Published in final edited form as:
J Multivar Anal. 2010 August 1; 101(7): 1594–1606. doi:10.1016/j.jmva.2010.01.015.

On Sparse Estimation for Semiparametric Linear Transformation Models

Hao Helen Zhang, **Wenbin Lu**, and **Hansheng Wang**
Hao Helen Zhang: hzhang2@stat.ncsu.edu; Wenbin Lu: lu@stat.ncsu.edu; Hansheng Wang: hansheng@gsm.pku.edu.cn

Abstract

Semiparametric linear transformation models have received much attention due to its high flexibility in modeling survival data. A useful estimating equation procedure was recently proposed by Chen et al. (2002) for linear transformation models to jointly estimate parametric and nonparametric terms. They showed that this procedure can yield a consistent and robust estimator. However, the problem of variable selection for linear transformation models is less studied, partially because a convenient loss function is not readily available under this context. In this paper, we propose a simple yet powerful approach to achieve both sparse and consistent estimation for linear transformation models. The main idea is to derive a profiled score from the estimating equation of Chen et al. (2002), construct a loss function based on the profile scored and its variance, and then minimize the loss subject to some shrinkage penalty. Under regularity conditions, we have shown that the resulting estimator is consistent for both model estimation and variable selection. Furthermore, the estimated parametric terms are asymptotically normal and can achieve higher efficiency than that yielded from the estimation equations. For computation, we suggest a one-step approximation algorithm which can take advantage of the LARS and build the entire solution path efficiently. Performance of the new procedure is illustrated through numerous simulations and real examples including one microarray data.

Key words and phrases

Censored survival data; Linear transformation models; LARS; Shrinkage; Variable selection

1. Introduction

In the last three decades, various semiparametric models have been proposed and extensively studied for the analysis of censored survival data. Among them, the proportional hazards model (Cox, 1972) and its associated partial likelihood principle (Cox, 1975) are commonly used in practice due to its nice theoretical properties and empirical performance. However, the proportional hazards assumption is often too restrictive and may be violated in some biomedical applications. Thus, other semiparametric models which relax such an assumption provide useful alternatives. For example, if the hazard functions of two treatment groups converge to the same limit, the proportional odds model (Pettitt, 1982, 1984; Bennett, 1983; Dabrowska and Doksum, 1988; Murphy et al. 1997) is preferable to the proportional hazards model. More

Correspondence to: Hao Helen Zhang, hzhang2@stat.ncsu.edu.

generally, a class of linear transformation models (Bickel et al., 1993; Cheng et al., 1995; Fine et al., 1998; Zeng and Lin, 2006; Zeng and Lin, 2007) have been proposed as a flexible alternative approach to modeling survival data. The linear transformation model is specified by

$$H(T) = -\beta' \mathbf{Z} + \varepsilon, \tag{1.1}$$

where H is an unknown monotone increasing function, $\mathbf{Z} = (Z_1,\cdots,Z_d)'$ are the d-dimensional covariates, $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_d)'$ is the regression parameter vector, and ε has a known continuous distribution that is independent of $\mathbf{Z}$. Linear transformation models form a rich class and include the proportional hazards (PH) model and the proportional odds (PO) model as special cases: the PH model corresponds to an error with the extreme value distribution and the PO model to an error following the logistic distribution. In addition, if ε follows the standard normal distribution, the model (1.1) naturally generalizes the usual Box-Cox transformation models.

In this paper, we consider the problem of model selection and estimation for (1.1) when the true model has a sparse representation, i.e. some components of $\boldsymbol{\beta}$ are exactly zero. Let $\mathscr{A} = \{j : \beta_j \neq 0, j = 1,\cdots,d\}$. Our goal is to discover the important index set $\mathscr{A}$ and estimate the corresponding coefficients consistently. Variable selection is fundamental to survival data analysis, since it helps medical researchers build more interpretable models without information loss and in the long run leads to better disease diagnosis and treatment. Traditional procedures include stepwise selection and best subset procedures. However, these procedures may suffer from high computational cost and selection variability (Breiman, 1996). Recently some shrinkage methods have been proposed for Cox's proportional hazards model based on the penalized partial likelihood, including the LASSO (Tibshirani 1997), the SCAD (Fan and Li, 2002) and the adaptive LASSO (Zou 2006;Zhang and Lu, 2007). For the proportional odds model, Lu and Zhang (2007) suggested the penalized marginal likelihood method for variable selection.

There has been less development for variable selection in semiparametric linear transformational models. This is partially due to substantial challenges in fitting linear transformation models: the lack of a convenient loss function and the need of estimating an infinite-dimensional parameter. Furthermore, most estimation procedures for linear transformation models are based on estimating equations (e.g., Cheng et al., 1995; Fine et al., 1998; Chen et al., 2002), which makes it difficult to incorporate a shrinkage penalty for variable selection as done for Cox's proportional hazards model. In this paper, we propose a simple yet powerful approach to achieve both sparse and consistent estimation for linear transformation models. The main idea is to derive a profiled score from the estimating equation of Chen et al. (2002), construct a loss function based on the profile score and its variance, and then minimize the constructed loss subject to some shrinkage penalty. Variable selection for estimating equations has drawn a lot of attention in other contexts and been recently studied by Fu (2003), Qu and Li (2006), and Johnson et al. (2008). In particular, Johnson et al. (2008) proposed an effective procedure which directly penalizes the estimation equation. It is noted that their procedure does not yield exactly zeros, while our estimator penalizes a quadratic loss constructed from the estimation equations and has sparsity property.

