Variable Selection for Linear Transformation Models

Hao Helen Zhang

Department of Statistics
North Carolina State University
hzhang@stat.ncsu.edu

Fields Institute, June 10, 2011

Table of contents

Background

In the context of censoring data for survival analysis,

- T is the failure time, or the time to event (e.g. death, relapse, cancer)
- C is the censoring time
- Z is the covariates or predictors
- Observe $\tilde{T} = \min(T, C)$ and the censoring indicator $\delta = I(T \leq C)$. The observations $(\tilde{T}_i, \delta_i, Z_i)$, $i = 1, \cdots, n$.

Example: Lymphoma dataset (Rosenwald et al. 2002)

- $n = 240$ diffuse large B-cell lymphoma (DLBCL) patients, and $p = 7,399$ genes for each patient.
- Patients' survival times were recorded, 138 patients died during the follow-up method.

Semi-parametric Survival Models

Widely-used survival models:

- Cox's proportional hazards (PH) model (Cox, 1972):

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_0' Z)$$

- Proportional odds (PO) model (Pettitt, 1982, 1984; Bennett, 1983):

$$\{1 - S(t|Z)\}/S(t|Z) = [\{1 - S_0(t)\}/S_0(t)]\exp(\beta_0' Z)$$

- Linear transformation (LT) models (Clayton and Cuzick, 1985; Cheng, Wei and Ying, 1995):

$$H_0(T) = -\beta_0' Z + \epsilon,$$

H_0 is an unknown increasing function, ϵ has a known continuous distribution and independent of $\mathbf{Z}$.

Variable selection problems for censored data

- Write the regression coefficients $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})'$.
- Index set for important variables: $I = \{1 \leq j \leq p : \beta_{j0} \neq 0\}$
- Index set for unimportant variables: $U = \{1 \leq j \leq p : \beta_{j0} = 0\}$
- Assume $|I| = p_0 < p$. Write $\beta_0 = (\beta'_{I0}, \mathbf{0}')'$.

The main goals of a variable selection procedure are:

- to identify I and U correctly;
- to provide good estimators for β_{I0}.

Oracle properties

An ideal variable selection procedure should asymptotically satisfy:

- produce parsimonious models automatically (with probability one)

$$\hat{\beta}_j \neq 0 \quad \text{for} \quad j \in I$$

$$\hat{\beta}_j = 0 \quad \text{for} \quad j \in U;$$

- achieve the optimal estimation rate

$$\sqrt{n}(\hat{\beta}_I - \beta_{I0}) \rightarrow_d N(0, \Sigma_{I0}),$$

  where Σ_{I0} is the covariance matrix knowing the true model.

Oracle procedure performs as well as if the correct true model were known.

Existing variable selection methods for censored data

- Best subset selection and stepwise selection
- Asymptotic testing procedures, such as score test and Wald test
- Bootstrap sampling procedures (Sauerbrei and Schumacher 1992)
- Bayesian variable selection (Faraggi and Simon 1998; Ibrahim, Chen and MacEachern 1999)
- Shrinkage methods (LASSO: Tibshirani 1997; SCAD: Fan and Li 2002; Adaptive-LASSO: Zhang and Lu 2007)

Penalized partial likelihood estimation for Cox's model

- Log partial likelihood (Cox 1975):

$$l_n(\beta) \equiv \sum_{i=1}^{n} \delta_i \{\beta' Z_i - \log[\sum_{j=1}^{n} I(\tilde{T}_j \geq \tilde{T}_i) \exp(\beta' Z_j)]\}.$$

- The penalized log partial likelihood estimation

$$\min_{\beta} -\frac{1}{n} l_n(\beta) + \sum_{j=1}^{p} J_\lambda(\beta_j).$$

Choices of penalty function

- Ridge regression (Hoerl and Kennard, 1970): $J_\lambda(\beta_j) = \lambda\beta_j^2$.
- Bridge regression (Frank and Friedman, 1993): $J_\lambda(\beta_j) = \lambda|\beta_j|^q, \quad q \geq 0$.
  - If $q = 0$, known as entropy penalty (Donoho and Johnstone, 1998).
  - If $q = 1$, known as LASSO (Tibshirani, 1996).
  - For $q \leq 1$, it tends to shrink small $|\beta|$'s to exactly zero.
  - J_λ is not convex for $q < 1$ while solutions are not sparse for $q > 1$.

