \ge 0, \tag{3.6}$$

4 Leave-one-out cross validation

From (2.4), a natural criterion to tune the smoothing parameter is to minimize the predicted average log likelihood with respect to future observations $(\tilde{x}_1, \tilde{z}_1, \tilde{\delta}_1), ..., (\tilde{x}_m, \tilde{z}_m, \tilde{\delta}_m)$, namely to minimize

$$-\frac{1}{m}\{\sum_{i=1}^{m}\tilde{\delta}_i h(\tilde{z}_i|\tilde{x}_i) + \sum_{i=1}^{m} S(\tilde{z}_i|\tilde{x}_i)\}.$$

While future observations are not available, we use a leave-out-one cross validated criterion in the following to approximate the predicted average log likelihood:

$$CV = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta^{[-i]}(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta(x_j))], \tag{4.1}$$

where $\eta^{[-i]}(x_i)$ stands for the fitted value at x_i when (2.4) is fitted without the ith data point. The second term in (4.1) is not cross validated. Similar arguments can be found in Xiang and Wahba (1996),Lin, Wahba, Xiang, Gao, Klein, and Klein (2000), Gao, Wahba, Klein, and Klein (2001). This is essentially likelihood cross validation.

For fixed θ's, the leave-out-one version of (2.4) is

$$-\frac{1}{n-1}\sum_{i\neq k}\delta_i\eta(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta(x_j))] + \lambda J(\eta), \tag{4.2}$$

when kth data point is left out.

We can approximate the minimizer of (4.2) from solution c_0 of (2.4) by one step iteration as following

$$c^{[-i]} \approx c_0 - H_{c_o}^{-1}\{(\frac{\partial A}{\partial c})_{c_0} + \frac{U^T\delta}{n} - \frac{U^T\delta^{[-i]}}{n-1}\}, \tag{4.3}$$

where $\delta^{[-i]} = (\delta_1, ..., \delta_{i-1}, 0, \delta_{i+1}, ..., \delta_n)^T$ and

$$H = \frac{\partial^2 A}{\partial c \partial c^t}.$$

Follow simple algebra,

$$\begin{aligned} c^{[-i]} &\approx c_0 - H_{c_o}^{-1}\{(\frac{\partial A}{\partial c})_{c_0} - \frac{U^T\delta}{n(n-1)} + \frac{\delta_i U_i^T}{n-1}\}, \\ &\approx c_0 - H_{c_o}^{-1}\{-\frac{U^T\delta}{n(n-1)} + \frac{\delta_i U_i^T}{n-1}\} \end{aligned} \tag{4.4}$$

since $c_0 \approx c_0 - H_{c_o}^{-1}(\frac{\partial A}{\partial c})_{c_0}$ with convergence. So

$$\eta^{[-i]}(x_i) \approx \eta(x_i) - \frac{1}{n-1} U_i H_{c_o}^{-1}(\delta_i U_i^T - \frac{U^T\delta}{n}). \tag{4.5}$$

It follows

$$\begin{aligned} CV =& -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta(x_j))] + \frac{1}{n}\sum_{i=1}^{n}(\delta_i\eta(x_i) - \delta_i\eta^{[-i]}(x_i)) \\ \approx& -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta(x_j))] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\delta_i U_i H_{c_o}^{-1}(\delta_i U_i^T - \frac{U^T\delta}{n}) \\ =& -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta(x_j))] + \{\frac{\mathrm{tr}(\Delta U H^{-1} U^T \Delta)}{n(n-1)} - \frac{\delta^T U H^{-1} U^T \delta^T}{n^2(n-1)}\} \\ =& PL + \{\frac{\mathrm{tr}(\Delta U^T H^{-1} U \Delta)}{n(n-1)} - \frac{\delta^T U^T H^{-1} U \delta^T}{n^2(n-1)}\} \end{aligned} \tag{4.6}$$

where $\Delta = \mathrm{diag}(\delta_1, ..., \delta_n)$ and PL stands for fitted partial likelihood. Similar criterion is obtained in Chapter 7.2 of Gu (2002) by approximating a leave-one-out cross validation estimate of the relative Kullback-Leibler loss.

A simple modification of (4.6) which we will call it approximate cross validation (**ACV**) is

$$\mathbf{ACV}(M) = L_M(\lambda, \theta, c) = PL + \frac{N}{n}\{\frac{\mathrm{tr}(U^T H^{-1} U)}{n(n-1)} - \frac{\mathbf{1}^T U^T H^{-1} U \mathbf{1}}{n^2(n-1)}\}. \tag{4.7}$$

The algorithm works as following:

1 Neglect M and fix $\theta = \theta_0 = (1, ..., 1)^T$, for λ on a reasonable range, calculate $L_M(\lambda, \theta_0, \hat{c}_{\lambda,\theta_0})$ and set $\lambda = \lambda_0$ which gives minimum of L_M.

2 For M on a reasonable range, use one step quadratic expansion (3.6) around $A(c_{\lambda_0,\theta_0}, \theta_0)$ to solve for $\hat{\theta}_{\lambda_0,M}$ $(\hat{\theta})$. Solve for $\hat{c}_{\lambda_0,\hat{\theta}}$ and calculate $L_M(\lambda_0, \hat{\theta}, \hat{c}_{\lambda_0,\hat{\theta}})$. Choose the $\hat{M}$ which minimizes L_M.

3 The final estimate is $\hat{\eta} = \sum_{\alpha=1}^{p}\hat{\eta}_\alpha$ with fitted components $\hat{\eta}_\alpha = \{\hat{\theta}_{\lambda_0,\hat{M}}\}_\alpha K_\alpha \hat{c}_{\lambda_0,\hat{\theta}}$.

5 Simulation

We measure the accuracy of the estimates by integrated square error $ISE = E_X\{\eta(X) - \hat{\eta}(X)\}^2$. For each replicate of the simulation, the ISE is estimated by Monte Carlo integration using 10000 test points from the same distribution as the training points. We run the simulation 100 times and average.