Our estimator is closely related to the least squares approximation (LSA) procedure of Wang and Leng (2007). For the likelihood or more general loss based estimation procedures, Wang and Leng (2007) proposed to penalize the second-order Taylor expansion of the loss function instead of the loss function itself subject to a shrinkage penalty. They showed that this quadratic approximation problem is not only easier to implement but also yields consistent and sparse estimators for parametric models. In this paper, we have generalized the idea to estimating

procedures where a loss function is not readily available, for example, the estimation equation estimator. The new estimator is generally different from LSA, but we show that its one-step estimation is asymptotically equivalent to the LSA. Compared to existing work for linear transformation models, the new procedure makes several unique contributions: (i) it lays down a general framework to construct a loss function based on the estimation equations, so that the penalized method can be adopted for sparse estimation; (ii) the profiled score takes care of the nonparametric component in a natural fashion; (iii) the new estimator has an improved efficiency over the estimator resulted from the estimation equations.

The remainder of this article is organized as follows. Section 2 proposes the new estimator for linear transformation models and studies asymptotic properties of the resulting estimator. Section 3 introduces the computational algorithms for computing the estimates. Section 4 derives the variance estimates of the estimates and discusses the selection of the regularization parameter. Sections 5 is devoted to simulation studies and real data analysis. Final remarks are given in Section 6. Major technical derivations are contained in the Appendix section.

2. New Estimation for Linear Transformation Models

2.1. Methods

Assume that the failure time T is from model (1.1). In the presence of censoring, we observe the event time $\tilde{T}_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \le C_i)$, where C_i is the censoring time of subject i and $I(\cdot)$ is the indicator function. Here we assume that the censoring variable C_i is independent of T_i given $\mathbf{Z}_i$. Suppose a random sample of n individuals is chosen, then the observations consist of $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$, $i = 1,\ldots,n$. Without loss of generality, we assume that $\mathbf{Z}$'s are standardized such that $\sum_{i=1}^{n} Z_{ij}=0$ and $\sum_{i=1}^{n} Z_{ij}^2=1$, for $j = 1,\cdots,d$.

Let $N_i(t) = \delta_i I(\tilde{T}_i \le t)$ and $Y_i(t) = I(\tilde{T}_i \ge t)$ respectively denote the counting and at-risk processes of the ith subject. In addition, define

$$M_i(t)=N_i(t) - \int_0^t Y_i(s)d\Lambda\{H_0(s)+\beta_0' Z_i\}, i=1,\cdots,n, \tag{2.1}$$

where $\Lambda(\cdot)$ is the known cumulative hazard function of ε and $(\boldsymbol{\beta}_0, H_0)$ are the true values of $(\boldsymbol{\beta}, H)$. Using the counting process and its associated martingale theory (Fleming and Harrington, 1991; Andersen et al., 1993), one can show that $M_i(t)$ is a mean zero martingale process. To estimate $\boldsymbol{\beta}$ and H, Chen et al. (2002) proposed a novel martingale-representation based estimating equation approach, which solves the following equations:

$$\begin{aligned} \sum_{i=1}^{n} \int_0^{\tau} \mathbf{Z}_i[\, dN_i(t) - Y_i(t)d\Lambda\{\beta'\mathbf{Z}_i+H(t)\}] &=0, \\ \sum_{i=1}^{n} [\, dN_i(t) - Y_i(t)d\Lambda\{\beta'\mathbf{Z}_i+H(t)\}] &=0, t \ge 0. \end{aligned} \tag{2.2}$$

Given $\boldsymbol{\beta}$, the left-hand side of the second equation of (2.2) is monotone in H and therefore has the unique solution, denoted by $\tilde{H}(\cdot; \boldsymbol{\beta})$. Denote the solutions to (2.2) as $\tilde{\boldsymbol{\beta}}_n$ and $\tilde{H}(\cdot; \tilde{\boldsymbol{\beta}}_n)$. For convenience, we call them as the EE (estimation equation) estimator in the rest of paper. Chen et al. (2002) studied their theoretical properties and showed that $\sqrt{n}(\tilde{\beta}_n - \beta_0) \to N(0, \Sigma)$ in distribution as $n \to \infty$, where Σ has a sandwich form $\Sigma = A^{-1}V(A^{-1})'$. They also suggested $\tilde{A}_n/n$ and $\tilde{V}_n/n$ respectively as a consistent estimator of A and V, where $\tilde{A}_n \equiv \tilde{A}_n\{\tilde{\boldsymbol{\beta}}_n, \tilde{H}(\cdot; \tilde{\boldsymbol{\beta}}_n)\}$

and $\tilde{V}_n \equiv \tilde{V}_n\{\tilde{\boldsymbol{\beta}}_n, \tilde{H}(\cdot; \tilde{\boldsymbol{\beta}}_n)\}$. See Chen et al. (2002) Section 2 for the expressions of A, V, $\tilde{A}_n$ and $\tilde{V}_n$. We can define $\tilde{\Sigma}_n=(n\tilde{A}_n^{-1})(\tilde{V}_n/n)(n\tilde{A}_n^{-1})'$, which is a consistent estimator of Σ.

Variable selection is often challenging for the estimation procedure based on solving (2.2), since there is not a convenient loss function available and the estimation involves an infinite dimensional parameter H. To tackle these difficulties, we develop a new estimate procedure in several steps. Firstly, we introduce the notion of the "profiled" score, which is computed by plugging H into the left-side of the first equation in (2.2):

$$U_n(\beta)=\sum_{i=1}^{n}\int_0^{\tau}\mathbf{Z}_i[\,dN_i(t) - Y_i(t)d\Lambda\{\beta'\mathbf{Z}_i+\tilde{H}(t;\beta)\}]. \tag{2.3}$$

