Nonparametric Model Selection for Cox's Proportional Hazard model

Chenlei Leng

June 10, 2003

Abstract

We extend the COSSO to Cox's proportional hazard models.

1 Introduction

Let $T, C, L, \mathbf{x}$ be respectively survival time, censoring time, (possible) left truncation time and their associated covarivates. Observing $(Z_i, \delta_i, \mathbf{x_i}), \mathbf{i} = \mathbf{1}, ...\mathbf{n}$, where $Z_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I_{[T_i \le C_i]}$, one is to estimate the hazard rate $\eta(\mathbf{x})$.

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

$$\prod_{i=1}^{n} h(Z_i|x_i)^{\delta_i} \prod_{i=1}^{n} S(Z_i|x_i). \tag{1.1}$$

The penalized log likelihood is then

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i \eta(Z_i, x_i) - \int_{L_i}^{Z_i} e^{\eta(t,x_i)}dt\} + \lambda J(\eta), \tag{1.2}$$

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

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 $x_{(1)}, ..., x_{(N)}$. Let R_j denote the risk set right before the time t_j^0:

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

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

$$\sum_{i=1}^{N} \text{log} h_i + \sum_{i=1}^{n} \delta_i \eta(x_i) - \sum_{i=1}^{n} [\exp(\eta(x_i)) \sum_{j=1}^{N} h_j I_{j \in R_i}]. \tag{1.3}$$

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

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j \in R_i} \exp(\eta(x_j))]\} + \lambda J(\eta). \tag{1.4}$$

2 COSSO type estimate

The functional decomposition of η of d variables is

$$\eta = \eta_0 + \sum_{j=1}^{d} \eta_j(x^{(j)}) + \sum_{j<k} g_{j,k}(x^{(j)}, x^{(k)}) + ... + \eta_{1,...,d}(x^{(1)}, ..., x^{(d)}). \tag{2.1}$$

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 (2.1). 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

$$\bigotimes_{j=1}^{d} W_2^{(j)}[0,1] = \bigotimes[\{1\} \bigoplus \{W^{(j)}\}] = \{1\} \bigoplus \sum_{j=1}^{d}[\{W^{(j)}\}] \bigoplus \sum_{j<k}[\{W^{(j)}\} \bigotimes \{W^{(k)}\}] + ... \tag{2.2}$$

The tensor sum in (2.2) can be truncated using the same functional ANOVA argument. We will denote the truncated series as $\{1\} \bigoplus \mathcal{W} = \{1\} \bigoplus \sum_{\alpha=1}^{p}[\{W^\alpha\}]$. 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 $K^{(j)}(s,t)K^{(k)}(s,t)$. We will use K_α for the reproducing kernel of the term η_α in the functional decomposition of η.

The COSSO estimate minimizes

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j \in R_i} \exp(\eta(x_j))]\} + \lambda J(\eta) \text{ with } J(\eta) = \sum_{\alpha=1}^{p} \|P^\alpha \eta\|. \tag{2.3}$$

3 Categorical Variables

Following Chapter 2.2 of Gu (2002), we use $K(x,y) = I_{(x=y)} - 1/k$ as the reproducing kernel for the domain $\{1,...,k\}$.

4 Algorithm

The COSSO problem (2.3) is equivalent to

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \lambda_0\sum_{\alpha=1}^{p}\theta_\alpha^{-1}\|P^\alpha\eta\|^2\} \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \leq M, \theta_\alpha \geq 0, \alpha = 1,...,p. \tag{4.1}$$

We seek a data adaptive solution $\eta(x) = \sum_{i=1}^{N}\sum_{\alpha=1}^{p}\theta_\alpha K_\alpha(x, x_{(i)})c_i$.

Denote equation (4.1) as $A(c,\theta)$. Using the notation $K_\theta = \sum_\alpha K_\alpha$, $\{Q_{N\times N}\}_{(k,l)} = K_\theta(x_{(k)}, x_{(l)})$, $\{U_{n\times N}\}_{(k,l)} = K_\theta(x_k, x_{(l)})$, we can write (4.1) as

$$A(c,\theta) = -\frac{1}{n}\mathbf{1}^T Qc + \frac{1}{n}\sum_{i=1}^{N}\log(\sum_{j\in R_i} e^{U_j c}) + \lambda c^T Qc, \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \leq M, \theta_\alpha \geq 0, \tag{4.2}$$

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

$$\begin{aligned}\frac{\partial A}{\partial c} &= -\frac{1}{n}Q\mathbf{1} + \mathbf{\frac{1}{n}\sum_{i=1}^{N}\frac{\sum_{j\in R_i} U_j^T e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}} + 2\lambda Qc};\\ \frac{\partial^2 A}{\partial c\partial c^t} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i} U_j^T U_j e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}} - \frac{\sum_{j\in R_i} U_j^T e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}}\frac{\sum_{j\in R_i} U_j e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}}\} + 2\lambda Q.\end{aligned} \tag{4.3}$$