We will use the examples in Lin and Zhang (2002). Namely, the following 4 functions are used as building block of the hazard function:

$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{sin(2\pi t)}{2 - sin(2\pi t)};$$

$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t).$$

To include categorical variable, We add

$$g_5(t) = t, \quad t = 0, 1.$$

We consider additive model on R^{10} with the true hazard function $t|x$ being $\eta(x) = 5g_1(x_1)+3g_2(x_2)+4g_3(x_3)+6g_4(x_4)+3g_5(I_{(x_5>0.6)})$. Including two more categorical variables $I_{(x_6<0.8)}$ and $I_{(x_7>0.2)}$. Two cases $cov(x^i, x^j) = 0, i \neq j$ and $cov(x^i, x^j) = 0.5, i \neq j$ are considered. We simulate $n = 100,\ 200$ data points from the exponential hazard function

$$h(t|x) = \exp(\eta(x)).$$

The distribution of censoring time is exponentially distributed with mean $U\exp(-\eta(x))$, where U is randomly generated from the uniform distribution from $[1, 3]$. The empirical censoring rate is about 35%. We use either all the observations or randomly chosen $m = 50$ observations as basis functions. The result is summarized in Table 5.1 for the independent case and Table 5.2 for the compound symmetry case. We summarize ISE in Table 5.3. We can see using a subset of observations as the basis does not degrade the performance. We plot the 5th, 50th, 95th best estimates according to ISE in Figure 5.1, Figure 5.2, Figure 5.3, and Figure 5.4 when all the observations are used as basis functions. The corresponding estimates while 50 basis functions are used are shown in Figure 8.1, Figure 8.2, Figure 8.3 and Figure 8.4 listed in the Appendix. Also included in the appedix are the estimates for $n = 800$. The ISE versus the sample size plot is shown in Figure 5.5.

6 Real data

6.1 Lung cancer data

The data is from the Veteran's Administration lung cancer trail, listed in Kalbfleisch and Prentice (1980), pp.223-224. There are 137 patients in the study and there are 9 censored observations among those. The time variable is survival in days and the regressors are:
treatment, 1=standard 2=test
celltype, 1=squamous, 2=smallcell, 3=adeno, 4=large
Karnofsky performance score (100=good)
months from diagnosis to randomisation
age in years

n(m)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
100(100)	100	95	100	100	100	8	8	16	19	13
100(50)	100	94	100	100	100	6	8	14	13	14
200(200)	100	100	100	100	100	2	6	11	13	8
200(50)	100	100	100	100	100	2	6	13	10	8
400(50)	100	100	100	100	100	2	2	7	10	7
800(50)	100	100	100	100	100	0	0	1	5	5

Table 5.1: Example 1: Frequency of variables selected and model sizes when $cov(x^i, x^j) = 0$. When $n = 100$ and $m = 100$, 53 out of 100 models are selected correctly. When $n = 100$ and $m = 50$, 55 out of 100 models are selected correctly. When $n = 200$ and $m = 200$, 70 out of 100 models are selected correctly. When $n = 200$ and $m = 50$, 69 out of 100 models are selected correctly.When $n = 400$ and $m = 50$, 76 out of 100 models are selected correctly. When $n = 800$ and $m = 50$, 89 out of 100 models are selected correctly

n(m)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
100(100)	98	67	100	100	100	8	15	12	17	19
100(50)	96	70	100	99	99	7	16	21	15	24
200(200)	100	100	100	100	100	4	2	12	14	11
200(50)	100	99	100	100	100	4	2	6	14	10
400(50)	100	100	100	100	100	1	2	3	5	4
800(50)	100	100	100	100	100	0	1	3	9	5

Table 5.2: Example 1: Frequency of variables selected and model sizes when $cov(x^i, x^j) = 0.5$. When $n = 100$ and $m = 100$, 32 out of 100 models are selected correctly.When $n = 100$ and $m = 50$, 27 out of 100 models are selected correctly. When $n = 200$ and $m = 200$, 69 out of 100 models are selected correctly. When $n = 200$ and $m = 50$, 73 out of 100 models are selected correctly.When $n = 400$ and $m = 50$, 88 out of 100 models are selected correctly. When $n = 800$ and $m = 50$, 85 out of 100 models are selected correctly.

n(m)	cov=0	cov=0.5
100(100)	3.91(0.11)	4.08(0.22)
100(50)	3.86(0.13)	4.12(0.20)
200(200)	1.17(0.05)	1.02(0.05)
200(50)	1.10(0.05)	0.89(0.05)
400(50)	0.36(0.02)	0.32(0.02)
800(50)	0.14(0.01)	0.16(0.01)

Table 5.3: Estimated integrated square error for the simulation.

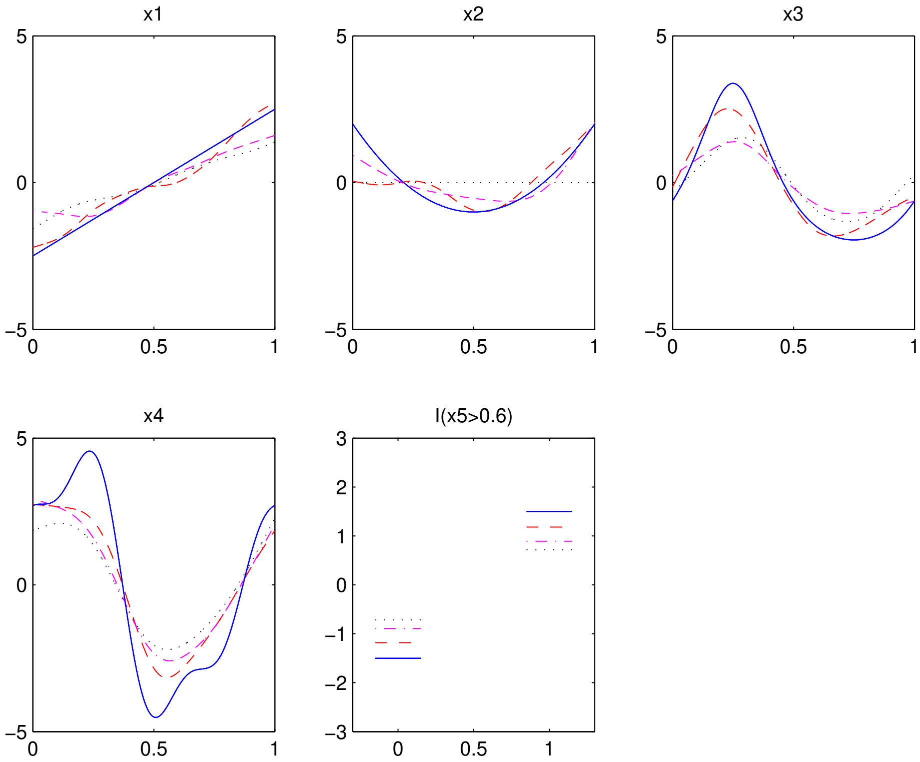

Figure 5.1: The estimated effects when $n = 100, m = 100$ and $cov(x^i, x^j) = 0$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

