DEPARTMENT OF STATISTICS

North Carolina State University

2501 Founders Drive, Campus Box 8203

Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2604

# Penalized Asymptotic Likelihood Approach for Linear Transformation Model Selection

Hao Helen Zhang

Department of Statistics, North Carolina State University, Raleigh, NC, 27695.

Wenbin Lu

Department of Statistics, North Carolina State University, Raleigh, NC, 27695.

Hansheng Wang

Guanghua School of Management, Peking University, Beijing, China, 100871.

hzhang@stat.ncsu.edu, lu@stat.ncsu.edu, hansheng.wang@vip.163.com

Supported in part by National Science Foundation grants DMS-0405913 and DMS-0504269.

# Penalized Asymptotic Likelihood Approach for Linear Transformation Model Selection


## Summary

We propose a new and general approach, the penalized asymptotic likelihood (PAL) maximization, to joint parameter estimation and variable selection for models where a likelihood or a convenient loss function is not readily available. The main idea is to first construct a "likelihood" function based on an existing estimator and its asymptotic distribution function, and then maximize the constructed likelihood subject to some shrinkage penalty for a sparse estimation. In this paper, we focus on the general class of linear transformation models for survival data analysis, which include the Cox's proportional odds models and proportional hazards models as special cases. We present how to construct a penalized asymptotic likelihood for linear transformation models, based on the estimation equation procedure developed by Chen et al. (2002). The new procedure leads to an estimator with improved efficiency and sparse representation. Theoretical properties of the PAL estimators are established, including the root-$n$ consistency, selection consistency, and asymptotic normality. Furthermore, the new procedure enjoys easy implementation and its entire solution path can be obtained with a modified LARS algorithm. The performance of the PAL estimator is illustrated by simulated examples and the analysis of Veteran's Administration lung cancer data.



*Some key words*: Censored survival data; Variable selection; Linear transformation models; Asymptotic likelihood; Adaptive LASSO.

## 1. Introduction

One main issue in survival analysis is to study the dependence of the survival time $T$ of patients on covariates $\mathbf{Z} = (Z_1, \cdots, Z_d)$. Though the proportional hazards model (Cox, 1972) has been in wide use for survival data analysis, it may not be an appropriate choice for some types of survival data. For example, if the hazard functions for two treatment groups converge to the same limit, i.e., the homogeneity between different groups increases with time, then the proportional odds model is naturally a better choice (Pettitt, 1982, 1984; Bennett, 1983; Dabrowska and Doksum, 1988; Murphy et al., 1997). Recently, a general class of semiparametric linear transformation models have been proposed and extensively studied (Clayton and Cuzick, 1985; Bickel et al., 1993; Cheng et al., 1995; Fine et al., 1998). Let $T, C$ and $\mathbf{Z}$ denote respectively the survival time, the censoring time and the $p \times 1$ covariate vector. A linear transformation model

$$H(T) = -\boldsymbol{\beta}'\mathbf{Z} + \epsilon, \tag{1·1}$$

where $H$ is an unknown monotone increasing function, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$ is the regression parameter vector, and $\epsilon$ has a completely known continuous distribution that is independent of $C$ and $\mathbf{Z}$. We use $\boldsymbol{\beta}_0$ to denote the vector of true regression coefficients. Let $\Lambda(x)$ be the cumulative hazard function of $\epsilon$, so that $P(\epsilon > x) = \exp\{-\Lambda(x)\}$. If $\epsilon$ follows the extreme value distribution, i.e. $\Lambda(x) = \exp(x)$, then (1·1) becomes the proportional hazards model. If $\epsilon$ has the standard logistic distribution, i.e. $\Lambda(x) = \log\{1 + \exp(x)\}$, then (1·1) is equivalent to the proportional odds model. When there is no censoring and $\epsilon$ follows the standard normal distribution, (1·1)

generalizes the usual Box-Cox transformation models.

In this paper, we are interested in the sparse modeling problem when the true model (1·1) has a sparse representation, i.e. some components of $\boldsymbol{\beta}_0$ are exactly zero. Our goal is to discover the important index set $A = \{j : \beta_{0j} \neq 0, j = 1, \cdots, d\}$ and estimate the corresponding coefficients. This task is also known as the variable selection problem. Effective variable selection is fundamental to survival data analysis, and it generally leads to better health risk assessment and model interpretation. Various variable selection methods have been proposed for ordinary linear models, including traditional procedures like the best subset selection and stepwise selection procedures, and modern shrinkage procedures such as the non-negative garrote (Breiman 1995), the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001), and the adaptive LASSO (Zou 2006).

The nature of data has made the model estimation and variable selection more difficult for survival data analysis. Classical methods like stepwise selection procedures can be expensive in computation and suffer from high variability. Recently some shrinkage methods are proposed for Cox's proportional hazards model based on the partial likelihood, including the LASSO (Tibshirani 1997), the SCAD (Fan and Li, 2002) and the adaptive LASSO (Zhang and Lu, 2007). However, very little work has been done for variable selection in linear transformational models. The main difficulty in linear transformation model selection is due to the lack of partial likelihood functions or some convenient loss functions. Most estimation procedures for liner transformation models are based on estimating equations (Chen et al., 1995; Fine et

al., 1998; Chen et al., 2002), which are not convenient to incorporate the shrinkage penalty like in the case of Cox models.

In this paper, we develop a new method called the penalized asymptotic likelihood (PAL) maximization for linear transformation model selection. The main idea is to construct a "likelihood" function based on an existing estimator and its asymptotic distribution function, and then maximize the constructed likelihood subject to some shrinkage penalty. This new approach provides a general framework for conducting parameter estimation and variable selection simultaneously for any model where a likelihood or a convenient loss function is not readily available. We illustrate how this approach is implemented for linear transformation models based on the estimation equation procedure developed by Chen et al. (2002). Furthermore, we show the resulting estimator is root-$n$ consistent, asymptotically normal, and can identify the true important set $A$ correctly with the probability tending to one.

The remainder of this article is organized as follows. Section 2 reviews the estimation equation method proposed by Chen et al. (2002), and then introduces our new estimation procedure for linear transformation models. In Section 3, we study the theoretical properties of the proposed estimator, and discuss the issue of parameter tuning. We also derive the sandwich-type formula to estimate the standard errors of the estimates. Section 4 is devoted to simulation studies and one application to a real data set. Some final remarks are given in Section 6.

## 2. New Estimation for Linear Transformation Models

Assume the failure time $T$ is distributed as (1·1). In the presence of censoring, we observe the event time $\tilde{T}_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \leq C_i)$, where $C_i$ is the censoring time of subject $i$ and $I(\cdot)$ is the indicator function. Suppose a random sample of $n$ individuals is chosen, then the observations consist of $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. We further assume that the censoring variable $C$ is independent of $T$ given $Z$. Without loss of generality, we assume that $Z$'s are standardized such that $\sum_{i=1}^{n} Z_{ij}^2 = 1$.