Other examples: SCAD, adaptive LASSO

Adaptive LASSO estimation for Cox's model

We solve (Zhang and Lu, 2007)

$$\min_{\beta} -\frac{1}{n} l_n(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| w_j,$$

where $\mathbf{w} = (w_1, \cdots, w_p)'$ are the data-dependent weights.
Key Motivations:

- Large penalties are imposed on unimportant covariate effects, while small penalties for important ones. (Protect important covariates more)
- Let the data choose w_j's adaptively.

Extension to PO model

What if there is no partial likelihood available?

- For the PO model, Lu and Zhang (2007) considered the marginal likelihood.
- The marginal likelihood generally does not have a closed form, but it can be calculated using an importance sampling technique.
- We proposed to use the penalized marginal likelihood estimation:

$$\min_{\beta} -\frac{1}{n} l_{n,M}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| w_j,$$

where $l_{n,M}(\beta)$ is log marginal likelihood function.

Linear Transformation Models

Linear transformation (LT) models form a rich class of models due to the flexibility of H_0.

$$H_0(T) = -\beta_0' Z + \epsilon.$$

They include PH and PO models as special cases

- if ϵ follows extreme value distribution, LT reduces to PH models
- if ϵ follows the logistic distribution, LT reduces to PO models
- if ϵ follows the standard normal distribution, LT generalizes the usual Box-Cox transformation models.

Advantages and Challenges with LT

A unified estimation framework for survival data.

- can relax the independence assumption between the covariates and the censoring variable (needed for the validity of PH model).
- reduce to partial likelihood under PH models.

Variable selection for LT models is less studied in literature.

- Most estimation procedures for LT models are based on estimating equations (e.g., Chen et al., 1995; Fine et al., 1998; Chen et al., 2002).
- Challenge 1: a convenient loss function is not available for LT models.
- Challenge 2: involves a nonparametric component H.

Our proposal: construct a sensible loss function first!

Martingale based estimating equations

- For subject i, define the counting process $N_i(t) = \delta_i I(\tilde{T}_i \leq t)$ and at-risk process $Y_i(t) = I(\tilde{T}_i \geq t)$.
- Mean-zero Martingale process: $M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda\{H_0(s) + \beta_0' Z_i\}$, where $\Lambda(\cdot)$ is the known cumulative hazard function of ϵ.
- Martingale-based Estimating equations (Chen et al., 2002):

$$\sum_{i=1}^{n} [dN_i(t) - Y_i(t) d\Lambda\{\beta' Z_i + H(t)\}] = 0, \quad t \geq 0. \qquad (1)$$

$$\sum_{i=1}^{n} \int_0^{\tau} Z_i [dN_i(t) - Y_i(t) d\Lambda\{\beta' Z_i + H(t)\}] = \mathbf{0}, \qquad (2)$$

A joint estimation of parametric and nonparametric terms.

Computation

Let $0 < t_1 < \cdots < t_K < \infty$ be the observed K failure times in the data.

- *Step* 1. Set $\beta = \hat{\beta}^{(0)}$. Compute $\hat{H}^{(0)}$ as follows. First solve

$$\sum_{i=1}^{n} Y_i(t_1)\Lambda\{H(t_1) + \beta' Z_i\} = 1,$$

for $\hat{H}^{(0)}(t_1)$. Then solve sequentially

$$\sum_{i=1}^{n} Y_i(t_k)[\Lambda\{H(t_k) + \beta' Z_i\} - \Lambda\{\hat{H}^{(0)}(t_k-) + \beta' Z_i\}] = 1,$$

for $\hat{H}^{(0)}(t_k)$, where $k = 2, \cdots, K$.