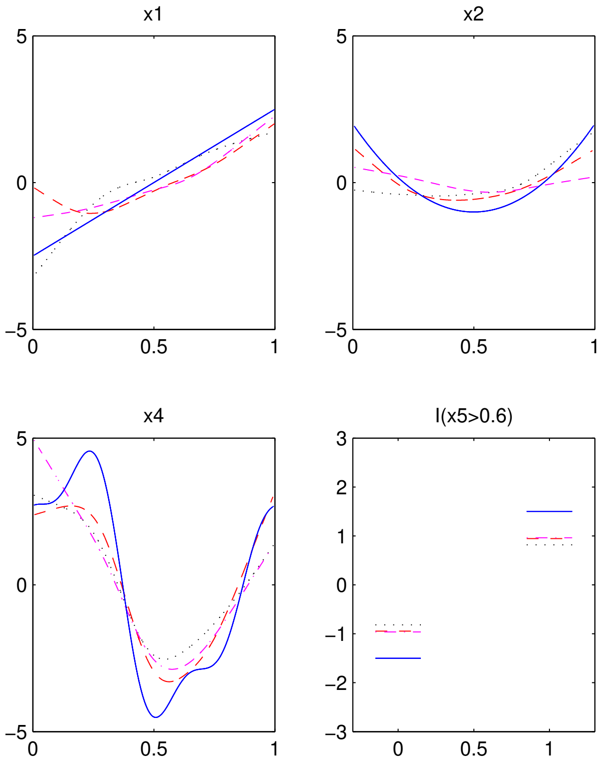

Figure 5.2: The estimated effects when $n = 100, m = 100$ and $cov(x^i, x^j) = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

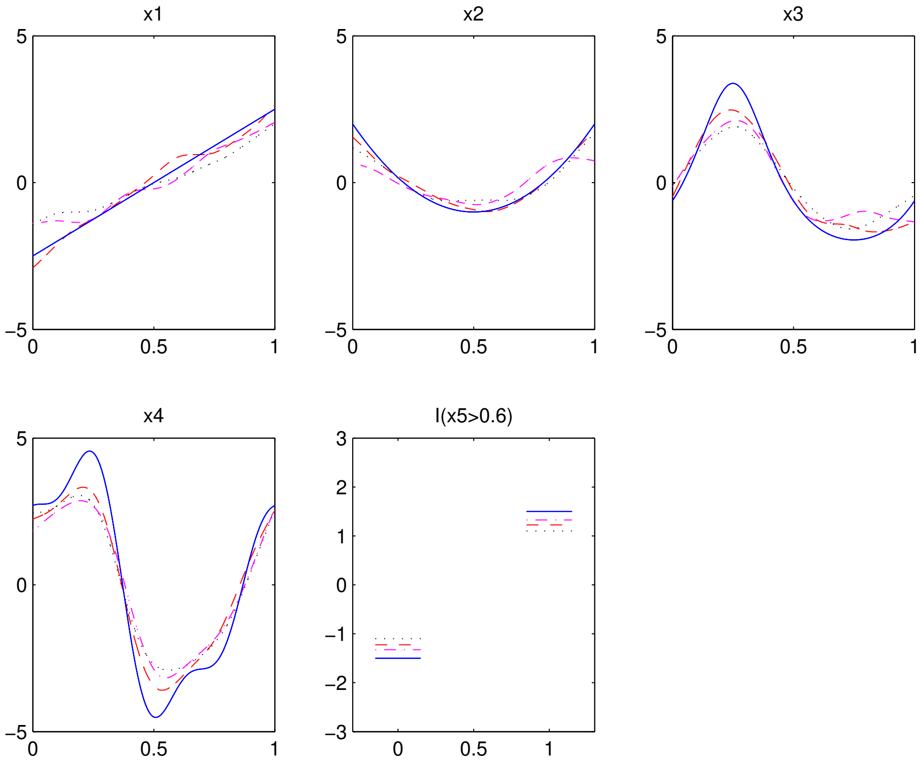

Figure 5.3: The estimated effects when $n = 200, m = 200$ and $cov(x^i, x^j) = 0$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

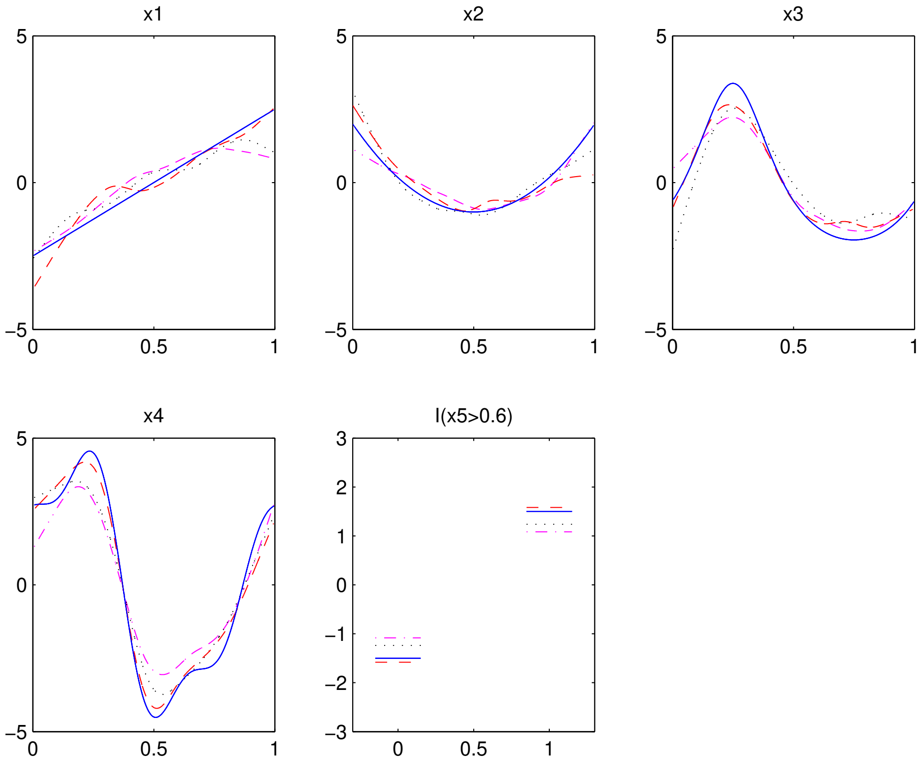

Figure 5.4: The estimated effects when $n = 200, m = 200$ and $cov(x^i, x^j) = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