We use $0 < t_1 < \cdots < t_k < \infty$ to denote the observed distinct failure times. Following the usual counting process, let $N_i(t) = \delta_i I(\tilde{T}_i \leq t)$ and $Y_i(t) = I(\tilde{T}_i \geq t)$ respectively denote the counting and at-risk processes of the $i$th subject. In addition, define

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda\{H_0(s) + \boldsymbol{\beta}_0' Z_i\}, \quad i = 1, \cdots, n, \tag{2·1}$$

where $\Lambda(\cdot)$ is the known cumulative hazard function of $\epsilon$ and $(\boldsymbol{\beta}_0, H_0)$ are the true values of $(\boldsymbol{\beta}, H, f)$. It is easy to show that $M_i(t)$ is a martingale process by the usual counting process and the associated martingale theory (Fleming and Harrington 1991; Andersen, Borgan, Gill and Keiding 1993). Chen et al.(2002) proposed a general estimation procedure by solving the following estimations:

$$\begin{aligned} \sum_{i=1}^{n} \int_0^{\infty} Z_i [dN_i(t) - Y_i(t) d\Lambda\{\boldsymbol{\beta}' Z_i + H(t)\} &= 0, \\ \sum_{i=1}^{n} [dN_i(t) - Y_i(t) d\Lambda\{\boldsymbol{\beta}' Z_i + H(t)\} &= 0 \quad (t \geq 0), \end{aligned} \tag{2·2}$$

where $H$ is an increasing function with $H(0) = 0$. For the special case of the Cox

model, (2·2) are equivalent to the Cox partial likelihood score equation. Let $(\widetilde{\boldsymbol{\beta}}, \widetilde{H})$ be the solution of (2·2). Chen et al. (2002) showed that, under suitable regularity conditions, the estimator $\widetilde{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal, i.e.

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \longrightarrow_d N\{0, \tilde{\Sigma}\}, \quad \text{as} \ \ n \to \infty,$$

where $\tilde{\Sigma} = \Sigma_*^{-1}\Sigma^*(\Sigma_*^{-1})'$, where $\Sigma^*$ and $\Sigma_*$ are defined in Chen et al. (2002). Chen et al. (2002) also suggested the following consistent estimators for $\Sigma_*$ and $\Sigma^*$,

$$\begin{aligned}
\widehat{\Sigma}^* &= n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\{Z_i - \bar{Z}(t)\}\{Z_i - \bar{Z}(t)\}'\lambda\{\widetilde{\boldsymbol{\beta}}' Z_i + \tilde{H}(t)\}Y_i(t)]d\tilde{H}(t), \\
\widehat{\Sigma}_* &= n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\{Z_i - \bar{Z}(t)\}Z_i'\dot{\lambda}\{\widetilde{\boldsymbol{\beta}}' Z_i + \tilde{H}(t)\}Y_i(t)]d\tilde{H}(t),
\end{aligned}$$

where $\bar{Z}(t)$ is given in Chen et al. (2002).

As we pointed out, it has been a challenge to conduct variable selection based on the estimation equation (2·2), since there is not a convenient loss function available to incorporate the shrinkage penalty like in the case of Cox models. Here we propose to first construct a likelihood function for $\widetilde{\boldsymbol{\beta}}$ based on its asymptotic distribution:

$$L_A(\widetilde{\boldsymbol{\beta}}) = const. + \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})'\widehat{\Sigma}^{-1}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})\}.$$

For variable selection purpose, a penalized asymptotic likelihood (PAL) estimator is obtained by minimizing

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})'\widehat{\Sigma}^{-1}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \lambda\sum_{j=1}^{d} w_j|\beta_j|, \tag{2·3}$$

where $\widehat{\Sigma} = n^{-1}\widehat{\Sigma}_*^{-1}\widehat{\Sigma}^*(\widehat{\Sigma}_*^{-1})'$, the weights $w_j$'s are pre-selected non-negative constants, and $\lambda > 0$ is the tuning parameter. We would like to point out, though the form (2·3)

is same as the LSA estimator proposed in Wang and Leng (2007), there are subtle differences between these two approaches. Firstly, the LSA approach starts with a loss function and the estimator is derived based on a second-order Taylor expansion of the loss function. But the PAL estimator does not require a loss function to start with. Secondly, if the asymptotic distribution of the initial estimator $\widetilde{\boldsymbol{\beta}}$ is not normal, these two approaches will result in different forms. Their identical expression in this context is mainly due to the asymptotic normality of the estimator of Chen et al. (2002).

This adaptive LASSO penalty has been studied in the context of linear models (Zou, 2006), LAD regression models (Wang et al., 2007), Cox' proportional hazard models (Zhang and Lu, 2007), regression with autoregressive errors (Wang et al., 2007) and proportional odds model (Lu and Zhang, 2007). These $w_j$'s can be regarded as leverage factors to adjust penalties on individual regression coefficients, taking large values for unimportant covariates and small values for important covariates. The choice of $w_j$'s is essential and their appropriate values will guarantee the optimality of the estimator. In this paper, we use $w_j = 1/|\tilde{\beta}_j|$. This weight has been used in other contexts, such as Zou (2006) and Zhang and Lu (2007), and shows very effective performance. So the penalized asymptotic likelihood (PAL) estimator is proposed as:

$$\widehat{\boldsymbol{\beta}}_n = \text{argmin}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})'\widehat{\Sigma}^{-1}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \lambda \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|. \qquad (2{\cdot}4)$$

## 3. Theoretical Properties for the New Estimator

In this section, we study the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ in (2·4). Without loss of generality, we assume that the true important index set $I = \{1, \cdots, d_0\}$, where $d_0$ is an integer and $0 \leq d_0 \leq d$. Therefore we have $\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0}) = \{(\boldsymbol{\beta}_0^{(1)})', (\boldsymbol{\beta}_0^{(2)})'\}'$. Correspondingly, we write $\widehat{\boldsymbol{\beta}}_n = (\hat{\beta}_{1n}, \cdots, \hat{\beta}_{dn}) = \{(\widehat{\boldsymbol{\beta}}_n^{(1)})', (\widehat{\boldsymbol{\beta}}_n^{(2)})'\}'$. Correspondingly, we decompose the asymptotic covariance matrix $\widetilde{\Sigma}$ into the following block matrix form

$$\widetilde{\Sigma} = \begin{bmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{bmatrix},$$

where $\widetilde{\Sigma}_{11}$ is the first $d_0 \times d_0$ submatrix of $\widetilde{\Sigma}$.

In the following, we will establish some relevant asymptotic theories through three theorems. Theorem 1 says as long as $\lambda$ goes to zero faster than $n^{-1/2}$, the estimator $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent. Theorem 2 says, with probability tending to one, all the zeros coefficients must be estimated as zero. These two theorems ensure that the new procedure can identify the true model consistently. Furthermore, Theorem 3 shows that the new estimator performs is asymptotically normal and performs as well as the estimator of Chen et. al. (2002) for estimating $\boldsymbol{\beta}_0^{(1)}$ when assuming we knew $\boldsymbol{\beta}_0^{(2)} = \mathbf{0}$.

Theorem 1 *($\sqrt{n}$-Consistency) Assume that $(Z_1, T_1, C_1), ..., (Z_n, T_n, C_n)$ are independently and identically distributed, and $T_i$ and $C_i$ are independent given $Z_i$. If $\sqrt{n}\lambda = O_p(1)$, then $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.*

Theorem 2 *(Selection-Consistency) Assume that $(Z_1, T_1, C_1), ..., (Z_n, T_n, C_n)$ are in-*

*dependently and identically distributed, and $T_i$ and $C_i$ are independent given $Z_i$. If $\sqrt{n}\lambda = O_p(1)$ and $n\lambda \to \infty$, then $P(\widehat{\boldsymbol{\beta}}_n^{(2)} = \mathbf{0}) \to 1$.*

THEOREM 3 *(Asymptotic Normality) Assume that $(Z_1, T_1, C_1), ..., (Z_n, T_n, C_n)$ are independently and identically distributed, and $T_i$ and $C_i$ are independent given $Z_i$. If $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(1)} - \boldsymbol{\beta}_0^{(1)}) \to N(\mathbf{0}, \widetilde{\Sigma}_{11} - \widetilde{\Sigma}_{12}\widetilde{\Sigma}_{22}^{-1}\widetilde{\Sigma}_{21})$$

*as $n$ goes to infinity.*

**Remark 1**. It is easy to see that the efficiency of the PAL estimator is better than that of the full model estimator obtained from the estimation equation. This is because the asymptotic covariance of $\hat{\boldsymbol{\beta}}_n^{(1)}$ is "smaller" than that of the of the unpenalized estimator $\widetilde{\boldsymbol{\beta}}_{(1)}$.