Computation algorithm

- *Step* 2. Solve equation

$$\sum_{i=1}^{n} Z_i[\delta_i - \Lambda\{\hat{H}^{(0)}(\tilde{T}_i) + \beta' Z_i\}] = 0.$$

  for $\hat{\beta}^{(1)}$.
- *Step* 3. Set $\beta = \hat{\beta}^{(1)}$ and repeat Steps 1 and 2 until prescribed convergence criteria are met.

Profiled score functions

- Given β, let $\tilde{H}(\cdot;\beta)$ denote the solution of (1).
- Plugging $\tilde{H}$ into equation (2) and define the profiled score functions of β

$$U_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} Z_i[dN_i(t) - Y_i(t)d\Lambda\{\beta' Z_i + \tilde{H}(t;\beta)\}]. \quad (3)$$

- Let $\tilde{\beta}$ denote the solution of $U_n(\beta) = 0$.

Asymptotic properties of $\tilde{\beta}$

- We have
  - (i) $n^{-1/2}U_n(\beta_0) \to N(0, V)$ in distribution, as $n \to \infty$.
  - (ii) $\sqrt{n}(\tilde{\beta} - \beta_0) \to N(0, \Sigma)$ in distribution with $\Sigma = A^{-1}V(A^{-1})'$, as $n \to \infty$.
  - (iii) The asymptotic variance-covariance matrix Σ can be consistently estimated by $\hat{\Sigma}_n = \hat{A}_n^{-1}\hat{V}_n(\hat{A}_n^{-1})'$ using the usual plugging method.
- See Chen et al. (2002) for the details.

Penalized profiled score squares

- We first define a weighted quadratic loss function as:

$$D_n(\beta) = (1/n) U_n'(\beta) \hat{V}_n^{-1} U_n(\beta).$$

Then propose to minimize

$$Q_n(\beta) = D_n(\beta) + \lambda_n \sum_{j=1}^{p} J(|\beta_j|). \tag{4}$$

- We use the adaptive Lasso penalty, using the weight $w_j = 1/|\tilde{\beta}_j|$.
- The PPS^2 estimator is defined as $\hat{\beta}_n = \text{argmin}_\beta Q_n(\beta)$.

Computation of PPS^2 estimators

- Consider the Taylor expansion of $U_n(\beta)$ around $\hat{\beta}^{[0]}$,

$$U_n(\beta) \approx U_n(\hat{\beta}^{[0]}) + n\hat{A}_n\{\hat{\beta}^{[0]}, \tilde{H}(\cdot; \hat{\beta}^{[0]})\}(\beta - \hat{\beta}^{[0]}),$$

- $Q_n(\beta)$ can approximated by a quadratic form

$$n(\hat{\beta}^{[0]} + \mathbf{b} - \beta)' \hat{A}_n^{[0]'} \hat{V}_n^{-1} \hat{A}_n^{[0]} (\hat{\beta}^{[0]} + \mathbf{b} - \beta) + \lambda_n \sum_{j=1}^{p} w_j |\beta_j|, \quad (5)$$

where $\hat{A}_n^{[0]'} = \hat{A}_n\{\hat{\beta}^{[0]}, \tilde{H}(\cdot; \hat{\beta}^{[0]})\}$ and
$\mathbf{b} = (\hat{A}_n^{[0]'} \hat{V}_n^{-1} \hat{A}_n^{[0]})^{-1} \hat{A}_n^{[0]'} \hat{V}_n^{-1} U_n(\hat{\beta}^{[0]})/n.$

