Variable Selection in Partially Linear Cox Models

Pang Du, Shuangge Ma and Hua Liang

Department of Statistics, Virginia Tech
Department of Epidemiology and Public Health, Yale University
Dept. of Biostat. & Comp. Biology, Univ. of Rochester Medical Center

Innovation and Inventiveness in Stat & Meth.
Statistical Workshop at Yale
May 15, 2009
New Haven, CT

Virginia Tech

Cox Proportional Hazards Model

- First introduced by Cox (1972).
- Assess covariate effect in survival analysis.
- Hazard function of a subject

$$h(t|Z) = h_0(t)\exp(Z^T\beta),$$

$h_0(t)$: unknown baseline hazard,
Z: covariate vector,
β: unknown coefficient vector.

- Relative risk: $\exp(Z^T\beta)$.

Relative Risk

- Parametric form: $\exp(Z^T\beta)$.
  - References: Kalbfleisch and Prentice (2002).
  - Drawback: Linear form may not be realistic.
- Nonparametric form: $\exp(g(Z))$, g unknown (smooth) function.
  - References: Zucker and Karr (1990), O'Sullivan (1993), Fan, Gijbels, and King (1997), Huang, Kooperberg, Stone, and Truong (2000), Huang and Liu (2006).
  - Drawback: Interpretability and curse of dimentionality.
- Semiparametric form: $\exp(U^T\beta + \eta(W))$, $Z^T = (U^T, W^T)$ and η unknown (smooth) function.
  - References: Huang (1999), Cai, Fan, Jiang, and Zhou (2007), Yin, Li, and Zeng (2008).
  - Common limitation: additive nonparametric part.

Variable Selection in Cox Model

- Classical methods: AIC and BIC.
  Drawback: Unstable and not incorporating stochastic errors.
- Modified AIC or BIC: Volinsky and Raftery (2000), Liang and Zou (2008).
- LASSO: Tibshirani (1997), Zou (2008).
- SCAD: Fan and Li (2002), Cai, Fan, Li, and Zhou (2005).

Common limitation: parametric relative risk.

Our Results

A Cox PH model with

- partially linear (log) relative risk
- nonparametric part
  - estimated by smoothing spline ANOVA
  - model selection through Kullback-Leibler geometry
- parametric part
  - estimated by penalized profile partial likelihood
  - variable selection by SCAD penalty

Asymptotic properties:

- nonparametric part: optimal convergence rate.
- parametric part: $\sqrt{n}$ convergence rate, asymptotic normality, and oracle property.

Model

For each subject i, one observes

- $X_i = \min(T_i, C_i)$ and $\Delta_i = I_{T_i \le C_i}$:
  T_i survival time
  C_i *independent* right censoring time
- Covariate $Z_i^T = (U_i^T, W_i^T)$:
  - U_i: covariate in parametric part
  - W_i: covariate in nonparametric part

Hazard function for a subject:

$$h(t|Z_i) = h_0(t)\exp(U^T\boldsymbol{\beta}_0 + \eta_0(W)),$$

Log partial likelihood:

$$l_p(\boldsymbol{\beta}, \eta) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \{ U_i^T\boldsymbol{\beta} + \eta(W_i) - \log \sum_{k=1}^{n} Y_k(X_i)\exp[U_k^T\boldsymbol{\beta} + \eta(W_k)]\}$$

Doubly Penalized Partial Likelihood (I)

Given $\boldsymbol{\beta}$, η_0 is estimated as the minimizer of

$$l_{\boldsymbol{\beta}}(\eta) \equiv -l_p(\boldsymbol{\beta}, \eta) + \frac{\lambda}{2} J(\eta)$$

wrt η in a Hilbert space of functions on $\mathcal{W}$:

- Negative log partial likelihood: representing goodness-of-fit,
- $J(\eta)$: roughness penalty. e.g., $J(\eta) = \int [\eta''(w)]^2 dw$ (when w is one-dimension).
- $\lambda > 0$: smoothing parameter balancing the tradeoff.

Doubly Penalized Partial Likelihood (II)

Given an estimate $\hat{\eta}$ of η_0, coefficient vector $\boldsymbol{\beta}_0$ is estimated as the minimizer of

$$l_{\hat{\eta}}(\boldsymbol{\beta}) \equiv l_p(\boldsymbol{\beta}, \hat{\eta}) - n \sum_{j=1}^{d} p_\theta(|\beta_j|)$$

- Log partial likelihood: representing goodness-of-fit,
- $p_\theta(|\beta_j|)$: SCAD penalty (Fan and Li, 2001).