The proofs of Theorem 1 and Theorem 3 are given in the Appendix. Proof of Theorem 2 is very similar to that in Wang and Leng (2007), so it will be omitted. We would like to point out that, since the proofs only require the root-$n$ consistency of $\tilde{\boldsymbol{\beta}}$, any root-$n$ consistent estimates of $\boldsymbol{\beta}_0$ can be used for the adaptive weights $w$'s without changing the asymptotic properties of the PAL estimator.

## 4. Variance Estimation and Parameter Tuning

The finite sample covariance of the PAL estimator is derived in the section. At the convergence, let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$, where $\widehat{\boldsymbol{\beta}}_1$ consists of all the nonzero coefficients.

Correspondingly, we partition the covariance matrix $\widehat{\Sigma}^{-1}$ as

$$\widehat{\Sigma}^{-1} = \widehat{\Omega} = \begin{bmatrix} \widehat{\Omega}_{11} & \widehat{\Omega}_{12} \\ \widehat{\Omega}_{21} & \widehat{\Omega}_{22} \end{bmatrix}.$$

In the following, we suggest two possible ways of estimating the covariance of the nonzero estimates $\widehat{\boldsymbol{\beta}}_n^{(1)}$.

The first covariance estimator of $\widehat{\boldsymbol{\beta}}_n^{(1)}$ is based on its asymptotic normality property given in Theorem 3,

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}_1) = \widehat{\Sigma}_{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21} = \widehat{\Omega}_{11}. \tag{4·1}$$

Secondly, we will derive a sandwich formula to approximate the covariance of the nonzero PAL estimate. Tibshirani (1996) proposed a standard error formula of the LASSO. Fan and Li (2001) suggested that the local quadratic approximation (LQA) can be used to derive a sandwich formula for computing the covariance of the nonzero estimates, and the consistency of the covariance estimate is proved by Fan and Peng (2004). In the next, we will apply the LQA approach to the PAL estimation procedure. For any nonzero $\beta_j$, we can approximate the $L_1$ penalty with a local quadratic function

$$\frac{|\beta_j|}{|\tilde{\beta}_j|} = \frac{1}{2}\frac{|\hat{\beta}_j^{[k]}|}{|\tilde{\beta}_j|} + \frac{1}{2}\frac{\beta_j^2}{|\tilde{\beta}_j||\hat{\beta}_j^{[k]}|},$$

where $\hat{\beta}_j^{[k]}$ is the estimate at a previous step. Then the PAL estimates can be obtained by iteratively solving a ridge regression problem:

$$(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1)'\widehat{\Omega}_{11}(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1) - 2(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1)'\widehat{\Omega}_{12}\widetilde{\boldsymbol{\beta}}_2 + \lambda\boldsymbol{\beta}_1' D(\widehat{\boldsymbol{\beta}}_1^{[k]})\boldsymbol{\beta}_1, \tag{4·2}$$

where $D(\boldsymbol{\beta}) = \text{diag}\{\frac{1}{2|\tilde{\beta}_j||\beta_j|}\}$. Therefore, the PAL estimates can be approximated by

$$\widehat{\boldsymbol{\beta}}_1 = \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1^{[k]})\right]^{-1} (\widehat{\Omega}_{11}\widetilde{\boldsymbol{\beta}}_1 + \widehat{\Omega}_{12}\widetilde{\boldsymbol{\beta}}_2). \tag{4·3}$$

This leads to a sandwich formula for estimating the covariance of nonzero coefficients

$$\begin{aligned}
\widehat{\text{Cov}}_S(\widehat{\boldsymbol{\beta}}_1) &= \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1} \widehat{\text{Cov}}\left(\widehat{\Omega}_{11}\widetilde{\boldsymbol{\beta}}_1 + \widehat{\Omega}_{12}\widetilde{\boldsymbol{\beta}}_2\right) \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1} \\
&= \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1} \left(\widehat{\Omega}_{11}\widehat{\Sigma}_{11}\widehat{\Omega}_{11} + 2\widehat{\Omega}_{11}\widehat{\Sigma}_{12}\widehat{\Omega}_{21} + \widehat{\Omega}_{12}\widehat{\Sigma}_{22}\widehat{\Omega}_{21}\right) + \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1} \\
&= \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1} \widehat{\Omega}_{11} \left[\widehat{\Omega}_{11} + \lambda D(\widehat{\boldsymbol{\beta}}_1)\right]^{-1}, \qquad (4·4)
\end{aligned}$$

after some algebra calculation.

**Remark 2**. In theory, the optimal tuning parameter $\lambda$ in (4·4) goes to zero very fast, therefore the covariance estimator based on the sandwich formula is asymptotically equivalent to the first estimator (4·1). For finite samples, the second estimator is typically smaller than the first estimator due to the non-vanishing term $\lambda D(\widehat{\boldsymbol{\beta}}_1)$, which is also confirmed in our numerical studies (shown in Section 5). Therefore, we recommend the use of the first estimator in practice.

To tune the parameter $\lambda$, we will use the BIC and AIC selection criteria proposed by Wang et al. (2007).

$$\text{BIC}_\lambda = (\widehat{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}})'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}) + \log n \cdot \text{df}_\lambda/n,$$

and

$$\text{AIC}_\lambda = (\widehat{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}})'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}) + 2 \cdot \text{df}_\lambda/n,$$

where $\text{df}_\lambda$ is the number of nonzero coefficients in $\widehat{\boldsymbol{\beta}}_\lambda$, a simple estimate for the degree of freedom (Zou et al. 2004). Wang and Leng (2007) proved that the BIC criterion is

consistent for their LSA estimator, i.e. the optimal $\lambda$ chosen by the BIC can identify the true model with probability tending to one. Similarly, we can show the BIC criterion for the PAL estimator is also consistent.

## 5. Numerical Studies

### 5·1. *Simulation Study*

Both the proportional hazards (PH) and proportional odds (PO) models are considered in our numerical study. For each example, we compare our new estimators with the estimating equation method (EE) of Chen et al. (2002), with regard to their overall mean squared error (MSE), point estimation accuracy, and the variable selection performance. Following Tibshirani (1997), we compute the MSE $= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\Sigma_X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and report the average MSE over 100 simulations for each method. Here $\Sigma_X$ is the population covariance matrix of the covariates. In term of variable selection performance, we compare the average numbers of correct and incorrect zero coefficients selected by each method. Both BIC and AIC are used to choose the tuning parameter $\lambda$. The numbers in parentheses are the standard errors. We further show the performance of our sandwich formula for covariance estimation.