Computational algorithm

- Step 0: Compute the full model estimator: $\tilde{\beta}$ and $\tilde{H}(\cdot) = \tilde{H}(\cdot; \tilde{\beta})$.
- Step 1: Choose an initial estimator $\hat{\beta}^{[0]}$. Set $w_j = 1/|\tilde{\beta}_j|$ for all j.
- Step 2: Solve equation (1) to obtain $\tilde{H}(\cdot; \hat{\beta}^{[0]})$.
- Step 3: Minimize (5) and denote the shrinkage estimate as $\hat{\beta}^{[1]}$.
- Step 4: Set $\hat{\beta}^{[0]} = \hat{\beta}^{[1]}$ and repeat Steps 2 and 3 until convergence.

Let $\hat{\beta}_n$ denote the resulting sparse estimator.

One-step estimator

- full iteration: computationally intensive
- one-step estimator, i.e. choosing $\hat{\beta}^{[0]} = \tilde{\beta}$. Note that $U_n(\tilde{\beta}) = 0$. Then (5) becomes

$$n(\tilde{\beta} - \beta)'\hat{\Sigma}_n^{-1}(\tilde{\beta} - \beta) + \lambda_n \sum_{j=1}^{p} w_j|\beta_j|. \tag{6}$$

- The minimization of (6) can be solved in R with the *lars* package (Efron et al., 2004). The entire solution path of the resulting PPS^2 estimator can be also obtained.

Theoretical properties of PPS^2 estimators

Write the solution $\hat{\beta}_n = (\hat{\beta}'_{I,n}, \hat{\beta}'_{U,n})'$. In addition, write the limiting covariance matrix Σ accordingly.

- Theorem 1 (root-n Consistency). If $\sqrt{n}\lambda_n = O_p(1)$, then $\|\hat{\beta}_n - \beta_0\| = O_p(n^{-1/2})$.
- Theorem 2 (Sparsity and normality). Assume that $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$, then
  (i) (Sparsity) $\hat{\beta}_{U,n} = \mathbf{0}$ with probability tending to one;
  (ii) (Asymptotic normality)
  $\sqrt{n}(\hat{\beta}_{I,n} - \beta_{I0}) \to N\{0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\}$ as n goes to infinity.

Efficiency and tuning

- The efficiency of the PPS^2 estimator $\hat{\beta}_{I,n}$ for nonzero components is better than that of the corresponding full model estimator $\tilde{\beta}_I$ since

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} < \Sigma_{11}.$$

- We use BIC for selecting λ, i.e. $\mathrm{BIC}_\lambda = (\hat{\beta}_\lambda - \tilde{\beta})'\hat{\Sigma}_n^{-1}(\hat{\beta}_\lambda - \tilde{\beta}) + \log n \cdot \mathrm{df}_\lambda / n$. Here df_λ is the number of nonzero coefficients in $\hat{\beta}_\lambda$, a simple estimate for the degree of freedom (Zou et al. 2007).

Variance estimation

Two approaches to estimate the variance of the estimator:

- estimates based on the asymptotic theory (Theorem 2)
- the sandwich variance estimation: For any nonzero β_j, we can approximate its weighted L_1 penalty with a local quadratic function

$$\frac{|\beta_j|}{|\tilde{\beta}_j|} \approx \frac{\beta_j^2}{|\tilde{\beta}_j||\beta_j|}.$$

Details can found in Zhang, Lu and Wang (2010).

Simulation studies for LT models

- We consider both the PH and PO models.
- We choose $\beta = (-1, -0.9, 0, 0, 0, -0.8, 0, 0, 0)'$, and the nine covariates $Z = (Z_1, ..., Z_9)$ are marginally standard normal with the pairwise correlation $\text{corr}(Z_j, Z_k) = \rho^{|j-k|}$ with $\rho = 0.5$.
- Censoring times are from uniform (0,c): 25% and 40% censoring rates
- Sample sizes n = 100, 200, simulation replications $M = 500$.
- We compare the PPS^2, PPL (Zhang and Lu, 2007), PML (Lu and Zhang, 2007) estimates.