Algorithm

Steps:

1. Conventional PH model with covariate (U, W)
   $\Rightarrow \hat{\boldsymbol{\beta}}^{(0)}$: estimated coefficient vector for U.
2. Plug $\hat{\boldsymbol{\beta}}^{(0)}$ into PPL(I) and minimize to solve for η
   $\Rightarrow \hat{\eta}^{(0)}$.
3. Plug $\hat{\eta}^{(0)}$ into PPL(II) and maximize to solve for $\boldsymbol{\beta}$
   $\Rightarrow \hat{\boldsymbol{\beta}}^{(1)}$.
4. Replace $\hat{\boldsymbol{\beta}}^{(0)}$ in Step 2 by $\hat{\boldsymbol{\beta}}^{(1)}$ and repeat Steps 2 and 3 until convergence to obtain the final estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}$.

Smoothing Spline ANOVA

ANOVA structure for η for bivariate W:

$$\eta(w_1, w_2) = \eta_\emptyset + \eta_1(w_1) + \eta_2(w_2) + \eta_{1,2}(w_1, w_2)$$

- Identifiability: side conditions for main effects η_1 and η_2, interaction $\eta_{1,2}$, e.g., $\int \eta_1 dw_1 = 0$, $\int \eta_{w_2} dw_2 = 0$
- $\eta_{1,2} = 0 \Rightarrow$ additive η.
- Easy extension to multivariate W.

Model Selection Tool

Goal: e.g., test on interaction effect
Reduced: $\eta(w_1, w_2) = \eta_\emptyset + \eta_1(w_1) + \eta_2(w_2)$
vs. Complete:
$\eta(w_1, w_2) = \eta_\emptyset + \eta_1(w_1) + \eta_2(w_2) + \eta_{1,2}(w_1, w_2)$

Model selection tool generalizing (Gu 2004):

- η_c: estimate under constant model
  $\hat{\eta}$: estimate under complete model
  $\tilde{\eta}$: KL-projection of $\hat{\eta}$ to space of reduced model

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\hat{\eta}, \eta_c)$$

- $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) < 0.05$
  $\Rightarrow$ Complete model not necessary

Asymptotic Theory

Under certain conditions,

- $\|\hat{\eta} - \eta_0\|_2 = O_p(n^{-m/(2m+1)})$, m is order of the relevant Sobolev space of functions.
- $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.
- With probability approaching one,
  - $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$
  - For $V_0(\pi_1)$, Σ_θ and $\mathbf{b}$ defined in the paper,

$$\sqrt{n}(V_0(\pi_1) + \Sigma_\theta)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (V_0(\pi_1) + \Sigma_\theta)^{-1}\mathbf{b}\} \rightarrow N(\mathbf{0}, V_0(\pi_1)).$$

Standard Errors and Smoothing Parameter Selection

Standard errors:

- η: Bayes model for smoothing splines.
- $\boldsymbol{\beta}$: asymptotic variance in the theorem.

Smoothing parameter selection:

- λ: a cross-validation score estimating Kullback-Leibler distance.
- θ: AIC.

Empirical Studies: Data

Common settings in all simulations:

- Exponential hazard: $h(t|U, W) = \exp[U^T \boldsymbol{\beta}_0 + \eta_0(W)]$.
- U: MVN with zero mean and $\text{Cov}(U_j, U_k) = 0.5^{|j-k|}$, $1 \leq j, k \leq 8$.
- $\boldsymbol{\beta}_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$

Variable Selection: Parametric Component

Two settings, each with 1000 data replicates.

- $\eta_{0a}(w) = 1.5\sin(2\pi w - \frac{\pi}{2})$ or
  $\eta_{0b}(w) = 4(w-0.3)^2 + 4.7e^{-w} - 3.4643.$
- Exponential censoring time: censoring rate 23% for η_{0a} and 40% for η_{0b}
- Sample sizes: $n = 150$ or 500.