The base design involves nine covariates $(Z_1, ..., Z_9)$, which are all marginally standard normal with the pairwise correlation $\text{corr}(z_j, z_k) = \rho^{|j-k|}$. We consider the moderate correlation between the covariates with $\rho = 0.5$. The true coefficients are $\boldsymbol{\beta} = (-1, -0.9, 0, 0, 0, -0.8, 0, 0, 0)^T$. Censoring times are generated from the uniform distribution over $[0, c_0]$, where $c_0$ is chosen to obtain the desired censoring rate. We

consider two types of censoring rate: 25% and 40%, and two sample sizes: $n = 100$ and $n = 200$.

Tables 1 and 2 summarize the MSEs and variable selection results for three methods under four different settings, when the PH and PO models are respectively fitted. Overall, the PAL with BIC works best in terms of the MSE and the overall variable selection. For example, under the PH model with $n = 100$ and 25% censoring rate, the LSE estimator tuned with BIC gives the smallest MSE (EE 0.244, PAL+AIC 0.149, PAL+BIC 0.122) and has the model size closest to the oracle (the true model size is 3, EE 9, PAL+AIC 4.66, PAL+BIC 3.61). Under the PO model with $n = 100$ and 40% censoring rate, the LSE estimator tuned with BIC gives the smallest MSE (EE 0.575, PAL+AIC 0.411, PAL+BIC 0.385) and selects covariates most accurately (the true model size is 3, EE 9, PAL+AIC 4.64, PAL+BIC 3.49).

(Insert Tables 1 and 2 here)

Tables 3 and 4 show the estimation bias for nonzero coefficients in various settings given by the PH and PO models. Overall speaking, the PAL estimators give a smaller finite sampling bias than the EE estimator, though all the three estimator are asymptotically unbiased in theory. We also note the PAL+BIC estimator tends to have a smaller bias then the PAL+AIC estimator, which is due to the selected $\lambda$ with BIC is generally larger than the parameter selected with AIC. When the sample size increases from $n = 100$ and $n = 200$, all the three estimators have decreased biases.

(Insert Tables 3 and 4 here)

To test the accuracy of the proposed standard error formula given in Section 4, we

compare the sample standard errors with their estimates. In Table 5, we summarize the average estimated standard errors ($\widehat{\text{SE}}$) given by the asymptotic estimator (4·1), the average estimated $\widehat{\text{SE}}_S$ given by the sandwich formula (4·4), and the sample standard errors from Monte Carlo simulations (SE), when $n = 200$ and the censored rate 25% and 40%, for both the PH and the PO model. The estimated standard errors of both methods are reasonably close to the sample standard errors, the (4·1) overall gives a better estimation than the sandwich formula. Also, the estimates for the PH model are generally better than those for the PO model. As expected, the estimates for the 25% censored rate setting are overall better than those for the 40% censored rate setting. We noted all the estimates tend to slightly under-estimate the actual MC standard error, which is mainly because the sandwich formula is derived when assuming a fixed $\lambda$, not taking into account the variance due to different tuning parameters $\lambda$ being chosen across runs. Similar patterns have been made for the shrinkage methods in other situations such as Tibshirani (1996) and Zhang and Lu (2007).

(Insert Table 5 here)

In Tables 6 and 7, we show the number of times of each variable being selected among 100 replicates in four settings, respectively for the PH model and the PO model. We observe that, the PAL+BIC estimator overall chooses unimportant variables with a much lower frequency than the PAL+AIC in all the settings.

(Insert Tables 6 and 7 here)

### 5·2. *Primary Biliary Cirrhosis Data*

The primary biliary cirrhosis (PBC) data was gathered from the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984. This data is provided in Therneau and Grambsch (2000), and a more detailed account can be found in Dickson *et al.* (1989). In this study, 312 patients from a total of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. For each patient, clinical, biochemical, serologic, and histological parameters are collected. Of those, 125 patients died before the end of follow-up. We study the dependence of the survival time on the following selected covariates: (1) continuous variables: age (in years), alb (albumin in g/dl), alk (alkaline phosphatase in U/liter), bil (serum bilirunbin in mg/dl), chol (serum cholesterol in mg/dl), cop (urine copper in $\mu$g/day), plat (platelets per cubic ml/1000), prot (prothrombin time in seconds), sgot (liver enzyme 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 of spiders), stage (histological stage of disease, graded 1, 2, 3 or 4), trt (1 for control, 2 for treatment).

We restrict our attention to the 276 observations without missing values. All the seventeen variables are included in the model. Table 8 summarizes the estimated coefficients by three methods and the corresponding standard errors for the PH model. As reported in Tibshirani (1997), the stepwise selection with the PH model chooses eight

variables: *age, oed,bil,alb, cop, sgot, prot* and *stage*. We found that the PAL+BIC identifies the exactly same set of important variables, and the PAL+AIC selects two additional variables *sex* and *chol*. Figure 1 depicts the solution path of the PAL estimator when fitting the PH model. Two vertical lines denote the PAL estimator corresponding to the BIC tuning (solid line) and the AIC tuning (broken line).

(Insert Table 8 and Figure 1 here)

Table 9 summarizes the estimated coefficients by three methods and the corresponding standard errors for the PO model. Interestingly, both the PAL estimators identify the same set of important risk factors as in the PH model. Also, the PAL+AIC selects two more variables:*sex* and *chol* than the PAL+BIC. Figure 2 depicts the solution path of the PAL estimator for the PO model fits. Two vertical lines denote the PAL estimator corresponding to the BIC tuning (solid line) and the AIC tuning (broken line).

(Insert Table 9 and Figure 2 here)

## 6. Discussion

The class of semiparametric linear transformation models has received much attention recently due to its high flexibility. In this paper, we have proposed a method to improve upon the estimation equation (EE) procedure proposed by Chen et al. (2002) by conducting variable selection and coefficient estimation simultaneously. Based on the numerical results, the PAL shows better performance than the standard EE method, in terms of both variable selection and model estimation. Theoretical properties, such

as root-$n$ consistency, variable selection consistency and asymptotic normality, of the PAL estimator have been established.

## Appendix 1

### *Proof of Theorem 1*

We follow similar steps as used in Fan & Li (2002) and Wang and Leng (2007). Since $Q(\boldsymbol{\beta})$ in (2·3) is strictly convex, there exists a unique global minimizer. Therefore it is sufficient to show that (2·3) has a $\sqrt{n}$-consistent local minimizer. According to Fan and Li (2001), we only need to show that, for any arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant $C$ such that that

$$\liminf_n P\left\{\inf_{\|\mathbf{r}\|=C} Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{r}) > Q(\boldsymbol{\beta}_0)\right\} \geq 1-\varepsilon, \tag{A.1}$$

where $\mathbf{r} = (r_1, \cdots, r_d)$. Simple algebra shows that

$$\begin{aligned}
& n\left[Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{r}) - Q(\boldsymbol{\beta}_0)\right] \\
= \;& \mathbf{r}'\widehat{\Sigma}^{-1}\mathbf{r} + 2\mathbf{r}'\widehat{\Sigma}^{-1}\left[\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right] + n\lambda\textstyle\sum_{j=1}^{d}|\beta_{j0} + n^{-1/2}r_j|/|\tilde{\beta}_j| - n\lambda\sum_{j=1}^{d}|\beta_{j0}|/|\tilde{\beta}_j| \\
= \;& \mathbf{r}'\widehat{\Sigma}^{-1}\mathbf{r} + 2\mathbf{r}'\widehat{\Sigma}^{-1}\left[\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right] + n\lambda\textstyle\sum_{j=1}^{d}|\beta_{j0} + n^{-1/2}r_j|/|\tilde{\beta}_j| - n\lambda\sum_{j=1}^{d_0}|\beta_{j0}|/|\tilde{\beta}_j| \\
\geq \;& \mathbf{r}'\widehat{\Sigma}^{-1}\mathbf{r} + 2\mathbf{r}'\widehat{\Sigma}^{-1}\left[\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right] + n\lambda\textstyle\sum_{j=1}^{d_0}\left[|\beta_{j0} + n^{-1/2}r_j| - |\beta_{j0}|\right]/|\tilde{\beta}_j| \\
\geq \;& \mathbf{r}'\widehat{\Sigma}^{-1}\mathbf{r} + 2\mathbf{r}'\widehat{\Sigma}^{-1}\left[\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right] - \sqrt{n}\lambda\textstyle\sum_{j=1}^{d_0}|r_j|/|\tilde{\beta}_j|. \qquad (A.2)
\end{aligned}$$