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 $G_{N\times p} = (\{K_1(x_{(k)}.x_{(l)})\}c, ..., \{K_p(x_{(k)}, x_{(l)})\}c)$ and $S_{n\times p} = (\{K_1(x_k.x_{(l)})\}c, ..., \{K_p(x_k, x_{(l)})\}c)$, equation (4.1) can be written as

$$-\frac{1}{n}\mathbf{1^T G}\theta + \mathbf{\frac{1}{n}\sum_{i=1}^{N}}\log(\sum_{\mathbf{j\in R_i}}\mathbf{e}^{\mathbf{S_j}\theta}) + \lambda\mathbf{c^T G}\theta \text{ s.t. } \sum_{\alpha=\mathbf{1}}^{\mathbf{p}}\theta_\alpha \leq \mathbf{M}, \theta_\alpha \geq \mathbf{0}, \tag{4.4}$$

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

$$\begin{aligned}\frac{\partial A}{\partial \theta} &= -\frac{1}{n}G^T\mathbf{1} + \mathbf{\frac{1}{n}\sum_{1}^{N}\frac{\sum_{j\in R_i} S_j^T e^(S_j\theta)}{\sum_{j\in R_i} e^(S_j\theta)}} + \lambda\mathbf{G^T c};\\ \frac{\partial^2 A}{\partial\theta\partial\theta^T} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i} S_j^T S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}} - \frac{\sum_{j\in R_i} S_j^T e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\frac{\sum_{j\in R_i} S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\}.\end{aligned} \tag{4.5}$$

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

$$\frac{1}{2}\theta^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta + [(\frac{\partial A}{\partial\theta})_{\theta_0} - (\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta_0]^T\theta \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \leq M, \theta_\alpha \geq 0, \tag{4.6}$$

5 Kullback-Leibler Loss and Leave-One-Out Cross Validation

Following Chapter 7.2 of Gu (2002), a leave-one-out cross validation eatimate of the relative Kullback-Leibler loss is

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \{\frac{\mathrm{tr}(U^T H^{-1} U)}{n(n-1)} - \frac{\mathbf{1^T U^T H^{-1} U 1}}{n^2(n-1)}\}. \tag{5.1}$$

where $H = \frac{\partial^2 A}{\partial c \partial c^t}$ at the solution.

6 Simulation

In this section, we use the examples used in Lin and Zhang (2002). Namely, the following 4 functions are used as building block of the hazrd function:

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

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

We include a function on categorical variables as

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

Example 1 We consider addtive model on R^{10} with the true hazard function $\eta(x) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4) + 3g_5(I_{(x_5>1.2u_1)})$, where $u_1 \sim$ Unif$[0,1]$. Including two more categorical variables $y_6 = I_{(x_6<0.8u_2)}$ and $y_7 = I_{(x_7>u_3)}$, where $u_2, u_3 \sim$ Unif$[0,1]$. 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 $n = 200$ data points from the hazrd function

$$h(t|y) = \exp(\eta(y)),$$

using the transformation $y_1 = x_1, ..., y_4 = x_4, y_5 = I_{(x_5>1.2u_1)}, y_6, y_7, y_8 = x_8, ..., y_{10} = x_{10}$.

The distribution of censoring time is exponentially distributed with mean $Uexp(\eta(x))$, where U is randomly generated from the uniform distribution from $[1,3]$. This is similiar to Fan and Li (2002). We do this 100 times. The result is summarized in Table 1 for the independet case and Table 2 for compound symmetry case. The estimated main effects for one of the runs when x^i's are correlated is shown in Figure 6.1 and 6.2. The estimate wrongly includes y_{10} in the model.

Variables										Model size				
1	2	3	4	5	6	7	8	9	10	5	6	7	8	9
100	100	100	100	100	2	1	4	10	5	84	13	1	1	1

Table 1: Frequency of variables selected and model sizes when x^i's are independent. 84 out of 100 models are selected correctly.

Variables										Model size					
1	2	3	4	5	6	7	8	9	10	5	6	7	8	9	10
100	100	100	100	100	6	6	13	14	22	68	18	5	6	0	3

Table 2: Frequency of variables selected and model sizes when $cov(x^i, x^j) = 0.5$. 68 out of 100 models are selected correctly.