Simulation results for PH model

Table 1. Mean squared error and model selection results

n	Censored	Method	Average MSE	Model Size	Number of zero coefficients	
				oracle (3)	correct (6)	incorrect (0)
100	25%	EE	0.244 (0.161)	9	0 (0)	0 (0)
		PPS^2	0.122 (0.119)	3.610 (0.920)	5.390 (0.920)	0.000 (0.000)
		PPL	0.130 (0.121)	3.136 (0.412)	5.858 (0.403)	0.006 (0.077)
	40%	EE	0.277 (0.186)	9	0 (0)	0 (0)
		PPS^2	0.143 (0.133)	3.620 (0.885)	5.380 (0.885)	0.000 (0.000)
		PPL	0.177 (0.161)	3.150 (0.456)	5.836 (0.435)	0.014 (0.118)
200	25%	EE	0.087 (0.052)	9	0 (0)	0 (0)
		PPS^2	0.051 (0.040)	3.250 (0.557)	5.750 (0.557)	0.000 (0.000)
		PPL	0.053 (0.050)	3.034 (0.181)	5.966 (0.181)	0.000 (0.000)
	40%	EE	0.110 (0.066)	9	0 (0)	0 (0)
		PPS^2	0.063 (0.049)	3.280 (0.604)	5.720 (0.604)	0.000 (0.000)
		PPL	0.062 (0.055)	3.048 (0.214)	5.952 (0.214)	0.000 (0.000)

Simulation results for PO model

Table 2. Mean squared error and model selection results

n	Censored	Method	Average MSE	Model Size	Number of zero coefficients	
				oracle (3)	correct (6)	incorrect (0)
100	25%	EE	0.481 (0.262)	9	0 (0)	0 (0)
		PPS^2	0.377 (0.303)	3.600 (0.932)	5.230 (0.874)	0.170 (0.403)
		PML	0.436 (0.419)	2.898 (0.684)	5.856 (0.389)	0.246 (0.539)
	40%	EE	0.575 (0.347)	9	0 (0)	0 (0)
		PPS^2	0.385 (0.314)	3.490 (0.916)	5.360 (0.811)	0.150 (0.386)
		PML	0.493 (0.484)	2.834 (0.735)	5.844 (0.400)	0.322 (0.599)
200	25%	EE	0.213 (0.109)	9	0 (0)	0 (0)
		PPS^2	0.122 (0.085)	3.340 (0.670)	5.660 (0.670)	0.000 (0.000)
		PML	0.231 (0.120)	3.026 (0.193)	5.968 (0.176)	0.006 (0.077)
	40%	EE	0.258 (0.168)	9	0 (0)	0 (0)
		PPS^2	0.132 (0.086)	3.310 (0.598)	5.690 (0.598)	0.000 (0.000)
		PML	0.218 (0.142)	3.030 (0.239)	5.952 (0.214)	0.018 (0.133)

Variance estimation results

Table 3. Estimated standard errors for the PPS^2 nonzero estimates ($n = 200$).

		$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_6$		
Model	Censoring	SE	$\widehat{\text{SE}}$	$\widehat{\text{SE}}_S$	SE	$\widehat{\text{SE}}$	$\widehat{\text{SE}}_S$	SE	$\widehat{\text{SE}}$	$\widehat{\text{SE}}_S$
PH	25%	0.113	0.109	0.105	0.121	0.105	0.100	0.110	0.092	0.088
	40%	0.126	0.120	0.114	0.135	0.116	0.109	0.122	0.103	0.097
PO	25%	0.187	0.165	0.152	0.211	0.164	0.147	0.165	0.146	0.131
	40%	0.196	0.176	0.161	0.225	0.177	0.156	0.187	0.155	0.138

SE: sample standard deviation of the estimates; $\widehat{\text{SE}}$: the average of estimated standard error based on theory; $\widehat{\text{SE}}_S$: the average of estimated standard error based on the sandwich formula.