Since $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, we have, for $1 \leq j \leq d_0$,

$$\frac{1}{|\tilde{\beta}_j|} = \frac{1}{|\beta_{j0}|} - \frac{\mathrm{sign}(\beta_{j0})}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0}) + o_p(|\tilde{\beta}_j - \beta_{j0}|) = \frac{1}{|\beta_{j0}|} + \frac{O_p(1)}{\sqrt{n}}.$$

In addition, since $\sqrt{n}\lambda_n = O_p(1)$, we have

$$\sqrt{n}\lambda\sum_{j=1}^{d_0}|r_j|/|\tilde{\beta}_j| = \sqrt{n}\lambda\sum_{j=1}^{d_0}\left\{\frac{|r_j|}{|\beta_{j0}|} + \frac{|r_j|}{\sqrt{n}}O_p(1)\right\} \leq C\sqrt{n}\lambda O_p(1) = CO_p(1).$$

Recall that $\|r\| = C$. In (6), the first term is uniformly larger than $\nu_*(\widehat{\Sigma}^{-1})C^2 \to_p \nu_*(\widetilde{\Sigma}^{-1})C^2$, where $\nu_*(M)$ refers the minimal eigenvalue of $M$. So, with the probability

tending to one, the first term in (A.2) is uniformly larger than $0.5\nu_*(\widetilde{\Sigma}^{-1})C^2$, which is quadratic in $C$. Furthermore, the second term in (A.2) is uniformly bounded by $C\|\widehat{\Sigma}^{-1}\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}})\|$, which is linear in $C$ with the coefficient $\|\widehat{\Sigma}^{-1}\sqrt{n}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}})\| = O_p(1)$. Therefore, as long as $C$ is sufficiently large, the first term in (A.2) always dominates the other two terms with arbitrarily large probability. Therefore (A.1) holds and it completes the proof.

*Proof of Theorem 3*

In Section 3, we decompose the asymptotic covariance matrix $\widetilde{\Sigma}$ into the following block matrix form

$$\widetilde{\Sigma} = \begin{bmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{bmatrix},$$

where $\widetilde{\Sigma}_{11}$ is the first $d_0 \times d_0$ submatrix of $\widetilde{\Sigma}$. Correspondingly, partition its inverse matrix $\Omega = \Sigma^{-1}$ as

$$\widetilde{\Omega} = \begin{bmatrix} \widetilde{\Omega}_{11} & \widetilde{\Omega}_{12} \\ \widetilde{\Omega}_{21} & \widetilde{\Omega}_{22} \end{bmatrix}.$$

Here $\widetilde{\Omega}_{11}^{-1} = \widetilde{\Sigma}_{11} - \widetilde{\Sigma}_{12}\widetilde{\Sigma}_{22}^{-1}\widetilde{\Sigma}_{21}$. The corresponding estimators $\widehat{\Sigma}$ and $\widehat{\Omega}$ can be similarly partitioned.

According to Theorem 2, with probability tending to one, $\widehat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}$, so $\widehat{\boldsymbol{\beta}}^{(1)}$ must be the global minimizer of the objective function

$$\begin{aligned} Q_0(\boldsymbol{\beta}^{(1)}) \;=\; & \left[\boldsymbol{\beta}^{(1)} - \widetilde{\boldsymbol{\beta}}^{(1)}\right]' \widehat{\Omega}_{11} \left[\boldsymbol{\beta}^{(1)} - \widetilde{\boldsymbol{\beta}}^{(1)}\right] - 2\left[\boldsymbol{\beta}^{(1)} - \widetilde{\boldsymbol{\beta}}^{(1)}\right]' \widehat{\Omega}_{12}\widetilde{\boldsymbol{\beta}}^{(2)} \\ & + \widetilde{\boldsymbol{\beta}}^{(2)\prime}\widehat{\Omega}_{22}\widetilde{\boldsymbol{\beta}}^{(2)} + \lambda \sum_{j=1}^{d_0} \frac{|\beta_j|}{|\tilde{\beta}_j|}. \end{aligned}$$

According to Theorem 1, with probability tending to one, each component of $\widehat{\boldsymbol{\beta}}^{(1)}$

must be nonzero, so their partial derivatives exist and they satisfy the following normal equation

$$0 = \frac{1}{2}\frac{\partial Q_0(\boldsymbol{\beta}^{(1)})}{\partial \boldsymbol{\beta}^{(1)}}|_{\boldsymbol{\beta}^{(1)}=\widehat{\boldsymbol{\beta}}^{(1)}} = \widehat{\Omega}_{11}\left[\boldsymbol{\beta}^{(1)} - \widetilde{\boldsymbol{\beta}}^{(1)}\right] - \widehat{\Omega}_{12}\widetilde{\boldsymbol{\beta}}^{(2)} + G(\widetilde{\boldsymbol{\beta}}^{(1)}), \tag{A.2}$$

where $G(\widetilde{\boldsymbol{\beta}}^{(1)}) = [0.5\lambda \text{ sign}(\hat{\beta}_1)/|\tilde{\beta}_1|, \cdots, 0.5\lambda \text{ sign}(\hat{\beta}_{d_0})/|\tilde{\beta}_{d_0}|]'$. Using the theorem's condition $\sqrt{n}\lambda \to 0$, for each component in $\sqrt{n}G(\widetilde{\boldsymbol{\beta}}^{(1)})$, we have

$$0.5\sqrt{n}\lambda \text{ sign}(\hat{\beta}_j)/|\tilde{\beta}_j| = o_p(1), \quad 1 \le j \le d_0.$$

Note that (A.2) implies that

$$\begin{aligned}
\sqrt{n}\left[\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right] &= \sqrt{n}\left[\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right] + \widehat{\Omega}_{11}^{-1}\widehat{\Omega}_{12}(\sqrt{n}\widetilde{\boldsymbol{\beta}}^{(2)}) - \widehat{\Omega}_{11}^{-1}\sqrt{n}G(\widetilde{\boldsymbol{\beta}}^{(1)}) \\
&= \sqrt{n}\left[\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right] + \widehat{\Omega}_{11}^{-1}\widehat{\Omega}_{12}(\sqrt{n}\widetilde{\boldsymbol{\beta}}^{(2)}) - o_p(1) \\
&= \sqrt{n}\left[\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right] + \widetilde{\Omega}_{11}^{-1}\widetilde{\Omega}_{12}(\sqrt{n}\widetilde{\boldsymbol{\beta}}^{(2)}) - o_p(1), \qquad \text{(A.3)}
\end{aligned}$$

The third equation in (A.3) is due to the fact $\sqrt{n}(\tilde{\boldsymbol{\beta}}^{(2)}) = O_p(1)$. Therefore, $\sqrt{n}\left[\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right]$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix

$$\widetilde{\Sigma}_{11} + 2\widetilde{\Omega}_{11}^{-1}\widetilde{\Omega}_{12}\widetilde{\Sigma}_{21} + \widetilde{\Omega}_{11}^{-1}\widetilde{\Omega}_{12}\widetilde{\Sigma}_{22}\widetilde{\Omega}_{21}\widetilde{\Omega}_{11}^{-1}$$

Straight linear algebra shows $\widetilde{\Omega}_{11}^{-1}\widetilde{\Omega}_{12} = -\widetilde{\Sigma}_{12}\widetilde{\Sigma}_{22}^{-1}$. Therefore, we have

$$\sqrt{n}\left[\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right] \to N(0, \widetilde{\Sigma}_{11} - \widetilde{\Sigma}_{12}\widetilde{\Sigma}_{22}^{-1}\widetilde{\Sigma}_{21}).$$

## References

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric of transformation models with censored data. *Biometrika* **89**, 659-668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835-845.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with Discussion). *Journal of the Royal Statistical Society, Series A* **148**, 82-117.