Variable Selection: Parametric Component

Methods in comparison:

(A) Complete oracle: contributing covariates (U_1, U_4, U_7, W) and form of η_0 are known;

(B) Partial oracle I: contributing covariates (U_1, U_4, U_7, W) known, η_0 unknown and assumed to be linear;

(C) Partial oracle II: contributing covariates (U_1, U_4, U_7, W) known, form of η_0 is unknown and estimated by PPL;

(D) Proposed method with SCAD penalty;

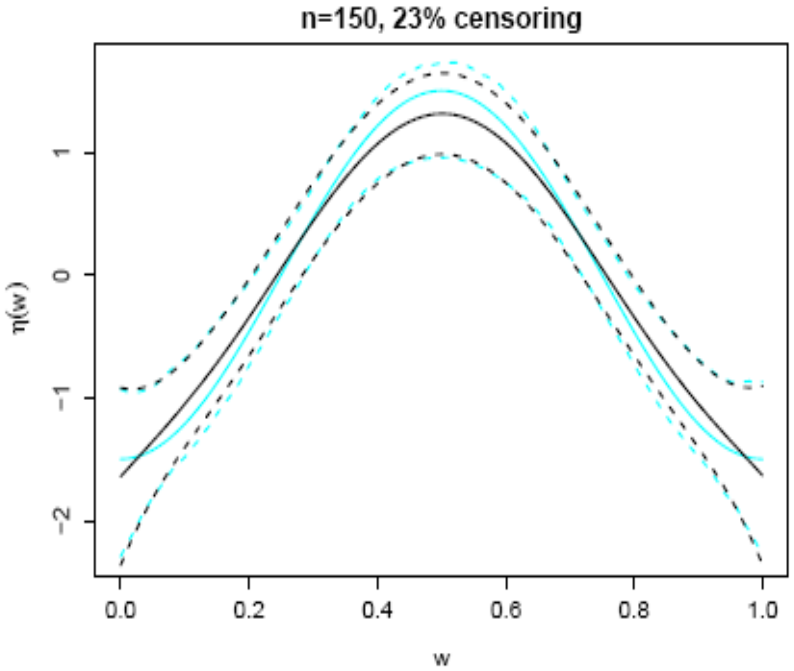
(E) Proposed method with adaptive LASSO penalty.

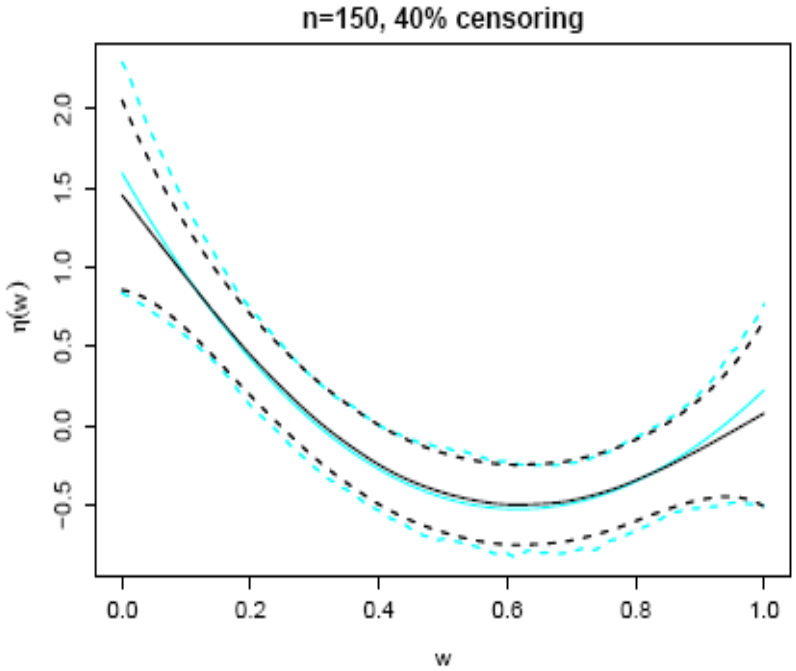
Table 1: Variable Selection for Parametric Component.