Primary biliary cirrhosis data

- Data gathered in the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984 (Therneau and Grambsch 2000).
- 312 eligible subjects with 125 deaths
- 17 predictors: 10 continuous and 7 discrete.
- Goal: to study the dependence of survival times on 17 covariates.
- Zhang and Lu (2007) studied variable selection for this data in the PH model using the penalized partial likelihood method with the adaptive Lasso penalty.

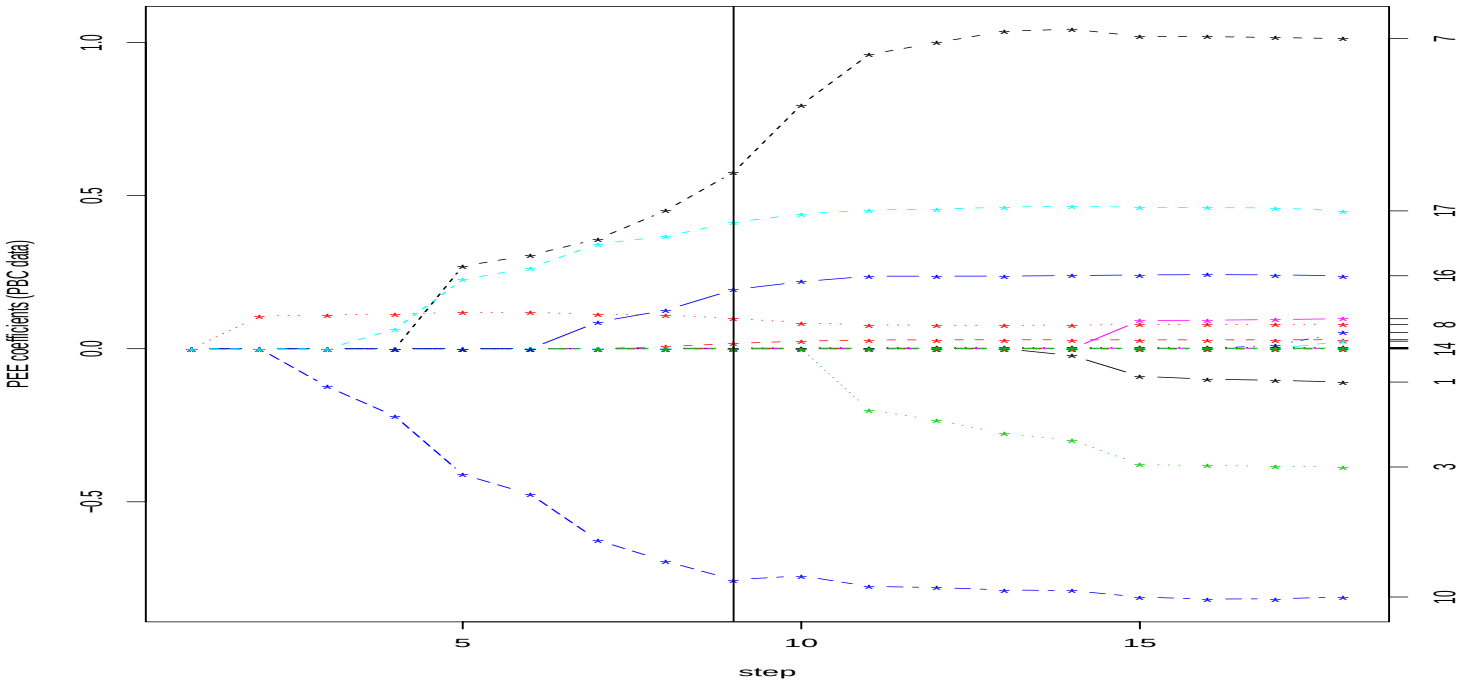
Analysis of PBC data

Table 4. Estimation and variable selection for PBC data with the PH model.