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

Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in the two-sample generalized odds rate model. *Journal of American Statistical Association* **83**, 744-749.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics* **32**, 407-451.

Fan, J. and Li, R. (2001). Variable selection via penalized likelihood. *Journal of American Statistical Association* **99**, 1348-1360.

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

Fine, J., Ying, Z. and Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika* **85**, 980-986.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397-416.

Kalbfleish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, Edition 2, New Jersey: Wiley.

Lam, K. F. and Kuk, Y. C. (1997). A marginal likelihood approach to estimation in frailty models. *Journal of American Statistical Association* **92**, 985-990.

Lam, K. F. and Leung, T. L. (2001). Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis* **7**, 39-54.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, New York: Wiley.

Lu, W. and Zhang, H. H. (2007) Variable selection for proportional odds model. *Statistics in Medicine*, **26**, 3771-3781.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of American Statistical Association* **92**, 968-976.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society, Series B* **44**, 234-243.

Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Applied Statistics* **33**, 169-175.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **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.* SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, 59.

Wang, H. and Leng, C. (2007) Unified LASSO estimation with least squares approximation. *Journal of American Statistical Association*, to appear.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business & Economics Statistics* **20**, 347-355.

Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's proportional hazards model. **94**, 691-703.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of American Statistical Association* **101**, 1418-1429.

Table 1: Model selection and fitting results for PH model

| $n$ | 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) |
| | | PAL+AIC | 0.149 (0.127) | 4.660 (1.327) | 4.340 (1.327) | 0.000 (0.000) |
| | | PAL+BIC | 0.122 (0.119) | 3.610 (0.920) | 5.390 (0.920) | 0.000 (0.000) |
| | 40% | EE | 0.277 (0.186) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.178 (0.148) | 4.560 (1.438) | 4.440 (1.438) | 0.000 (0.000) |
| | | PAL+BIC | 0.143 (0.133) | 3.620 (0.885) | 5.380 (0.885) | 0.000 (0.000) |
| 200 | 25% | EE | 0.087 (0.052) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.057 (0.040) | 4.150 (1.266) | 4.850 (1.267) | 0.000 (0.000) |
| | | PAL+BIC | 0.051 (0.040) | 3.250 (0.557) | 5.750 (0.557) | 0.000 (0.000) |
| | 40% | EE | 0.110 (0.066) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.073 (0.055) | 4.370 (1.353) | 4.630 (1.353) | 0.000 (0.000) |
| | | PAL+BIC | 0.063 (0.049) | 3.280 (0.604) | 5.720 (0.604) | 0.000 (0.000) |

PH stands for proportional hazards model.

EE stands for the estimation equation estimate.

PAL+AIC stands for the PAL estimation estimate obtained with AIC.

PAL+BIC stands for the PAL estimation estimate obtained with BIC.

Table 2: Model selection and fitting results for PO model

| $n$ | 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) |
| | | PAL+AIC | 0.364 (0.259) | 4.670 (1.407) | 4.260 (1.397) | 0.070 (0.256) |
| | | PAL+BIC | 0.377 (0.303) | 3.600 (0.932) | 5.230 (0.874) | 0.170 (0.403) |
| | 40% | EE | 0.575 (0.347) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.411 (0.322) | 4.640 (1.487) | 4.310 (1.440) | 0.050 (0.219) |
| | | PAL+BIC | 0.385 (0.314) | 3.490 (0.916) | 5.360 (0.811) | 0.150 (0.386) |
| 200 | 25% | EE | 0.213 (0.109) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.150 (0.101) | 4.390 (1.392) | 4.610 (1.392) | 0.000 (0.000) |
| | | PAL+BIC | 0.122 (0.085) | 3.340 (0.670) | 5.660 (0.670) | 0.000 (0.000) |
| | 40% | EE | 0.258 (0.168) | 9 | 0 (0) | 0 (0) |
| | | PAL+AIC | 0.182 (0.148) | 4.540 (1.417) | 4.460 (1.417) | 0.000 (0.000) |
| | | PAL+BIC | 0.132 (0.086) | 3.310 (0.598) | 5.690 (0.598) | 0.000 (0.000) |

PO stands for proportional odds model.

EE stands for the estimation equation estimate.

PAL+AIC stands for the PAL estimation estimate obtained with AIC.

PAL+BIC stands for the PAL estimation estimate obtained with BIC.

Table 3: Estimation bias for the nonzero estimates in the PH model.

| n | Censored | Method | Important Coefficients | | |
|---|---|---|---|---|---|
| | | | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ | $\widehat{\beta}_6$ |
| 100 | 25% | EE | -0.077 (0.179) | -0.081 (0.200) | -0.046 (0.197) |
| | | PAL+AIC | -0.017 (0.179) | -0.021 (0.181) | -0.001 (0.187) |
| | | PAL+BIC | 0.038 (0.185) | 0.034 (0.168) | 0.053 (0.175) |
| | 40% | EE | -0.089 (0.204) | -0.073 (0.210) | -0.041 (0.219) |
| | | PAL+AIC | -0.019 (0.206) | -0.013 (0.130) | -0.013 (0.201) |
| | | PAL+BIC | 0.033 (0.204) | 0.041 (0.179) | 0.072 (0.201) |
| 200 | 25% | EE | -0.018 (0.116) | -0.018 (0.121) | -0.025 (0.124) |
| | | PAL+AIC | 0.013 (0.115) | 0.013 (0.115) | 0.007 (0.114) |
| | | PAL+BIC | 0.037 (0.113) | 0.037 (0.121) | 0.041 (0.110) |
| | 40% | EE | -0.020 (0.126) | -0.028 (0.136) | -0.028 (0.145) |
| | | PAL+AIC | 0.012 (0.125) | 0.005 (0.134) | 0.007 (0.139) |
| | | PAL+BIC | 0.044 (0.126) | 0.034 (0.135) | 0.053 (0.122) |

PH stands for proportional hazards model.

EE stands for the estimation equation estimate.

PAL+AIC stands for the PAL estimation estimate obtained with AIC.

PAL+BIC stands for the PAL estimation estimate obtained with BIC.