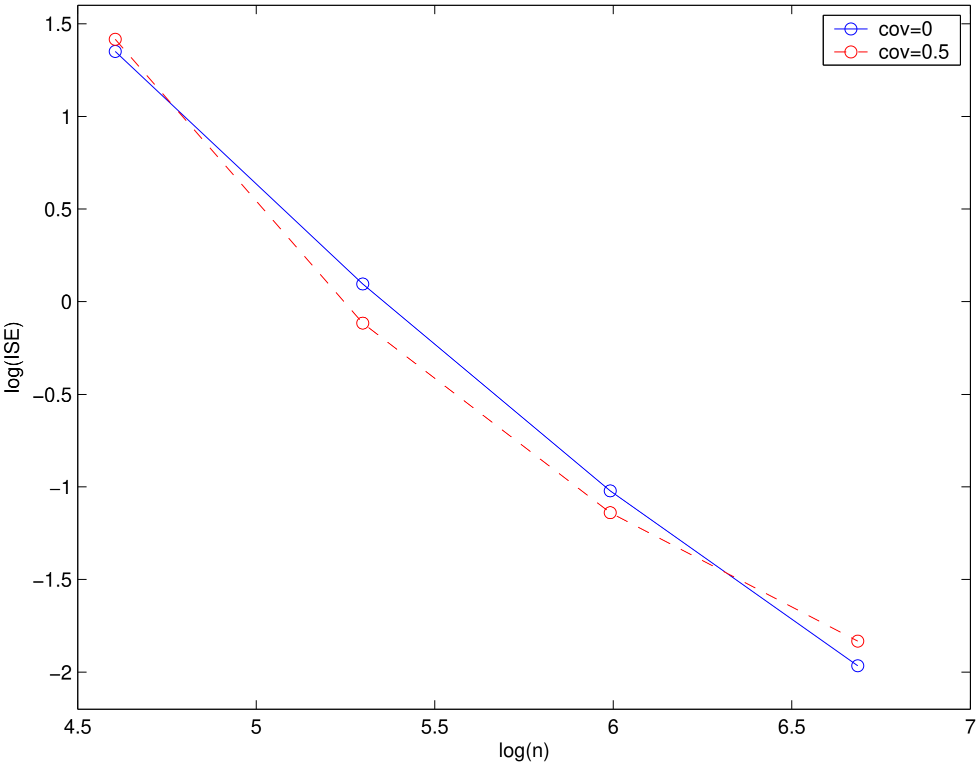

Figure 5.5: The ISE versus n plot in the log scale. The estimated slopes for $cov = 0$ and $cov = 0.5$ are -1.60 and -1.55, respectively.

prior therapy 0=no, 1=yes

Tibshirani (1997) treated cell type as continuous variable and found only Karnofsky performance score is selected in the final model using LASSO or stepwise regression. We treat cell type as categorical variable. For parametric Cox proportional model, we use Mallow's Cp as the criterion for stepwise selection. The stepwise procedure chooses karnofsky performance score and cell type in the final model. The estimated effects for cell type are -0.550 for squamous, 0.166 for small cell, 0.608 for adeno and -0.224 for large cell. Clearly, cell type can not treated as a continuous variable since the effect is not linear at all! Our procedure yields the estimate in Figure 6.1. The coefficients for different cell types are -0.517 for squamous, 0.203 for small cell, 0.554 for adeno and -0.240 for large cell. The estimates are quite close to those obtained by stepwise variable selection.

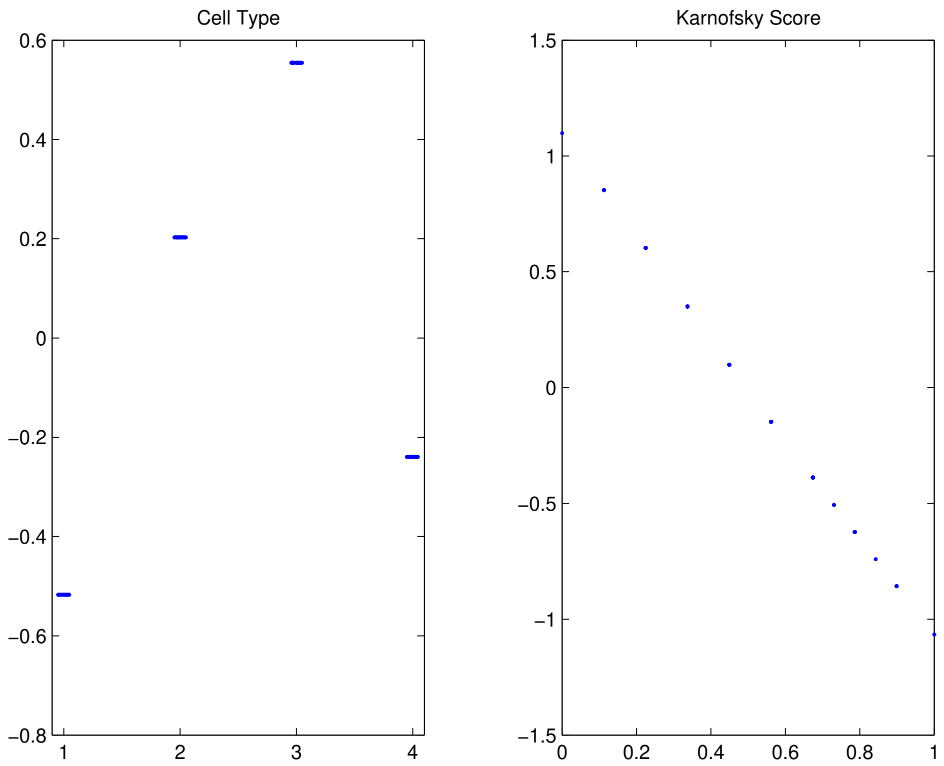

Figure 6.1: Fitted main effects for lung cancer data.

6.2 PBC data

We analyze the primary biliary cirrhosis (PBC) data from Therneau and Grambsch (2000). A slightly different data is analyzed by Cox proportional hazard model by Fleming and Harrington (1991). The data are from the Mayo Clinic trail in primary biliary cirrhosis of liver conducted between 1974 and 1984. PBC is a progressive disease thought to be of an autoimmune origin. The subsequent inflammatory process eventually leads to cirrhosis and destruction of the liver's bileducts and death of the patient. A more detailed discussion can be found in Dickson et al. (1989). 312 patients from a total of 424 patients who agreed to participate in the randomized trail are eligible for the analysis. Of those, 125 patients died before the end of follow-up. We constrain our attention to the 11 covariates as shown in Table 4.4.2 of Fleming and Harrington (1991), including continuous variables age (age), albumin (alb), alkaline phos (alk), bilirubin (bili), platelet count (plat), pro time (prot) as well as categorical variables ascites (asc), edema (ede), hepatomegaly (hep), sex (sex) and spiders (spid). See the appendix for detailed description of the variables. The PBC database is a valuable resource to liver specialists because PBC is a rare but fatal disease.