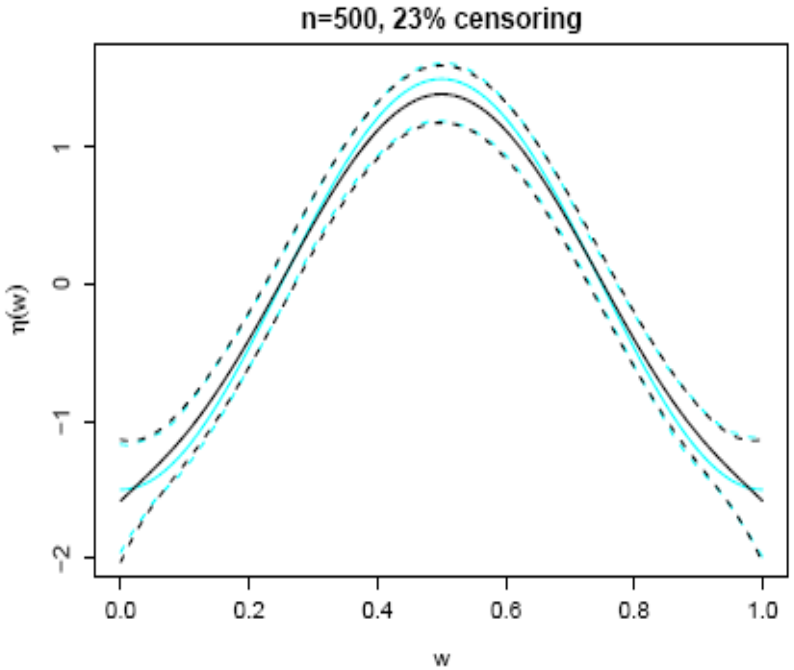
Method	MRME	No. of Non-zeros C	No. of Non-zeros IC	Proportion of Under-fit	Proportion of Correct-fit	Proportion of Over-fit
		$n = 150$, $\eta_0 = \eta_{0a}$ (23% censoring)				
B	0.168	-	-	-	-	-
C	0.475	-	-	-	-	-
D	0.409	2.998	0.825	0.002	0.476	0.522
E	0.387	2.998	0.959	0.002	0.444	0.554
		$n = 150$, $\eta_0 = \eta_{0b}$ (40% censoring)				
B	0.167	-	-	-	-	-
C	0.711	-	-	-	-	-
D	0.518	2.996	0.949	0.004	0.430	0.566
E	0.563	2.998	1.131	0.002	0.378	0.620
		$n = 500$, $\eta_0 = \eta_{0a}$ (23% censoring)				
B	0.056	-	-	-	-	-
C	0.431	-	-	-	-	-
D	0.396	3.000	0.717	0.000	0.525	0.475
E	0.375	3.000	0.736	0.000	0.540	0.460
		$n = 500$, $\eta_0 = \eta_{0b}$ (40% censoring)				
B	0.057	-	-	-	-	-
C	0.712	-	-	-	-	-
D	0.619	3.000	0.749	0.000	0.512	0.488
E	0.628	3.000	0.776	0.000	0.529	0.471

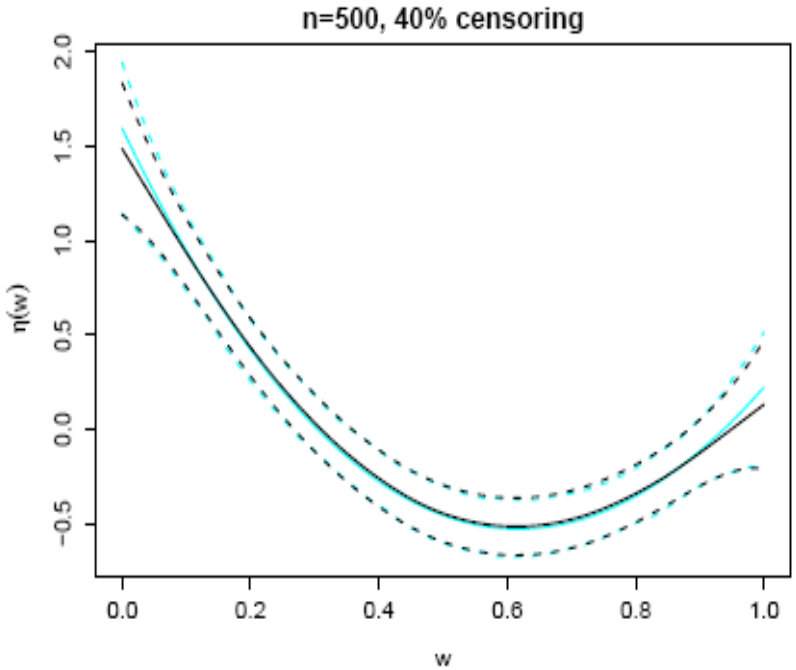
Table 2: Standard Deviations for The SCAD Estimate of $\boldsymbol{\beta}$.