Note that the score U_n depends on H implicitly. Secondly, we use U_n and its variance estimate to construct a loss function as

$$D_n(\beta)=U_n'(\beta)\tilde{V}_n^{-1}U_n(\beta), \tag{2.4}$$

where the inverse variance $\tilde{V}_n^{-1}$ of the profiled score U_n is the weight matrix. Later on, we show that this particular choice of weight can provide gain in estimation efficiency. We will refer to D_n as the weighted profiled score squares (WPSS). D_n is a continuous function in $\boldsymbol{\beta}$. To achieve sparse estimation, we finally propose minimizing

$$Q_n(\beta)=D_n(\beta)+n\lambda\sum_{j=1}^{d}w_j|\beta_j|, \tag{2.5}$$

where the weights w_j's are pre-selected non-negative constants and $\lambda > 0$ is the tuning parameter. When the weights w_j's are all equal to one, the selection procedure is based on the LASSO penalty (Tibshirani 1997). A general choice of w_j's in (2.5) leads to the adaptive LASSO penalty, recently studied in various contexts including linear models (Zou, 2006), LAD regression models (Wang et al., 2007), the Cox proportional hazard models (Zhang and Lu, 2007;Zou, 2008), the proportional odds model (Lu and Zhang, 2007) and regression models with auto-regressive errors (Wang et al., 2007). The weight w_j's are leverage factors used to adjust penalties on individual regression coefficients, taking large values for unimportant covariates and small values for important covariates. In this paper, we use $w_j = 1/|\tilde{\beta}_j|$, where $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1,\cdots,\tilde{\beta}_d)'$. As shown in the next section, any root-n consistent estimator of $\boldsymbol{\beta}$ can be used to construct the weights w, and they will assure the consistency the new estimator for both model estimation and variable selection in theory

2.2. Asymptotic Properties

Now consider the following estimator

$$\widehat{\beta}_n=\text{argmin}_{\beta}\{D_n(\beta)+n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j|\}. \tag{2.6}$$

In this section, we study the asymptotic properties of $\hat{\boldsymbol{\beta}}_n$. Without loss of generality, we assume that the true important index set $\mathscr{A} = \{1,\cdots,q\}$, where q is an integer and $0 \le q \le d$. Therefore

we have $\beta_0=(\beta_{01}',\beta_{02}')'$, where $\boldsymbol{\beta}_{01}$ contains the first q nonzero components. We further decompose the covariance matrix

$$\Sigma=A^{-1}V(A^{-1})'=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}.$$

where Σ_{11} is the first $q\times q$ submatrix of Σ. In addition, write $\widehat{\beta}_n=(\widehat{\beta}_{n1}',\widehat{\beta}_{n2}')'$, where $\hat{\boldsymbol{\beta}}_{n1}$ consists of all the nonzero coefficients.

In order to study the asymptotic properties of the new estimator, we assume the following regularity conditions used in Chen et al. (2002):

(c1) The covariates Z are bounded with probability 1;

(c2) $\boldsymbol{\beta}_0$ belongs to the interior of a known compact set $\mathcal{B}_0$ and H_0 has a continuous and positive derivative;

(c3) $\lambda(\cdot)\equiv\dot{\Lambda}(\cdot)$ is positive, $\psi(\cdot)\equiv\dot{\lambda}(\cdot)/\lambda(\cdot)$ is continuous, and $\lim_{t\to-\infty}\lambda(t)=0=\lim_{t\to-\infty}\psi(t)$;

(c4) τ is finite, satisfying $P(T>\tau)>0$ and $P(C=\tau)>0$;

(c5) A and V are finite and non-degenerate.

(c6) $\frac{1}{n}\frac{\partial^2 U_n(\beta)}{\partial\beta^2}|_{\beta=\beta_0}=W+o_p(1)$ for some finite and positive definite W.

In the following theorems, we establish the $\sqrt{n}-$consistency, selection consistency, and asymptotic normality of the proposed estimator. The proofs are given in the Appendix sections.

THEOREM 1 *Under the regularity conditions,*

- **i.** *($\sqrt{n}-$consistency) If $\sqrt{n}\lambda=O(1)$, then $\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$.*
- **ii.** *(Selection consistency) If $\sqrt{n}\lambda=O(1)$ and $n\lambda\to\infty$, then $P(\hat{\boldsymbol{\beta}}_{n2}=\mathbf{0})\to 1$.*

Remark 1: Based on the theoretical proof given in the Appendix, we can conclude that any root-n consistent estimator of $\boldsymbol{\beta}$ can be used to construct the weights w's. Both Theorems 1 and 2 hold as long as the reciprocal of weights are root-n consistent for $\boldsymbol{\beta}$.

THEOREM 2 *(Asymptotic normality) Under the regularity conditions, if $\sqrt{n}\lambda\to 0$ and $n\lambda\to\infty$, then as $n\to\infty$,*

$$\sqrt{n}(\widehat{\beta}_{n1}-\beta_{01})\to N(\mathbf{0},\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Remark 2: It is easy to see that, the efficiency of the new estimator for nonzero components is improved over that of the corresponding full model estimator obtained from the estimation equation because $\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}<\Sigma_{11}$. The efficiency gain of the new estimator is due to the weight matrix $\tilde{V}_n^{-1}$ used in (2.4). If the weight is chosen as the constant matrix, say, the identify matrix, such an improvement in efficiency is not warranted.

3. Computational Algorithm

To solve the minimization problem (2.6), we start with an initial estimator $\hat{\boldsymbol{\beta}}^{[0]}$ and approximate $U_n(\boldsymbol{\beta})$ by its first order Taylor expansion around $\hat{\boldsymbol{\beta}}^{[0]}$. Based on the theoretical results of Chen et al. (2002), we have the following linear approximation for U_n at the initial point

$$\frac{1}{n}U_n(\beta) \approx \frac{1}{n}U_n(\widehat{\beta}^{[0]}) + \frac{1}{n}\tilde{A}_n\{\widehat{\beta}^{[0]}, \tilde{H}(\cdot;\widehat{\beta}^{[0]})\}(\beta - \widehat{\beta}^{[0]}),$$

where $\tilde{A}_n/n$ can be regarded as the asymptotic derivative of $\frac{1}{n}U_n$ with respect to $\boldsymbol{\beta}$. Then the objective function D can be locally approximated by a quadratic form