We take log transformation to alk, bili and prot since they are skewed to the right. We apply our procedure to the data. The cross validation curve selects $M = 5$ (Figure 6.2). The fitted main effects are shown in Figure 6.3. Comparing with Table 4.4.3 in Fleming and Harrington (1991), our model has one extra term asc. From the plot, the data should be fitted using proportional hazard models with parametric main effects. To assess the model, we use 100 bootstrap samples. The result is summarized in Table 5.3. The bootstrap has an average number of 6.71 terms. And the final fitted model contains all the terms which appear more than half in the bootstrap simulation.

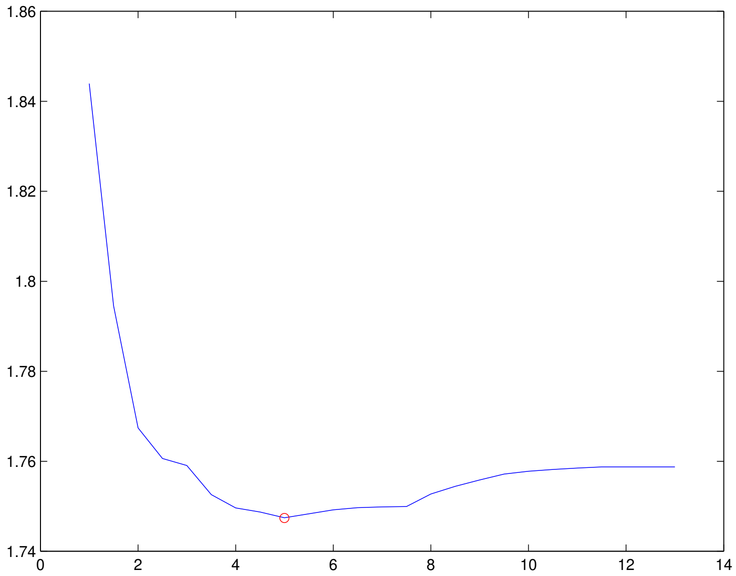

Figure 6.2: Cross validation curve for PBC data. The minimum is obtained at $M = 5$.

age	alb	log(alk)	asc	log(bili)	edema	hep	plat	log(prot)	sex	spid
100	**96**	17	**62**	**100**	**83**	42	27	**82**	36	26

Table 6.1: Frequency of variables for PBC data in the 100 bootstrap runs. The components in the fitted final model are shown in boldface. Average model size is 6.71.

6.3 Re-analysis of PBC data

There are 17 covariates in PBC data. A more ambitious task is to select important variables among them. Tibshirani (1997) analyzed the data using LASSO, which is a linear model selection procedure.

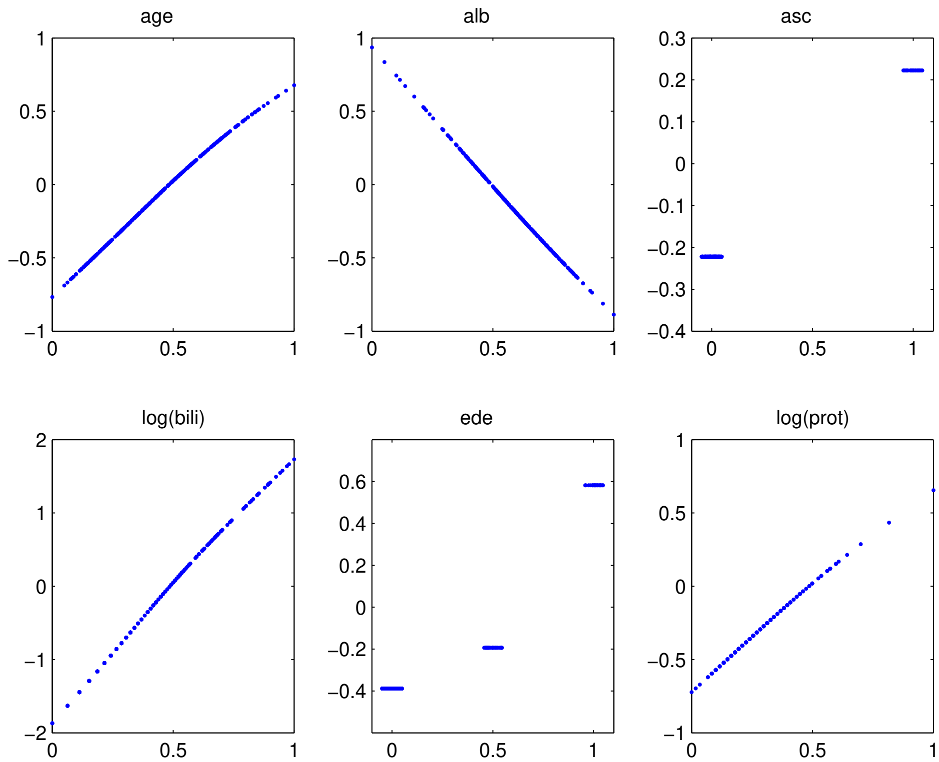

Figure 6.3: Fitted effects for PBC data.

We carry out the same analysis using all the variables as listed in the Appendix. Namely, there are 10 continuous variables and 7 categorical variables. We apply our procedure to 276 observations with no missing data in the covariates. As reported in Tibshirani (1997), stepwise selection has age, ede, bili, alb, cop, sgot, prot and stage; totally 8 variables appear in the final selected model. The LASSO procedure selects three more variables, sex, asc and spid. Our procedure selects two more variables sex and chol. Quite interestingly, the stepwise model selected only those covariates with absolute Z-scores larger than 2.00, and our model selected only those covariates with absolute Z-scores larger than 1.00, where Z-scores refer to as the scores obtained via 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. It remains unclear to what extent accounting for non linearity contributes to the discrepancy. The fitted effects of our model is shown in Figure 6.4. Except for cop, possible linear effects are quite obvious for the other covariates.