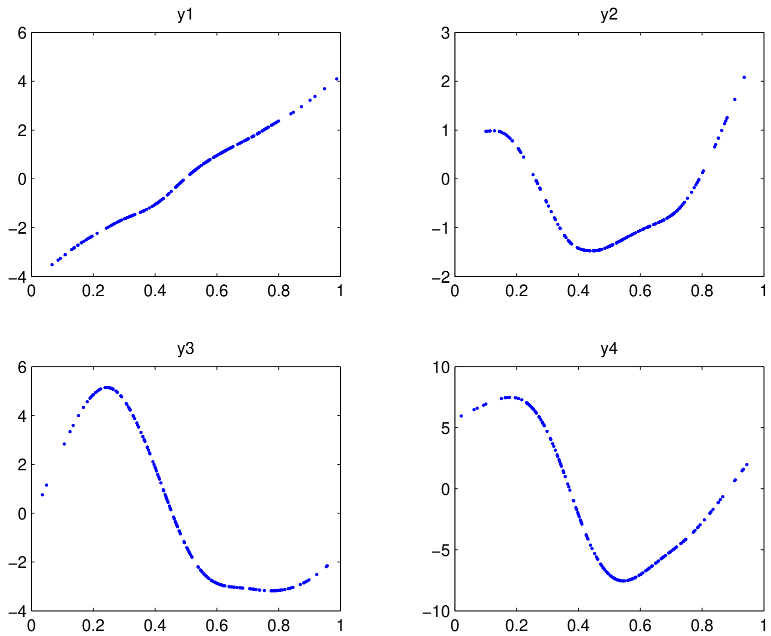

Figure 6.1: Fitted main effects for $y_1, ..., y_4$.

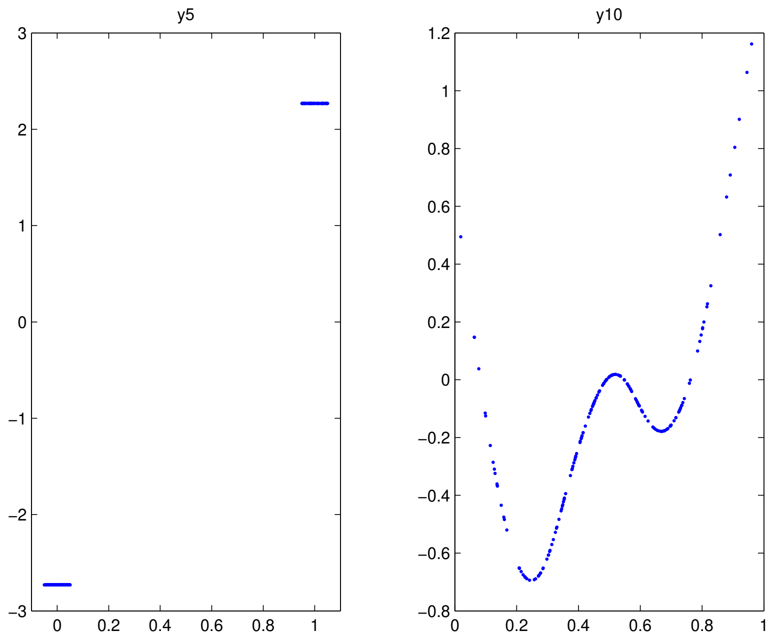

Figure 6.2: Fitted main effects for y_5 and y_{10}.

7 Example

We apply the analysis to the primary biliary cirrhosis data using the 332 patiet who agreed to randomized treatment ((Fleming and Harrington 1991),p.161). 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 ascategorical variables ascites (asc),edema (ede), hepatomegaly (hep), sex (sex) and spiders (spid). We take log transformation to alk, biliand prot since they are skewed to the right. We apply our procedure to the data. The fitted main effects are shown in Figure 7.1. 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.

References

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.

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

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

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

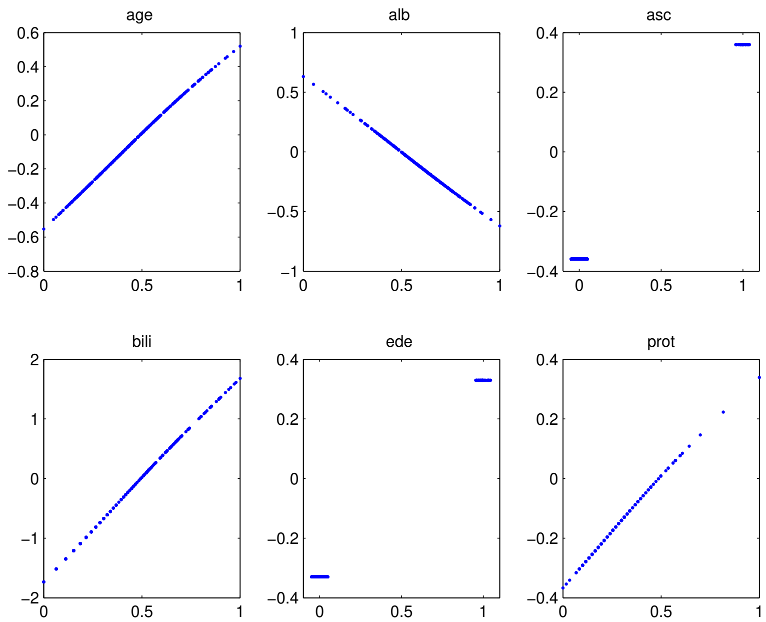

Figure 7.1: Fitted effects for PBC data.