Covariate	EE	PPS^2	PPL
trt	-0.109 (0.234)	0 (-)	0 (-)
age	0.029 (0.012)	0.017 (0.007)	0.019 (0.010)
sex	-0.386 (0.346)	0 (-)	0 (-)
asc	0.053 (0.469)	0 (0)	0 (-)
hep	0.024 (0.263)	0 (-)	0 (-)
spid	0.098 (0.279)	0 (-)	0 (-)
oed	1.013 (0.486)	0.576 (0.241)	0.671 (0.377)
bil	0.079 (0.024)	0.099 (0.018)	0.095 (0.020)
chol	0.001 (0.000)	0 (-)	0 (-)
alb	-0.811 (0.286)	-0.755 (0.211)	-0.612 (0.280)
cop	0.003 (0.001)	0.003 (0.001)	0.002 (0.001)
alk	0.000 (0.000)	0 (-)	0 (-)
sgot	0.004 (0.002)	0.002 (0.001)	0.002 (0.001)
trig	-0.001 (0.001)	0 (-)	0 (-)
plat	0.001 (0.001)	0 (-)	0 (-)
prot	0.238 (0.103)	0.193 (0.066)	0.103 (0.108)
stage	0.450 (0.171)	0.413 (0.121)	0.367 (0.142)

Solution path for the PPS^2 estimates

- For PBC data using PH model



Lung cancer data

- Data is from the Veteran's Administration lung cancer trial (Kalbfleish and Prentice 2002).
- 137 males with advanced inoperable lung cancer were randomized to either a standard treatment or chemotherapy
- There are six covariates: Treatment (1=standard, 2=test), Cell type (1=squamous, 2=small cell, 3=adeno, 4=large), Karnofsky score, Months from Diagnosis, Age, and Prior therapy (0=no, 10=yes).
- Lu and Zhang (2007) studied variable selection for this data in the PO model using the penalized marginal likelihood method with the adaptive Lasso penalty.