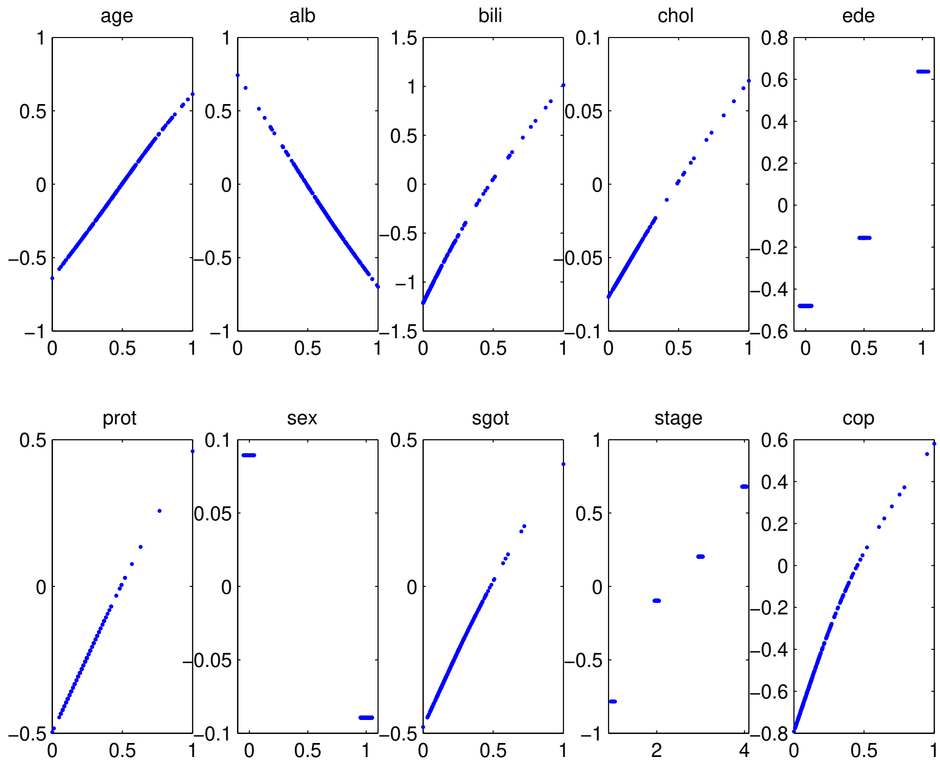

Figure 6.4: Fitted main effects while 17 covariates are considered.

7 Extending the model to non proportional models

Gu (1996) and Gu (1998) studied penalized likelihood hazard regression without the proportionality assumption. We sketch the derivation under this setting.

Consider the model

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i\eta(Z_i,x_i)-\int_{L_i}^{Z_i}e^{\eta(t,x_i)}dt\}+\lambda J(\eta), \tag{7.1}$$

it is the model in Equation (2.2). Making no proportional assumption and using the notation $K_\theta = \sum_\alpha K_\alpha$, $\{Q_{N\times N}\}_{(k,l)} = K_\theta((t_{(k)},x_{(k)}),(t_{(l)},x_{(l)}))$, $\{U_{n\times N}(t)\}_{(k,l)} = K_\theta((t,x_k),(t_{(l)},x_{(l)}))$, we can write (7.1) as

$$A(c,\theta)=-\frac{1}{n}\mathbf{1}^TQc+\frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}\exp(U_i(t)c)dt+\lambda c^TQc,\ \text{s.t.}\ \sum_{\alpha=1}^{p}\theta_\alpha\le M,\theta_\alpha\ge 0, \tag{7.2}$$

where $U_i(t)$ is the ith row of $U(t)$. When θ is fixed, it is easy to see

$$\begin{aligned}\frac{\partial A}{\partial c} &= -\frac{1}{n}Q\mathbf{1}+\frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}U_i^T(t)\exp(U_i(t)c)dt+2\lambda Qc;\\ \frac{\partial^2 A}{\partial c\partial c^t} &= \frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}U_i^T(t)U_i(t)\exp(U_i(t)c)dt+2\lambda Q.\end{aligned} \tag{7.3}$$

The updating equation for c is then $c=c_0-(\frac{\partial^2 A}{\partial c\partial c^t})_{c_o}^{-1}(\frac{\partial A}{\partial c})_{c_0}$.

Denote $G_{N\times p}=(\{K_1((t_{(k)},x_{(k)}),(t_{(l)},x_{(l)})\}c,...,\{K_p(t_{(k)},x_{(k)}),(t_{(l)},x_{(l)}))\}c)$ and $S_{n\times p}(t)=(\{K_1((t,x_k),(t_{(l)},x_{(l)}))\}c,...,\{K_p((t_k,x_k),(t_{(l)},x_{(l)}))\}c)$, equation (7.1) can be written as

$$-\frac{1}{n}\mathbf{1}^TG\theta+\frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}\exp(S_i(t)\theta)dt+\lambda c^TG\theta\ \ \text{s.t.}\ \sum_{\alpha=1}^{p}\theta_\alpha\le M,\theta_\alpha\ge 0, \tag{7.4}$$

where $S_i(t)$ is the ith row of $S(t)$. When c is fixed, it is easy to see

$$\begin{aligned}\frac{\partial A}{\partial \theta} &= -\frac{1}{n}G^T\mathbf{1}+\frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}S_i^T(t)\exp(S_i(t)\theta)dt+\lambda G^Tc;\\ \frac{\partial^2 A}{\partial\theta\partial\theta^T} &= \frac{1}{n}\sum_{i=1}^{n}\int_{L_i}^{Z_i}S_i^T(t)S_i(t)\exp(S_i(t)\theta)dt.\end{aligned} \tag{7.5}$$

Using Taylor expansion,

$$A(c,\theta)\approx A(c,\theta_0)+(\theta-\theta_0)^T(\frac{\partial A}{\partial\theta})_{\theta_0}+\frac{1}{2}(\theta-\theta_0)^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}(\theta-\theta_0).$$

The updating equation for θ is then

$$\frac{1}{2}\theta^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta+[(\frac{\partial A}{\partial\theta})_{\theta_0}-(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta_0]^T\theta\ \ \text{s.t.}\ \sum_{\alpha=1}^{p}\theta_\alpha\le M,\theta_\alpha\ge 0, \tag{7.6}$$

A cross validation criterion similar to (4.6) can be derived. And we can use the same algorithm as listed in section 4 to select the best model according to the criterion. The full likelihood model has great potential. If the time and covariate interactions are estimated as zeros, the proportional hazard model assumption is more appropriate for further analysis.

8 Appendix

8.1 Proofs

Proof of Lemma 2.1. Denote (2.7) as $A(\eta)$. We first prove $A(\eta)$ is convex in η.