$$\begin{aligned}\tilde{D}_n(\beta) &= [\,U_n(\widehat{\beta}^{[0]}) + \tilde{A}_n\{\widehat{\beta}^{[0]}, \tilde{H}(\cdot;\widehat{\beta}^{[0]})\}(\beta - \widehat{\beta}^{[0]})]'\tilde{V}_n^{-1} \times [\,U_n(\widehat{\beta}^{[0]}) + \tilde{A}_n\{\widehat{\beta}^{[0]}, \tilde{H}(\cdot;\widehat{\beta}^{[0]})\}(\beta - \widehat{\beta}^{[0]})] \\ &= (\beta - \widehat{\beta}^{[0]})'\tilde{A}_n^{[0]\prime}\tilde{V}_n^{-1}\tilde{A}_n^{[0]}(\beta - \widehat{\beta}^{[0]}) + 2U_n(\widehat{\beta}^{[0]})'\tilde{V}_n^{-1}\tilde{A}_n^{[0]}(\beta - \widehat{\beta}^{[0]}) + \text{constant},\end{aligned}$$

where $\tilde{A}_n^{[0]} = \tilde{A}_n\{\widehat{\beta}^{[0]}, \tilde{H}(\cdot;\widehat{\beta}^{[0]})\}$. With some algebraic derivation, this leads to the following quadratic optimization problem

$$\min_{\beta}\{(\beta - \widehat{\beta}^{[0]} - \mathbf{b})\tilde{A}_n^{[0]\prime}\tilde{V}_n^{-1}\tilde{A}_n^{[0]}(\beta - \widehat{\beta}^{[0]} - \mathbf{b}) + n\lambda\sum_{j=1}^{d}|\beta^j|/|\tilde{\beta}_j|\}, \tag{3.1}$$

where $\mathbf{b} = (\tilde{A}_n^{[0]\prime}\tilde{V}_n^{-1}\tilde{A}_n^{[0]})^{-1}\tilde{A}_n^{[0]\prime}\tilde{V}_n^{-1}U_n(\widehat{\beta}^{[0]})$. Since $\tilde{D}_n(\boldsymbol{\beta})$ is quadratic in $\boldsymbol{\beta}$, the corresponding minimization problem can be easily solved using standard packages for computing LASSO, such as the shooting algorithm (Fu 1998), the algorithm proposed by Osborne et al. (2000), and *lars* algorithm (Efron et al., 2004). For the PEE estimator, we propose the following iterative algorithm:

ALGORITHM:

step 1: Choose an initial estimator $\hat{\boldsymbol{\beta}}^{[0]}$.

step 2: Solve the second equation of (2.2) to obtain $\tilde{H}(\cdot; \hat{\boldsymbol{\beta}}^{[0]})$.

step 3: Minimize (3.1) and denote the solution as $\hat{\boldsymbol{\beta}}^{[1]}$.

step 4: Set $\hat{\boldsymbol{\beta}}^{[0]} = \hat{\boldsymbol{\beta}}^{[1]}$.

step 5: Go to step 2 until convergence.

Note that the algorithm above needs to update $\tilde{H}$ iteratively by solving (2.2) at each step, which can be computationally expensive in practice. Interestingly, if the initial estimator $\hat{\boldsymbol{\beta}}^{[0]}$ is chosen good enough, one does not have to iterate the algorithm until its convergence and one-step iteration is often sufficient. In particular, we suggest using the initial estimate $\hat{\boldsymbol{\beta}}^{[0]} = \tilde{\boldsymbol{\beta}}_n$, from $U_n(\tilde{\boldsymbol{\beta}}_n) = 0$, due to its consistency. It is known that $\tilde{\boldsymbol{\beta}}_n$ is $\sqrt{n}-$consistent, which assures that the initial estimate is pretty close to the true parameter. The optimization problem in (3.1) then becomes

$$\min_{\beta}\{(\beta - \tilde{\beta}_n)\tilde{A}_n\tilde{V}_n^{-1}\tilde{A}_n(\beta - \tilde{\beta}_n) + n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j|\}. \tag{3.2}$$

Remark 3: Interestingly, the one-step solution is asymptotically equivalent to the LSA procedure (Wang and Leng, 2007) if there were a loss function to start with. In this sense, the new procedure can be regarded as a generalization of the LSA to complicated models where the LSA is not directly applicable due to the unavailability of a loss function.

The one-step procedure is in the same spirit of the one-step M-estimation (Le Cam, 1956). A good overview of the one-step M-estimation can be found in Le Cam and Yang (1990) and van der Vaart (1998). Similar discussions are also given by Fan and Li (2001) and Zou and Li (2008) for the SCAD estimator. Our empirical experience shows that one-step iteration performs very well for the new estimator. Another advantage of the one-step procedure is that the entire solution path can be obtained using the *lars* package (Efron et al., 2004) in R. Consequently, we suggest using the one-step estimator in practice and will demonstrate its empirical performance in Section 6.

4. Variance Estimation and Parameter Tuning

In the following, we suggest two estimation formula for the covariance of the nonzero estimates $\hat{\boldsymbol{\beta}}_{n1}$. These two estimators are asymptotically equivalent. The first estimator is based on the asymptotic normality result given in Theorem 3. Correspondingly, we can partition $\tilde{\Sigma}_n^{-1}$ as

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

Then the covariance of $\hat{\boldsymbol{\beta}}_{n1}$ can be approximated as

$$\widehat{\mathrm{Cov}}(\hat{\beta}_{n1})=(\tilde{\Sigma}_{11}-\tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21})/n=\tilde{\Omega}_{11}^{-1}/n. \tag{4.1}$$

Next, we derive a sandwich formula to approximate the covariance of $\hat{\boldsymbol{\beta}}_{n1}$. 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 SCAD estimates. In the following, we apply the LQA approach to derive the covariance estimate for the nonzero PEE estimates. For any nonzero β_j, we can approximate its weighted L_1 penalty with a local quadratic function