Table 4: Estimation bias for the nonzero estimated obtained in the PO model.

| n | Censored | Method | Important Coefficients | | |
|---|---|---|---|---|---|
| | | | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ | $\widehat{\beta}_6$ |
| 100 | 25% | EE | -0.056 (0.298) | 0.016 (0.284) | -0.032 (0.291) |
| | | PAL+AIC | -0.014 (0.301) | 0.091 (0.305) | 0.054 (0.290) |
| | | PAL+BIC | 0.045 (0.339) | 0.161 (0.332) | 0.128 (0.315) |
| | 40% | EE | -0.099 (0.301) | -0.008 (0.288) | -0.068 (0.297) |
| | | PAL+AIC | -0.043 (0.307) | 0.087 (0.316) | 0.042 (0.270) |
| | | PAL+BIC | 0.029 (0.337) | 0.167 (0.323) | 0.134 (0.302) |
| 200 | 25% | EE | -0.034 (0.184) | -0.009 (0.212) | -0.036 (0.198) |
| | | PAL+AIC | -0.003 (0.186) | 0.028 (0.217) | 0.013 (0.187) |
| | | PAL+BIC | 0.036 (0.187) | 0.072 (0.211) | 0.068 (0.165) |
| | 40% | EE | -0.055 (0.199) | -0.041 (0.234) | -0.060 (0.221) |
| | | PAL+AIC | -0.018 (0.203) | -0.009 (0.224) | -0.012 (0.221) |
| | | PAL+BIC | 0.027 (0.196) | 0.043 (0.225) | 0.060 (0.187) |

PO stands for proportional odds model.

EE stands for the estimation equation estimate.

PAL+AIC stands for the PAL estimation estimate obtained with AIC.

PAL+BIC stands for the PAL estimation estimate obtained with BIC.

Table 5: Estimated and MC standard errors for the PAL nonzero estimates ($n = 200$).

| | | | $\hat{\beta}_1$ | | | $\hat{\beta}_2$ | | | $\hat{\beta}_6$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | Model | Tune | SE | $\widehat{\text{SE}}$ | $\widehat{\text{SE}}_S$ | SE | $\widehat{\text{SE}}$ | $\widehat{\text{SE}}_S$ | SE | $\widehat{\text{SE}}$ | $\widehat{\text{SE}}_S$ |
| 25% | PH | AIC | 0.115 | 0.110 | 0.108 | 0.115 | 0.108 | 0.104 | 0.114 | 0.098 | 0.094 |
| | | BIC | 0.113 | 0.109 | 0.105 | 0.121 | 0.105 | 0.100 | 0.110 | 0.092 | 0.088 |
| | PO | AIC | 0.186 | 0.167 | 0.160 | 0.217 | 0.169 | 0.159 | 0.187 | 0.155 | 0.146 |
| | | BIC | 0.187 | 0.165 | 0.152 | 0.211 | 0.164 | 0.147 | 0.165 | 0.146 | 0.131 |
| 40% | PH | AIC | 0.125 | 0.122 | 0.118 | 0.134 | 0.119 | 0.115 | 0.139 | 0.110 | 0.105 |
| | | BIC | 0.126 | 0.120 | 0.114 | 0.135 | 0.116 | 0.109 | 0.122 | 0.103 | 0.097 |
| | PO | AIC | 0.203 | 0.179 | 0.171 | 0.224 | 0.183 | 0.173 | 0.221 | 0.167 | 0.155 |
| | | BIC | 0.196 | 0.176 | 0.161 | 0.225 | 0.177 | 0.156 | 0.187 | 0.155 | 0.138 |

PH and PO are defined the same as in Table 1.

SE stands for the sample standard errors of the estimated coefficients.

$\widehat{\text{SE}}$ stands for the average of estimated SE based on (4·1).

$\widehat{\text{SE}}_S$ stands for the average of estimated SE based on the sandwich formula (4·4).

Table 6: Frequency of variables selected for PH model.

| n | Method | Censored | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ | $z_6$ | $z_7$ | $z_8$ | $z_9$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | PAL+AIC | 25% | 100 | 100 | 27 | 25 | 25 | 100 | 29 | 29 | 31 |
| | | 40% | 100 | 100 | 25 | 23 | 24 | 100 | 31 | 25 | 28 |
| | PAL+BIC | 25% | 100 | 100 | 12 | 6 | 10 | 100 | 12 | 7 | 14 |
| | | 40% | 100 | 100 | 10 | 9 | 10 | 100 | 14 | 9 | 10 |
| 200 | PAL+AIC | 25% | 100 | 100 | 19 | 23 | 21 | 100 | 24 | 14 | 14 |
| | | 40% | 100 | 100 | 18 | 23 | 30 | 100 | 26 | 20 | 20 |
| | PAL+BIC | 25% | 100 | 100 | 6 | 6 | 3 | 100 | 5 | 2 | 3 |
| | | 40% | 100 | 100 | 4 | 3 | 7 | 100 | 5 | 5 | 4 |

Table 7: Frequency of variables selected for PO model.

| n | Method | Censored | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ | $z_6$ | $z_7$ | $z_8$ | $z_9$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | PAL+AIC | 25% | 100 | 97 | 28 | 32 | 29 | 96 | 31 | 25 | 29 |
| | | 40% | 100 | 96 | 27 | 28 | 31 | 99 | 32 | 24 | 27 |
| | PAL+BIC | 25% | 97 | 93 | 15 | 14 | 16 | 93 | 13 | 11 | 8 |
| | | 40% | 96 | 95 | 11 | 11 | 15 | 94 | 13 | 7 | 7 |
| 200 | PAL+AIC | 25% | 100 | 100 | 20 | 23 | 23 | 100 | 26 | 25 | 22 |
| | | 40% | 100 | 100 | 28 | 28 | 29 | 100 | 22 | 25 | 22 |
| | PAL+BIC | 25% | 100 | 100 | 6 | 4 | 8 | 100 | 6 | 4 | 6 |
| | | 40% | 100 | 100 | 7 | 4 | 8 | 100 | 4 | 4 | 4 |

Table 8: Estimated coefficients and standard errors for PBC data with PH Model.

| Covariate | EE | PAL+AIC | PAL+BIC |
|---|---|---|---|
| trt | -0.109 (0.234) | 0 (-) | 0 (-) |
| age | 0.029 (0.012) | 0.029 (0.011) | 0.017 (0.007) |
| sex | -0.386 (0.346) | -0.200 (0.220) | 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) | 0.961 (0.428) | 0.576 (0.241) |
| bil | 0.079 (0.024) | 0.078 (0.022) | 0.099 (0.018) |
| chol | 0.001 (0.000) | 0.001 (0.000) | 0 (-) |
| alb | -0.811 (0.286) | -0.776 (0.266) | -0.755 (0.211) |
| cop | 0.003 (0.001) | 0.003 (0.001) | 0.003 (0.001) |
| alk | 0.000 (0.000) | 0 (-) | 0 (-) |
| sgot | 0.004 (0.002) | 0.003 (0.002) | 0.002 (0.001) |
| trig | -0.001 (0.001) | 0 (-) | 0 (-) |
| plat | 0.001 (0.001) | 0 (-) | 0 (-) |
| prot | 0.238 (0.103) | 0.237 (0.096) | 0.193 (0.066) |
| stage | 0.450 (0.171) | 0.453 (0.162) | 0.413 (0.121) |