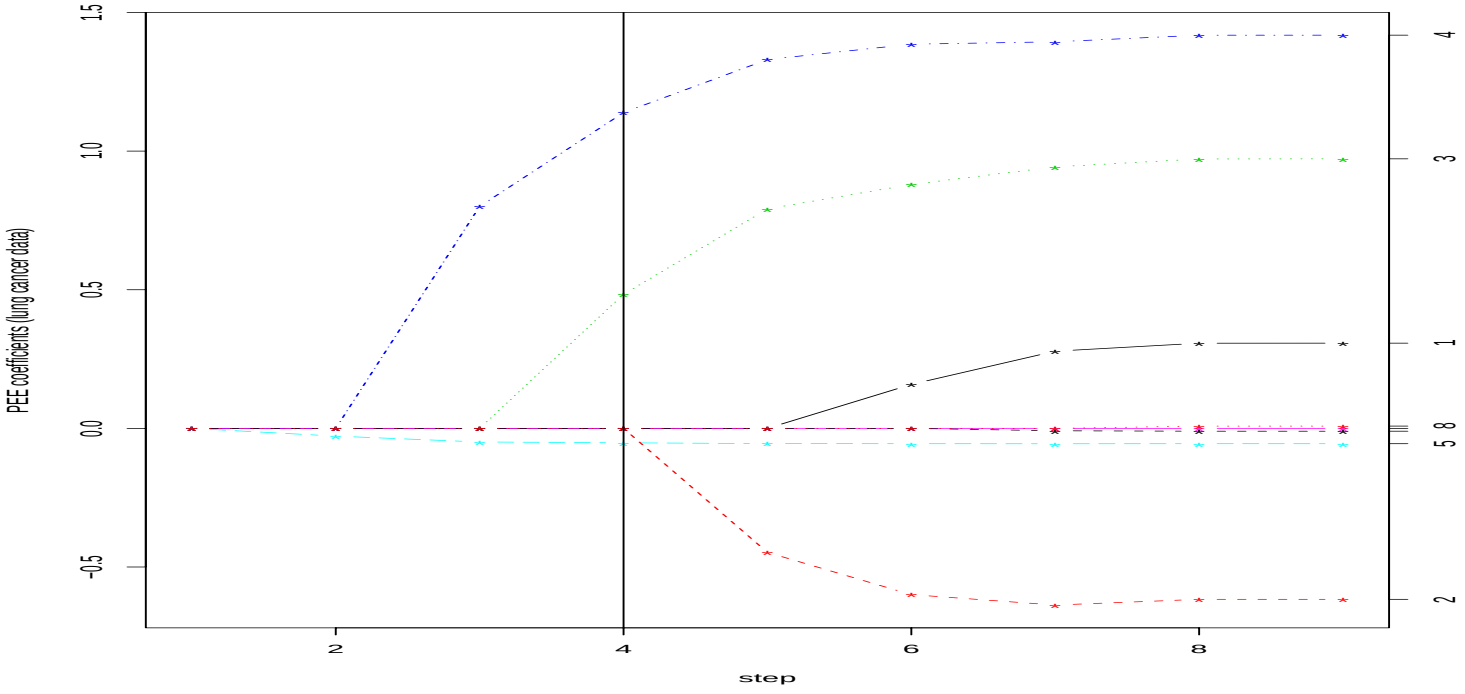
Analysis of lung cancer data

Table 5. Estimation and variable selection results for lung cancer data with the PO model.

Covariate	EE	PPS^2	PML
Treatment	0.307 (0.317)	0 (-)	0 (-)
squamous vs large	-0.617 (0.482)	0 (-)	0 (-)
small vs large	0.972 (0.473)	0.483 (0.197)	0.706 (0.356)
adeno vs large	1.418 (0.371)	1.139 (0.261)	0.841 (0.397)
Karnofsky	-0.055 (0.009)	-0.052 (0.008)	-0.053 (0.008)
Months from Diagnosis	0.000 (0.015)	0 (-)	0 (-)
Age	-0.010 (0.017)	0 (-)	0 (-)
Prior therapy	0.008 (0.040)	0 (-)	0 (-)

Solution path for the PPS^2 estimates

- For lung cancer data using PO model



Microarray Data (DLBCL) Analysis

About the dataset (Rosenwald et al. 2002)

- $n = 240$ diffuse large B-cell lymphoma (DLBCL) patients, and $p = 7,399$ genes for each patient.
- Patients' survival times were recorded, 138 patients died during the follow-up method.
- a common practice is to first conduct a preliminary gene filtering based on some univariate analysis. We choose the top 50 genes selected using univariate Cox score.

Results:

- randomly divide the data set into two sets: 160 for training and the remaining 80 for testing.
- The PEE selects totally 20 genes and the PPL selects 13 genes, and they share 9.

Summary and Discussions

- a general class of survival models in a unified framework with desired theoretical properties
- the profiled score takes care of the nonparametric component in a natural fashion
- can improve efficiency over the original estimator from the estimation equations.

Future work:

- extensions to general methods of estimation equations (on-going work)

References and acknowledgements

- References:
  - Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's Proportional Hazards Model. Biometrika, 94, 1-13.
  - Lu, W. and Zhang, H. H. (2007). Variable Selection for Proportional Odds Model. Statistics in Medicine, 26, 3771-3781.
  - Zhang, H. H., Lu, W. and Wang, H. (2010) On sparse estimation for semiparametric linear transformation models. Journal of Multivariate Analysis, 101, 1594-1606.
- Acknowledgements:
  - Collaborators: Wenbin Lu (NCSU), Hansheng Wang (Beijing University)
  - The research was partially supported by NSF and NIH grants.