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

The nonzero PEE estimates is obtained by one-step optimization problem in (3.2), which can be approximated by the following ridge-type regression

$$(\beta_1-\tilde{\beta}_{n1})'\tilde{\Omega}_{11}(\beta_1-\tilde{\beta}_{n1})-2(\beta_1-\tilde{\beta}_{n1})'\tilde{\Omega}_{12}\tilde{\beta}_{n2}+\lambda\beta_1'E\beta_1, \tag{4.2}$$

where $\tilde{\beta}_n=(\tilde{\beta}_{n1}',\tilde{\beta}_{n2}')'$ and $E=\mathrm{diag}\{1/\tilde{\beta}_1^2,\cdots,1/\tilde{\beta}_d^2\}$. The solution of (4.2) is

$$\hat{\beta}_{n1}=\left[\tilde{\Omega}_{11}+\lambda E_1\right]^{-1}\tilde{\Omega}_{11}(\tilde{\beta}_{n1}+\tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{12}\tilde{\beta}_{n2}), \tag{4.3}$$

where E_1 is the submatrix of E corresponding to the nonzero estimates. This leads to a sandwich formula for the covariance estimation:

$$\begin{aligned}\widehat{\mathrm{Cov}}(\widehat{\beta}_{n1}) &= \left(\tilde{\Omega}_{11}+\lambda E_1\right)^{-1}\tilde{\Omega}_{11}\widehat{\mathrm{Cov}}\left(\tilde{\beta}_{n1}+\tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{12}\tilde{\beta}_{n2}\right)\tilde{\Omega}_{11}\left(\tilde{\Omega}_{11}+\lambda E_1\right)^{-1} \\ &= \left(\tilde{\Omega}_{11}+\lambda E_1\right)^{-1}\tilde{\Omega}_{11}\left(\tilde{\Omega}_{11}+\lambda E_1\right)^{-1}/n. \end{aligned} \tag{4.4}$$

Remark 3. In theory, the optimal parameter λ in (4.4) goes to zero very fast, so the covariance estimator based on the sandwich formula is asymptotically equivalent to the asymptotic estimator given in (4.1). For finite samples, the sandwich estimator is generally smaller than the asymptotic estimator, due to the non-vanishing term λD_1. This pattern is also observed in our numerical studies presented in next section.

To tune the parameter λ, many selection criteria such as cross validation (CV), generalized cross validation (GCV), BIC and AIC selection can be used. Wang and Leng (2007) proved that the BIC criterion is consistent for the LSA estimator, i.e. the optimal λ chosen by the BIC can identify the true model with probability tending to one. Similarly, we can show that the BIC criterion for the PEE estimator is also consistent. Our empirical experience also suggests that the BIC gives the best performance for parameter tuning. So BIC is applied for parameter tuning in all the following numerical examples. To be specific,

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

5. Numerical Studies

5.1. Simulation Examples

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 original estimating equation method (EE) of Chen et al. (2002). In addition, for the PH models, we also compare with the penalized partial likelihood (PPL) estimator proposed by Zhang and Lu (2007); for the PO models, we compare with the penalized marginal likelihood (PML) estimator of Lu and Zhang (2007). BIC is used for choosing the regularization parameter for each method.

We compare all the methods with regard to their overall mean squared error (MSE), point estimation accuracy, and the variable selection performance. Following Tibshirani (1997), we compute the MSE $\equiv (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T\Sigma_X(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ and report the average MSE over 500 simulations for each method. Here Σ_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. The numbers in parentheses are the standard errors. We also demonstrate and compare the performance of the proposed two formula for covariance estimation: the estimator (4.1) based on the asymptotic results and the sandwich formula (4.4).

The base design involves nine covariates ($Z_1, \ldots, Z_9$), which are marginally standard normal with pairwise correlation $\mathrm{corr}(z_j, z_k) = \rho^{|j-k|}$. A moderate correlation between covariates with $\rho = 0.5$ is considered. The true coefficients $\boldsymbol{\beta}_0 = (-1, -0.9, 0, 0, 0, -0.8, 0, 0, 0)'$. Censoring times are generated from the uniform distribution over $[0, c_0]$, where c_0 is chosen to get the desired censoring rate. We consider two censoring rates: 25% and 40%, and two sample sizes: $n = 100$ and $n = 200$.

Table 1 summarizes the model estimation and variable selection results for EE, PEE, and PPL for the PH model under four different settings. Overall, the PEE gives the smallest MSE in all the settings, showing substantial improvement over the original EE estimator, and the PPL is

slightly worse than the PEE. For example, when $n = 100$ and the censoring rate is 40%, their average MSEs are respectively: EE 0.277, PEE 0.143, and PPL 0.177. When $n = 200$ and the censoring rate is 25%, their average MSEs are respectively: EE 0.087, PEE 0.051, and PPL 0.053. With regard to variable selection, the PPL gives the model sizes closest to the truth 3, the PEE gives slightly larger sizes, and the EE always gives the full model. For example, when $n = 100$ and the censoring rate is 40%, their model sizes are respectively: PEE 3.620 and PPL 3.150. When $n = 200$ and the censoring rate is 25%, their model sizes are respectively: PEE 3.250 and PPL 3.034. Note that the PPL is based on the partial likelihood estimation and has the oracle property (Zhang and Lu, 2007), so it is expected to be asymptotically optimal. In this finite sample setting, we have observed that the new estimator PEE performs well and gives comparable results with the PPL. When the sample size n increases, all the methods demonstrate better performance. Table 2 summarizes the model estimation and variable selection results for EE, PEE, and PML for the PO model. Again, we observe that the PEE gives the smallest MSE in all the settings. Similar patterns are discovered as in the PH example; see details in Table 2.

In summary, we observe that the PEE estimate gives much better improvement than the original EE, in terms of both model estimation and variable selection. Compared with other likelihood based methods, the PEE also gives comparable results. The unique features of the PEE include: it can handle both the PH and PO models in one unified framework; the estimator is easy to compute; its entire solution path can be obtained by taking advantage of existing software LARS (Efron. et al., 2004).