For any $f, g \in \mathcal{H}$ and $a, b \in R$ satisfying $a + b = 1$, clearly $J(\eta)$ is convex and $\delta\eta$ is linear in η. It suffices to prove $\sum_{i=1}^{n} \log[\sum_{j \in R_i} \exp(\eta(x_j))]$ is convex. For each i, by Holder's inequality,

$$\sum_{j \in R_i} \exp((f+g)(x_j)) \leq \sum_{j \in R_i} \exp(f(x_j) \sum_{j \in R_i} \exp(g(x_j)),$$

which implies

$$\log[\sum_{j \in R_i} \exp((f+g)(x_j))] \leq \log[\sum_{j \in R_i} \exp(f(x_j)] + \log[\sum_{j \in R_i} \exp(g(x_j))].$$

This completes the proof of the convexity of η.

The null space of $J(\eta)$ is a constant. When there is no covariate, the minimizer of $A(\eta)$ is a constant. This is the Nelson-Aalen estimator (Nelson 1972; Aalen 1978), which always exists. Following theorem 2.9 in Gu (2002), we complete the proof. $\square$

Proof of Theorem 2.2. Let the minimizer be

$$\eta = c_0 + \sum_{i=1}^{n} \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_i) c_i + \rho$$

where ρ is some element in $\mathcal{H}$ orthogonal to $1, \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_1), ..., \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_n)$. Denote $c = (c_1, ...c_n)^T$, (2.7) becomes

$$-\frac{1}{n} \sum_{i=1}^{N} \{\eta(x_{(i)}) - \log[\sum_{j \in R_i} \exp(\eta(x_j))]\} + \lambda(c'Qc + \rho^2)^{1/2}, \tag{8.1}$$

and we must find c_0, c, ρ to minimize (8.1). It is then obvious $||\rho||^2 = 0$. c_0 can be further absorbed into the baseline hazard, which involves a rescaling of λ. $\square$

8.2 PBC data

1 Continuous variables
age: age in years
alb: serum albumin in gm/dl
alk: alkaline phosphatase in U/liter
bil: serum bilirunbin in mg/dl
chol: serum cholesterol in mg/dl
cop: urine copper in μg/day
plat: platelets per cubic ml/1000
prot: standardized prothrombin time in seconds
sgot: liver enzyme (now called AST) in U/ml
trig: triglycerides in mg/dl

2 Categorical variables

asc: 0, absence of ascites; 1, presence of ascites

ede: 0 no edema; 0.5 untreated or successfully treated; 1 unsuccessfully treated edema

hep: 0, absence of hepatomegaly; 1, presence of hepatomegaly

sex: 0, male; 1, female

spid: 0, absence of spiders; 1, presence spiders

stage: histological stage of disease (needs biopsy), graded 1, 2, 3 or 4

trt: 1 for control, 2 for treatment

8.3 Additional plots

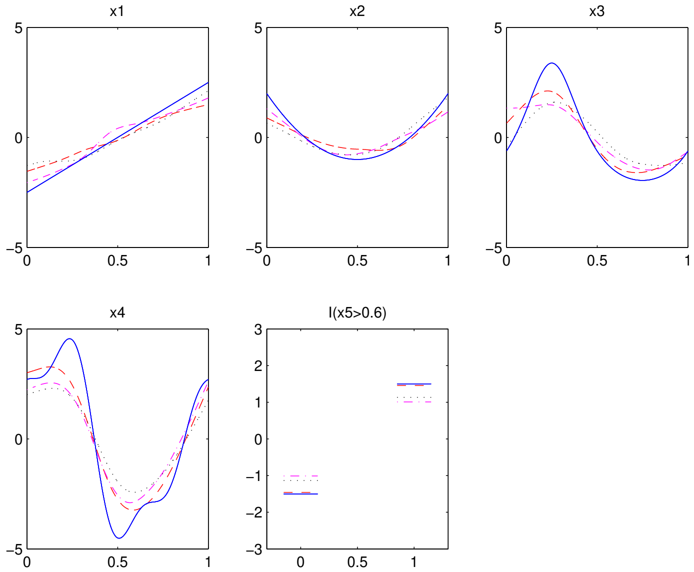

Figure 8.1: The estimated effects when $n = 100, m = 50$ and $cov(x^i, x^j) = 0$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics 6*, 701–726.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc. 68*, 337–404.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 37*, 373–384.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*, 89–99.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological 34*, 187–220.

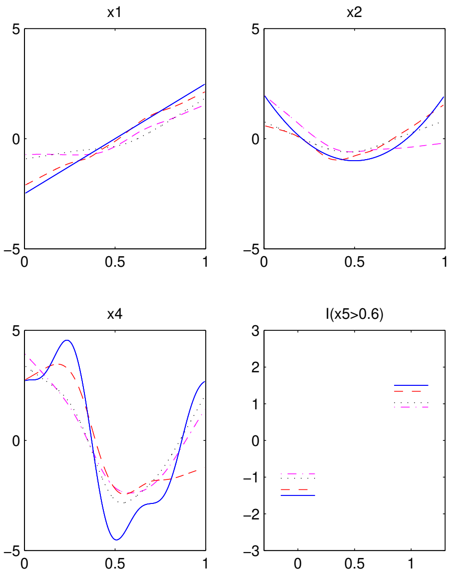

Figure 8.2: The estimated effects when $n = 100, m = 50$ and $cov(x^i, x^j) = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

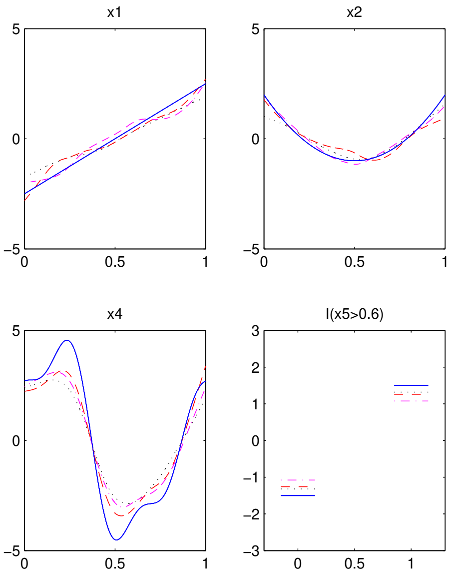

Figure 8.3: The estimated effects when $n = 200, m = 50$ and $cov(x^i, x^j) = 0$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