	$\hat{\boldsymbol{\beta}}_1$		$\hat{\boldsymbol{\beta}}_4$		$\hat{\boldsymbol{\beta}}_7$	
n, censor%	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
150, 23%	0.124	0.113(0.015)	0.141	0.121(0.017)	0.135	0.109(0.015)
150, 40%	0.159	0.135(0.017)	0.188	0.145(0.021)	0.155	0.128(0.019)
500, 23%	0.065	0.059(0.005)	0.073	0.063(0.005)	0.062	0.057(0.005)
500, 40%	0.075	0.070(0.006)	0.088	0.076(0.006)	0.078	0.066(0.006)









Virginia Tech

Variable Selection: Nonparametric Component

Four settings, each with 1000 data replicates.

- True model: W_1.
  Fitted model: $W_1 + W_2$.
  Reduced models: W_1 and W_2.
  - $\eta_0(w_1, w_2) = \eta_{0a}(w_1)$ or
    $\eta_0(w_1, w_2) = \eta_{0b}(w_1)$.
  - Exponential censoring time: censoring rate 23% for η_{0a} and 40% for η_{0b}
  - Sample sizes: $n = 150$ or 500.

Variable Selection: Nonparametric Component

- True model: $W_1 + W_2$.
  Fitted model: $W_1 * W_2$.
  Reduced models: W_1, W_2 and $W_1 + W_2$.
  - $\eta_0(w_1, w_2) = 0.7\eta_{0a}(w_1) + 0.3\eta_{0a}(w_2)$ or $\eta_0(w_1, w_2) = \eta_{0a}(w_1) + \eta_{0a}(w_2)$.
  - Uniform censoring time: respective censoring rates 25% and 39%
  - Sample sizes: $n = 150$ or 300.

Table 3: Variable Selection for Nonparametric Component.

Sample Size	Proportion of Selecting W_1	Proportion of Selecting W_2	Proportion of Selecting $W_1 : W_2$	Proportion of Under-fit	Proportion of Correct-fit	Proportion of Over-fit
	True model: $\eta_0(w_1, w_2) = \eta_{0a}(w_1)$, 23% censoring					
$n = 150$	1.000	0.512	-	0.000	0.488	0.512
$n = 500$	1.000	0.001	-	0.000	0.999	0.001
	True model: $\eta_0(w_1, w_2) = \eta_{0b}(w_1)$, 40% censoring					
$n = 150$	1.000	0.101	-	0.000	0.899	0.101
$n = 500$	1.000	0.257	-	0.000	0.743	0.257
	True model: $\eta_0(w_1, w_2) = 0.7\eta_{0a}(w_1) + 0.3\eta_{0b}(w_2)$, 25% censoring					
$n = 150$	1.000	0.738	0.286	0.262	0.451	0.287
$n = 300$	1.000	0.613	0.152	0.387	0.461	0.152
	True model: $\eta_0(w_1, w_2) = \eta_{0a}(w_1) + \eta_{0b}(w_2)$, 39% censoring					
$n = 150$	1.000	0.987	0.272	0.013	0.715	0.272
$n = 300$	1.000	0.999	0.132	0.001	0.867	0.132

Sexually Transmitted Diseases Data

An STD study in Klein and Moeschberger (1997)

- 877 individuals with an initial diagnosis of gonorrhea or chlamydia were followed for reinfection.
- Covariates
  - $W_j, j = 1, 2$: age, years of schooling
  - $U_k, 1 \leq k \leq 22$: 5 demographic variables, 7 sexual behavior variables, 10 symptom variables.
- Goal: identify factors that are related to time until reinfection given an initial infection.