To test the accuracy of the standard error formula proposed in Section 4, we compare the sample standard errors (SEs) with their estimates. In Table 3, we summarize the average estimated $\widehat{\mathrm{SE}}$ given by the asymptotic estimator (4.1), the average estimated $\widehat{\mathrm{SE}}_s$ given by the sandwich formula (4.4), and those from Monte Carlo simulations (SE), when $n = 200$ and the censored rate 25% and 40%, for both PH and PO models. The estimated standard errors of both methods are reasonably close to the sample standard errors. Overall, the asymptotic estimator (4.1) gives a better estimation than the sandwich formula. We also noted all the estimates tend to slightly under-estimate the actual Monte Carlo standard errors. This is mainly because these two formula are derived when either assuming a fixed λ or letting λ converge to zero quickly, which does not take into account the variability due to different λ's chosen across runs. Similar patterns were observed for shrinkage methods in other situations (e.g., Tibshirani, 1997;Zhang and Lu, 2007).

5.2. Primary Biliary Cirrhosis Data Analysis

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 µg/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 control, 2 treatment). We restrict our attention to the 276 observations without missing values. All seventeen variables are included in the model.

This data has been previously analyzed in literature with various estimation and variable selection methods. Tibshirani (1997) fitted the PH model with the stepwise selection and with the LASSO penalty based on the partial likelihood (PL) approach. Zhang and Lu (2007) further studied the PPL estimation with the SCAD and the adaptive LASSO penalty. We fit the PEE for the PH model and compare results with other methods. Table 4 summarizes the estimated coefficients and the standard errors for various models. We found that the PEE selects eight variables: *age*, *oed*,*bil*,*alb*, *cop*, *sgot*, *prot* and *stage*, which is the same set of variables chosen by the PPL and the stepwise selection. Figure 1 depicts the solution path of the PEE estimator.

5.3. Lung Cancer Data Analysis

The data comes from the Veteran's Administration lung cancer trial (Prentice, 1973). In this trial, 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).

This data set has been analyzed by many authors. It was found that the proportional hazards model may not fit the data well. For example, Bagdonavicius et al. (2003) considered the generalized linear proportional hazards (GLPH) model (Bagdonavicius and Nikulin, 2002), a natural alternative to the proportional hazards model. Their method rejected the proportional hazards model in the favor of the GLPH model. In addition, Lam and Kuk (2001), fitted the proportional odds model to a subset of the data of 97 patients with no prior therapy based on the marginal likelihood approach, and Chen et al. (2002) fitted the linear transformation model to the same subset of data using the martingale based estimating equations. Only two variables Cell type and Karnofsky score were included in their analysis. They concluded that both Cell type and Karnofsky score are significant.

For variable selection, Lu and Zhang (2007) fitted the PO model with all the covariates, using the penalized marginal likelihood (PML) with the LASSO and the adaptive LASSO penalty. Here we fit the same model with the PEE approach and BIC is used for parameter tuning. Table 5 summarizes the estimated coefficients and their standard errors by different methods. We see that both the PEE and the PML select Cell type (small vs large, adeno vs large) and Karnofsky score as important variables. This result is in good agreement with Lam and Kuk (2001) and Chen et al. (2002). The bottom plot of Figure 1 depicts the solution path of the PEE estimator, obtained by fitting the LARS package (Efron et. al. 2004) in R.

5.4. Microarray Data (DLBCL) Analysis

We now apply the PEE method to the high dimensional microarray gene expression data of Rosenwald et al. (2002). The data consists of 240 diffuse large B-cell lymphoma (DLBCL) patients, and the expressions of 7, 399 genes for each patient. Patients' survival times were recorded, and among them, 138 patients died during the follow-up method. There are two purposes for this study; firstly, to predict patients' survival time using gene expression information; secondly, to identify important genes contributing to survival outcomes. This data was analyzed by Li and Luan (2005). For data of such high dimensionality, a common practice is to first conduct a preliminary gene filtering based on some univariate analysis, and then apply a more sophisticated model-based analysis. Following Li and Luan (2005), we concentrate on the top 50 genes selected using univariate Cox score.

The data are randomly divided into two sets: the first 160 patients for the training set and the remaining 80 patients for the testing set. The PH model is assumed. We apply both the PEE and the PPL, and BIC is used for parameter tunings. The PEE selects totally 20 genes and the PPL selects 13 genes. We notice that 9 out of 13 genes selected by PPL are also identified by the PEE. To further confirm the contribution of the selected genes by the PEE, we also evaluate the prediction performance of the PH model built with the training set on both the training and the testing data sets. Figure 2 shows that the Kaplan-Meier estimates of survival functions for the high-risk and low-risk groups of patients, defined by the predicted risk scores. The cut-off value was determined by the median of the estimated scores from the training set, and the same cutoff was applied to the testing data. It is seen that the model both fits the training data and predicts the testing data pretty well, achieving a good separation of the two-risk groups. The log-rank test of differences between two survival curves gives *p*-values of 0 and 0.0384 for the training and testing data, respectively.

6. Discussion

The class of semiparametric linear transformation models has become more popular due to its high flexibility. In this paper, we have proposed a method to improve upon the martingale equations based estimation procedure of Chen et al. (2002) and achieve sparse estimation. It was shown that the new estimator achieves a higher efficiency than the estimator of Chen et al. (2002). The numerical results also demonstrate the competitive performance of the new estimator for both variable selection and model estimation.

The proposed penalized estimating equation estimator was constructed based on a set of estimating equations, i.e. the martingale difference equation for the unknown transformation function and the martingale integral equation for the regression parameters as in Chen et al. (2002). As a consequence, the estimator of the regression parameters is consistent and asymptotically normal but in general not efficient. In the two listed papers (Dabrowska, 2005; 2006), a general class of M-estimators for the semiparametric transformation models was considered. This class also includes a special choice of the score equation corresponding to an asymptotically efficient estimator of the regression parameters. Actually, the martingale estimating equation based estimator considered in this paper is a special case of the general class of M-estimators. Therefore, to construct more efficient estimators, it is possible to construct the loss function based on the score equations for the general class of M-estimators. However, the corresponding computation can be much more intensive than the estimating equations considered in this paper. This is an interesting problem which deserves further investigation.