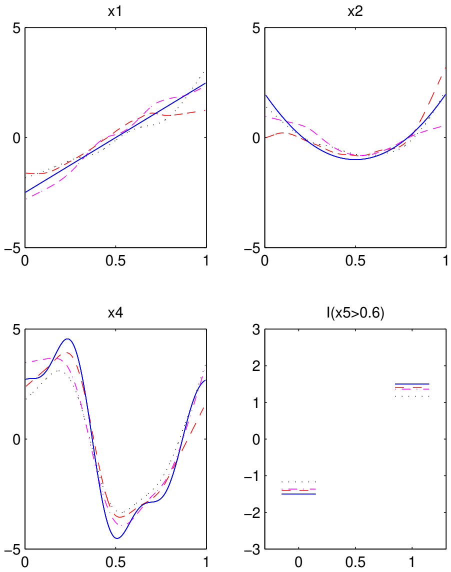

Figure 8.4: The estimated effects when $n = 200, m = 50$ and $cov(x^i, x^j) = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

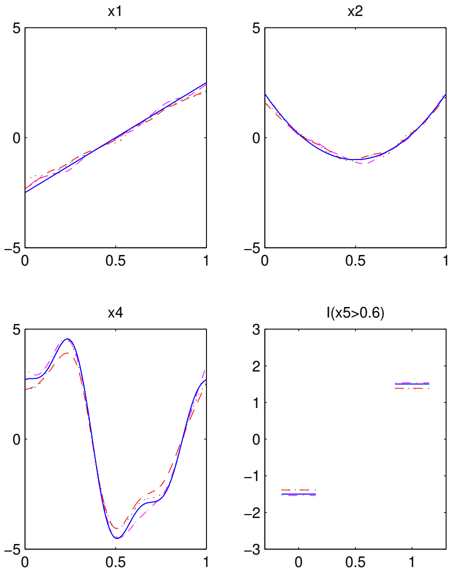

Figure 8.5: The estimated effects when $n = 800, m = 50$ and $cov(x^i, x^j) = 0$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

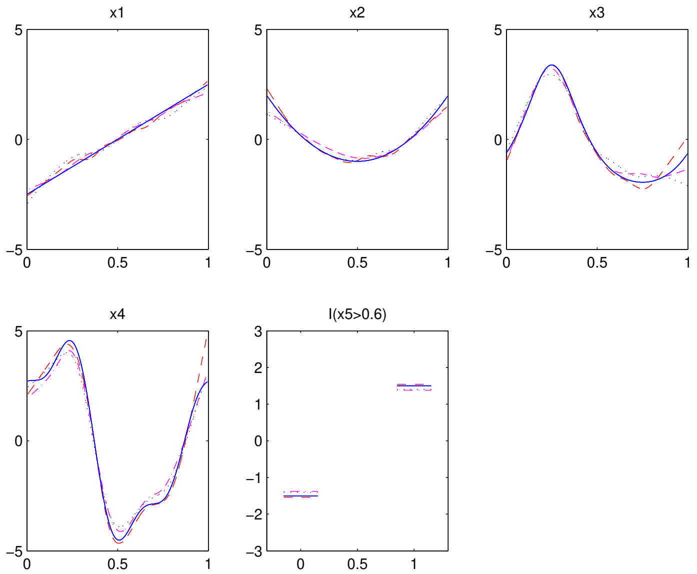

Figure 8.6: The estimated effects when $n = 800, m = 50$ and $cov(x^i, x^j) = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate 95th best; the magenta dash-dot lines indicate 50th best; the black dot lines are 5th best.

Dickson, E., P. Grambsch, T. Fleming, L. Fisher, and A. Langworthy (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology 10*, 1–7.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications (ISBN 041298321).* Chapman and Hall Ltd.

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

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics 30*(1), 74–99.

Faraggi, D. and R. Simon (1998). Bayesian variable selection method for censored survival data. *Biometrics 54*, 1475–1485.

Fleming, T. R. and D. P. Harrington (1991). *Counting processes and survival analysis.* John Wiley and Sons.

Gao, F., G. Wahba, R. Klein, and B. Klein (2001). Smoothing spline Anova for multivariate Bernoulli observations with application to ophthalmology data. *Journal of the American Statistical Association 96*(453), 127–160.

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.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics 50*, 640–652.

Green, P. J. and B. W. Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall Ltd.

Gu, C. (1996). Penalized likelihood hazard estimation: A general procedure. *Statistica Sinica 6*, 861–876.

Gu, C. (1998). Structural multivariate function estimation: Some automatic density and hazard estimates. *Statistica Sinica 8*, 317–336.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag.

Gu, C. and C. Qiu (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics 21*, 217–234.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Chapman & Hall Ltd.

Ibrahim, J. G., M.-H. Chen, and S. N. MacEachern (1999). Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics 27*, 701–717.

Kalbfleisch, J. D. and R. L. Prentice (1980). *The statistical analysis of failure time data*. John Wiley and Sons.

Kimeldorf, G. S. and G. Wahba (1971). Some results on tchebycheffian spline functions. *J. Math. Anal. Applic. 33*, 82–95.

Klein, J. P. and M. L. Moeschberger (1997). *Survival analysis: techniques for censored and truncated data*. Springer-Verlag Inc.

Lin, X., G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein (2000). Smoothing spline Anova models for large data sets with Bernoulli observations and the randomized Gacv. *The Annals of Statistics 28*(6), 1570–1600.

Lin, Y. and H. Zhang (2002). Component selection and smoothing in smoothing spline analysis of variance model. Technical Report 1072, University of Wisconsin, Madison.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics 14*, 945–966.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing [Formerly: SIAM Journal on Scientific Computing] 9*, 531–542.

O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *The Annals of Statistics 21*, 124–145.

Ruppert, D. and R. J. Carroll (2000). Spatially-adaptive penalties for spline fitting. *The Australian and New Zealand Journal of Statistics 42*(2), 205–223.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics 10*, 795–810.

Tapia, R. A. and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins University Press.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling survival data: extending the Cox model*. Springer-Verlag Inc.

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

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

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wahba, G., Y. Wang, C. Gu, R. Klein, and B. Klein (1995). Smoothing spline Anova for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *The Annals of Statistics 23*, 1865–1895.

Wand, M. and M. Jones (1995). *Kernel Smoothing*. Chapman Hall.

Wood, S., R. Kohn, T. Shively, and W. Jiang (2002). Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society, Series B, Methodological 64*(1), 119–139.

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica 6*, 675–692.

Zhang, H., G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein (2002). Variable selection and model building via likelihood basis pursuit. Technical Report 1059, University of Wisconsin, Madison.

Zucker, D. M. and A. F. Karr (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics 18*, 329–353.