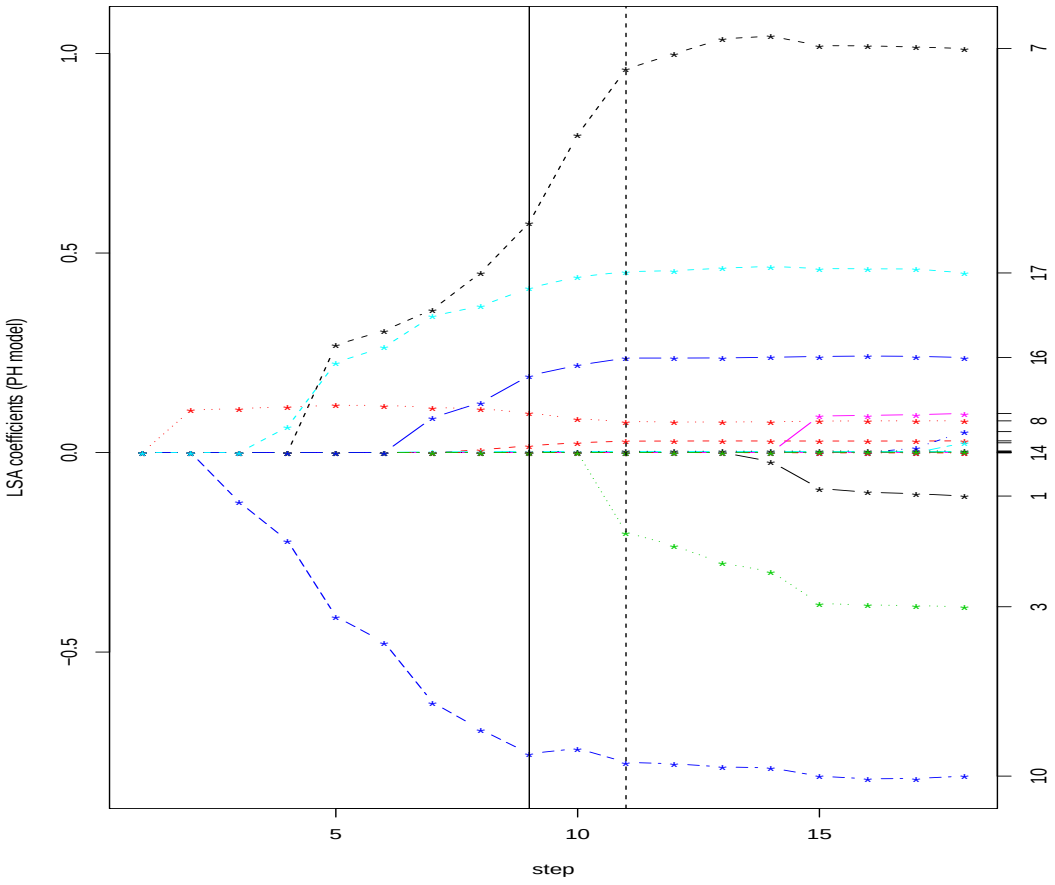


Figure 1: The solution path of PAL estimates for the PBC data when fitting the PH model. Two vertical lines denote the PAL estimates obtained with BIC (solid line) tuning and AIC (broken line) tuning.

Table 9: Estimated coefficients and standard errors for PBC data with PO Model.

| Covariate | EE | PAL+AIC | PAL+BIC |
|---|---|---|---|
| trt | -0.132 (0.322) | 0 (-) | 0 (-) |
| age | 0.041 (0.017) | 0.035 (0.016) | 0.024 (0.011) |
| sex | -0.572 (0.492) | -0.278 (0.296) | 0 (-) |
| asc | 0.349 (0.855) | 0 (0) | 0 (-) |
| hep | 0.057 (0.361) | 0 (-) | 0 (-) |
| spid | 0.252 (0.406) | 0 (-) | 0 (-) |
| oed | 1.189 (0.797) | 1.331 (0.647) | 0.905 (0.382) |
| bil | 0.110 (0.034) | 0.109 (0.030) | 0.131 (0.027) |
| chol | 0.001 (0.001) | 0.001 (0.000) | 0 (-) |
| alb | -1.090 (0.380) | -1.083 (0.350) | -1.036 (0.301) |
| cop | 0.004 (0.002) | 0.004 (0.002) | 0.004 (0.001) |
| alk | 0.000 (0.000) | 0 (-) | 0 (-) |
| sgot | 0.005 (0.003) | 0.004 (0.002) | 0.002 (0.001) |
| trig | -0.001 (0.003) | 0 (-) | 0 (-) |
| plat | 0.001 (0.002) | 0 (-) | 0 (-) |
| prot | 0.308 (0.143) | 0.318 (0.130) | 0.272 (0.098) |
| stage | 0.552 (0.226) | 0.598 (0.208) | 0.554 (0.168) |

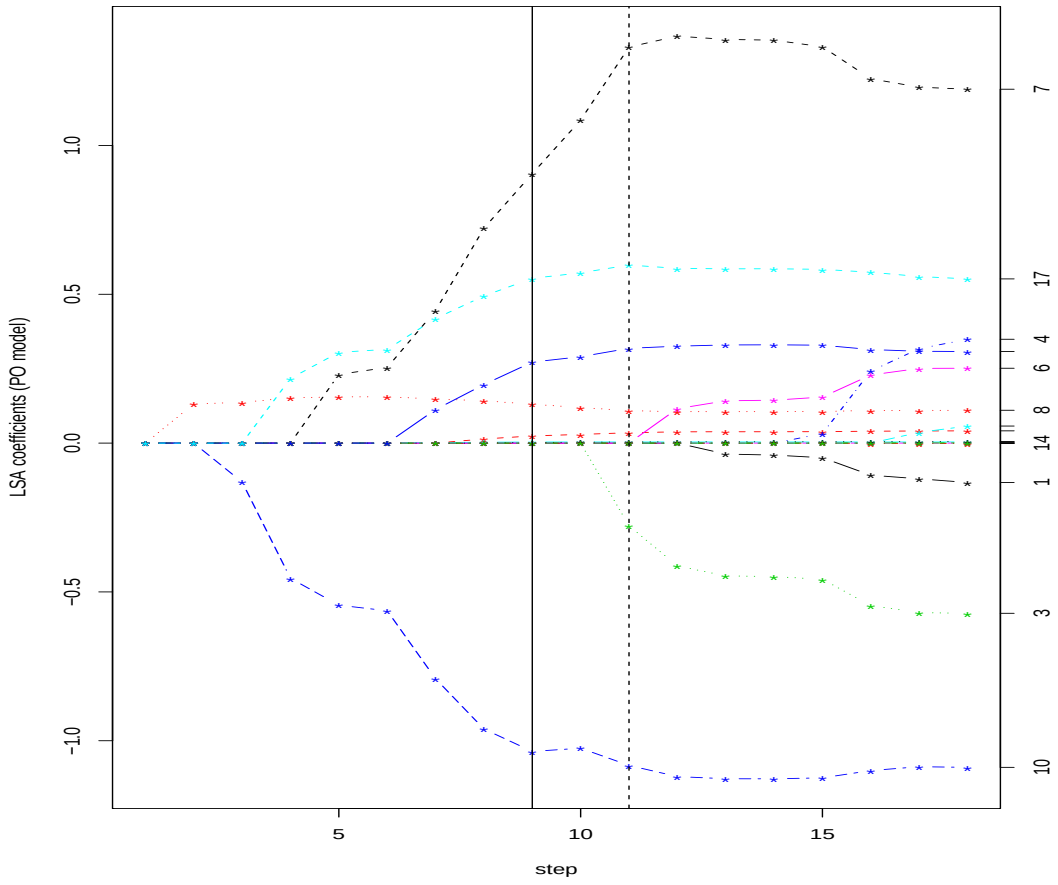


Figure 2: The solution path of PAL estimates for the PBC data when fitting the PO model. Two vertical lines denote the PAL estimates obtained with BIC (solid line) tuning and AIC (broken line) tuning.