STD Data: Model

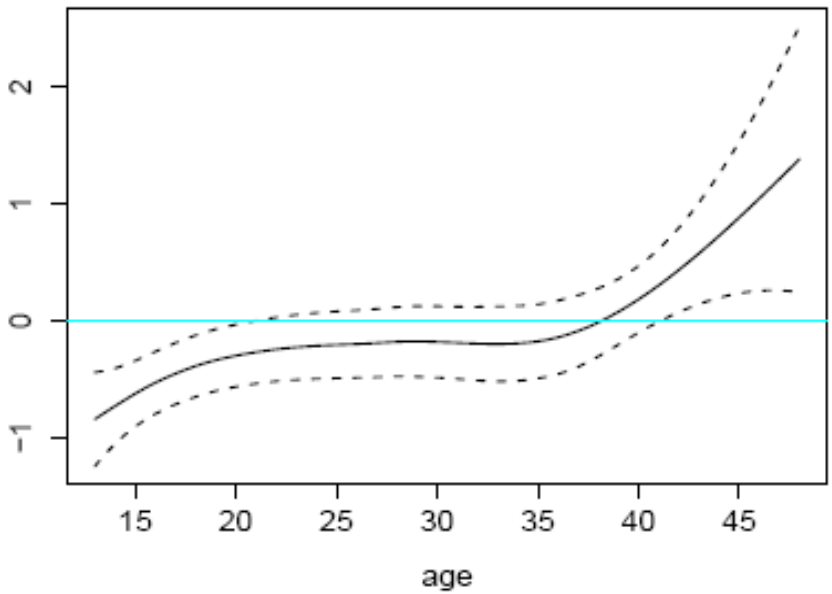
Initial model:

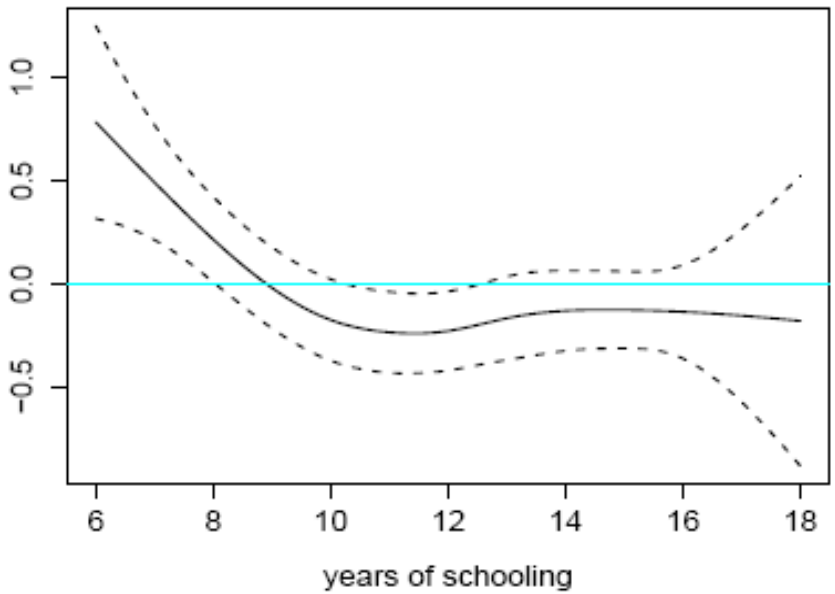
$$h_i(t|Z) = h_0(t) \exp\left\{ \sum_{j=1}^{3} \eta_j(W_{ji}) + \sum_{k=1}^{22} U_{ki}\beta_k \right\},$$

$\eta_3(W_{3i}) = \eta_3(W_{1i}, W_{2i})$: interaction term between W_1 and W_2.
Model selection tool for η
$\Rightarrow$ negligible interaction between W_1 and W_2.

Table 4: Fitted Coefficients and Their Standard Errors for Sexually Transmitted Diseases Data.

npart	raceW	maritalM	maritalS	typeC	typeB
0(-)	0(-)	0(-)	0.487(0.212)	-0.412(0.149)	-0.337(0.144)
oralY	oralM	rectY	rectM	abdom	disc
-0.336(0.201)	-0.341(0.235)	0(-)	0(-)	0.253(0.151)	0(-)
dysu	condS	condN	itch	lesion	rash
0.193(0.152)	0(-)	-0.327(0.114)	0(-)	0(-)	0(-)
lymph	involve	discE	node		
0(-)	0.423(0.166)	-0.460(0.220)	0(-)		


2
1
0
−1
15
20
25
30
35
40
45
age


1.0
0.5
0.0
−0.5
6
8
10
12
14
16
18
years of schooling

Summary and Remarks

- Cox PH model with partially linear relative risk
  - Doubly penalized partial likelihood
  - nonparametric part modeled and variables selected by smoothing spline ANOVA
  - Variable selection in parametric part by SCAD penalty
  - Asymptotic properties
  - Standard errors and smoothing parameter selection
- Easy extension to time-dependent covariate.