APPENDIX. PROOFS OF THEOREMS

LEMMA 1: Under the regularity conditions (*c*1) – (*c*6), we have

$$\frac{1}{n}\frac{\partial U_n(\beta)}{\partial \beta}|_{\beta=\beta_0} = A(\beta_0, H_0) + o_p(1),$$
$$\frac{1}{n}\frac{\partial U_n(\beta)}{\partial \beta}|_{\beta=\tilde{\beta}_n} = \frac{1}{n}\tilde{A}_n\{\tilde{\beta}_n, \tilde{H}(\cdot;\tilde{\beta}_n)\} + o_p(1),$$

where A and $\tilde{A}_n$ are given in Chen et al. (2002). Since the proof is similar as Chen et al. (2002), we omit it here.

Proof of Theorem 1. Recall that

$$Q_n(\beta)=D_n(\beta)+n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j|, \tag{A.1}$$

where $D_n(\beta)=U_n'(\beta)\tilde{V}_n^{-1}U_n(\beta)$. It is sufficient to show that (A.1) has a $\sqrt{n}-$consistent local minimizer. Following Fan and Li (2002), we only need to show that, for any arbitrarily small $\varepsilon > 0$, there exists a sufficiently large constant C such that

$$\liminf_n P\left\{\inf_{\|\mathbf{r}\|\geq C} Q_n(\beta_0+n^{-1/2}\mathbf{r})>Q_n(\beta_0)\right\}\geq 1-\varepsilon, \tag{A.2}$$

where $\mathbf{r} = (r_1,\cdots,r_d)$. Lemma 1 suggests that $U_n(\boldsymbol{\beta})$ has the following asymptotic representation

$$\frac{1}{n}U_n(\beta)=\left[\frac{1}{n}\tilde{A}_n\{\tilde{\beta}_n,\tilde{H}(\tilde{\beta}_n)\}\right](\beta-\tilde{\beta}_n)+(\beta-\tilde{\beta}_n)\cdot o_p(1).$$

Then we have

$$\begin{aligned} D_n(\beta) &= U_n'(\beta)\tilde{V}_n^{-1}U_n(\beta)\\ &=\left(\{\tilde{A}_n+no_p(1)\}(\beta-\tilde{\beta}_n)\right)'\tilde{V}_n^{-1}\left(\{\tilde{A}_n+no_p(1)\}(\beta-\tilde{\beta}_n)\right)\\ &=(\beta-\tilde{\beta}_n)'\{\tilde{A}_n+no_p(1)\}'\tilde{V}_n^{-1}\{\tilde{A}_n+no_p(1)\}(\beta-\tilde{\beta}_n)\end{aligned}$$

Thus, for any vector $\mathbf{r}$ we have

$$\begin{aligned} &D_n(\beta_0+n^{-1/2}\mathbf{r})-D_n(\beta_0)\\ &=\mathbf{r}'\{n^{-1}\tilde{A}_n+o_p(1)\}'\tilde{V}_n^{-1}\{\tilde{A}_n+no_p(1)\}\mathbf{r}+2\mathbf{r}'\{n^{-1}\tilde{A}_n+o_p(1)\}'\tilde{V}_n^{-1}\{\tilde{A}_n+no_p(1)\}\sqrt{n}(\beta_0-\tilde{\beta}_n)\\ &=\mathbf{r}'\{\tilde{\Sigma}_n^{-1}+o_p(1)\}\mathbf{r}+2\mathbf{r}'\{\tilde{\Sigma}_n^{-1}+o_p(1)\}\sqrt{n}(\beta_0-\tilde{\beta}_n).\end{aligned} \tag{A.3}$$

In addition, the penalty term can be bounded as

$$n\lambda\sum_{j=1}^{d}|\beta_{j0}+n^{-1/2}r_j|/|\tilde{\beta}_j|-n\lambda\sum_{j=1}^{d}|\beta_{j0}|/|\tilde{\beta}_j|\geq n\lambda\sum_{j=1}^{q}(|\beta_{j0}+n^{-1/2}r_j|-|\beta_{j0}|)/|\tilde{\beta}_j|\geq-\sqrt{n}\lambda\sum_{j=1}^{q}|r_j|/|\tilde{\beta}_j|. \tag{A.4}$$

Combining (A.3) and (A.4), we have

$$Q_n(\beta_0+n^{-1/2}\mathbf{r})-Q_n(\beta_0)\geq\mathbf{r}'\{\tilde{\Sigma}_n^{-1}+o_p(1)\}\mathbf{r}+2\mathbf{r}'\{\tilde{\Sigma}_n^{-1}+o_p(1)\}\left[\sqrt{n}(\beta_0-\tilde{\beta}_n)\right]-\sqrt{n}\lambda\sum_{j=1}^{q}|r_j|/|\tilde{\beta}_j|. \tag{A.5}$$

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

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

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

$$\sqrt{n}\lambda\sum_{j=1}^{q}|r_j|/|\tilde{\beta}_j|=\sqrt{n}\lambda\sum_{j=1}^{q}\left\{\frac{|r_j|}{|\beta_{j0}|}+\frac{|r_j|}{\sqrt{n}}O_p(1)\right\}\le\|\mathbf{r}\|\sqrt{n}\lambda O_p(1)=\|\mathbf{r}\|\cdot O_p(1).$$

Let $\nu_*(M)$ refers the minimal eigenvalue of M. Recall that $\|\mathbf{r}\|\ge C$. In (A.5), the first term is uniformly larger than $\nu_*(\tilde{\Sigma}_n^{-1})C^2\to_p\nu_*(\Sigma^{-1})C^2$. So, with the probability tending to one, the first term in (A.5) is uniformly larger than $0.5\nu_*(\Sigma^{-1})C^2$, which is quadratic in C. Furthermore, the second term in (A.5) is uniformly bounded by $C\|\tilde{\Sigma}_n^{-1}\sqrt{n}(\beta_0-\tilde{\beta}_n)\|$, which is linear in C with the coefficient $\|\tilde{\Sigma}_n^{-1}\sqrt{n}(\beta_0-\tilde{\beta}_n)\|=O_p(1)$. Therefore, as long as C is sufficiently large, the first term in (A.5) always dominates the other two terms with arbitrarily large probability. Therefore (A.2) holds and it completes the proof.

Proof of Theorem 2. We will show the sparsity of the PEE estimator, i.e., $\hat{\boldsymbol{\beta}}_{n2}=\mathbf{0}$ with probability one as $n\to\infty$. It is sufficient to show that for any sequence $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\|=O_p(n^{-1/2})$ and any constant C,

$$Q_n(\beta_1,\mathbf{0})=\min_{\|\beta_2\|\le Cn^{-1/2}}Q_n(\beta_1,\beta_2).$$

For any $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\|=O_p(n^{-1/2})$, we will show that, $\partial Q(\boldsymbol{\beta})/\partial\beta_j$ and β_j have the same sign for $\beta_j\in(-C\,n^{-1/2},C\,n^{-1/2})$ for $j=q+1,\cdots,d$, with probability tending to 1. For each $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$, by Lemma 1, we have the following asymptotic representations

$$\begin{aligned}\tfrac{1}{n}U_n(\beta)&=\tfrac{1}{n}U_n(\beta_0)+A\{\beta_0,\tilde{H}(\cdot;\beta_0)\}(\beta-\beta_0)+(\beta-\beta_0)\cdot o_p(1).\\ D_n(\beta)&=(\beta-\beta_0)'\{nA+o_p(n)\}'\widehat{V}_n^{-1}\{nA+o_p(n)\}(\beta-\beta_0),\end{aligned}$$

which lead to

$$\begin{aligned}\tfrac{\partial D_n}{\partial\beta}&=2\{A+o_p(1)\}'(n\tilde{V}_n^{-1})\{A+o_p(1)\}n(\beta-\beta_0)\\&=2\{A+o_p(1)\}\{V^{-1}+o_p(1)\}\{A+o_p(1)\}n(\beta-\beta_0),\end{aligned}$$

Thus, for $j=q+1,\cdots,d$, we have

$$\frac{\partial Q_n(\beta)}{\partial\beta_j}=\frac{\partial D_n(\beta)}{\partial\beta_j}+n\lambda_n\frac{\mathrm{sign}(\beta_j)}{|\tilde{\beta}_j|}=O_p(n^{1/2})+(n\lambda_n)n^{1/2}\frac{\mathrm{sign}(\beta_j)}{|n^{1/2}\tilde{\beta}_j|}.$$

Note that $n^{1/2}(\tilde{\beta}_j-0)=O_p(1)$, we have

$$\frac{\partial Q_n(\beta)}{\partial\beta_j}=n^{1/2}\left\{O_p(1)+n\lambda_n\frac{\mathrm{sign}(\beta_j)}{|O_p(1)|}\right\}. \tag{A.6}$$

Since $n\lambda_n\to\infty$, the sign of $\dfrac{\partial Q_n(\beta_j)}{\partial\beta_j}$ in (A.6) is completely determined by the sign of β_j when n is large, and they always have the same sign.

Proof of Theorems 3. According to Theorem 2, with probability tending to one, $\widehat{\beta}_{n2}=\mathbf{0}$, so $(\widehat{\beta}_{n1}'=\mathbf{0}')'$ must be the global minimizer of the objective function

$$\begin{aligned} Q_n(\beta) &= U_n'(\beta)\tilde{V}_n^{-1}U_n(\beta)+n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j| \\ &=(\beta-\tilde{\beta}_n)'\{\tilde{A}_n+no_p(1)\}'\tilde{V}_n^{-1}\{\tilde{A}_n+no_p(1)\}(\beta-\tilde{\beta}_n)+n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j| \\ &=n(\beta-\tilde{\beta}_n)'\{\tilde{\Sigma}_n^{-1}+o_p(1)\}(\beta-\tilde{\beta}_n)+n\lambda\sum_{j=1}^{d}|\beta_j|/|\tilde{\beta}_j|. \end{aligned}$$

Note that $\tilde{\Sigma}_n^{-1}=\tilde{\Omega}_n=\begin{bmatrix}\tilde{\Omega}_{11} & \tilde{\Omega}_{12}\\ \tilde{\Omega}_{21} & \tilde{\Omega}_{22}\end{bmatrix}$. Then $\hat{\boldsymbol{\beta}}_{n1}$ is the minimizer of

$$Q_n^0(\beta_{n1})=n(\beta_{n1}-\tilde{\beta}_{n1})'\{\tilde{\Omega}_{11}+o_p(1)\}(\beta_{n1}-\tilde{\beta}_{n1})-2n(\beta_{n1}-\tilde{\beta}_{n1})'\{\tilde{\Omega}_{12}+o_p(1)\}\tilde{\beta}_{n2}+n\tilde{\beta}_{n2}'\{\tilde{\Omega}_{22}+o_p(1)\}\tilde{\beta}_{n2}+n\lambda\sum_{j=1}^{q}\frac{|\beta_j|}{|\tilde{\beta}_j|}.$$

Therefore, we have the following normal equation

$$0=\frac{1}{2}\frac{\partial Q_n^0(\beta_{n1})}{\partial\beta_{n1}}\Big|_{\beta_{n1}=\widehat{\beta}_{n1}}=n\{\tilde{\Omega}_{11}+o_p(1)\}\left(\widehat{\beta}_{n1}-\tilde{\beta}_{n1}\right)-n\{\tilde{\Omega}_{12}+o_p(1)\}\tilde{\beta}_{n2}+nG(\widehat{\beta}_{n1}), \tag{A.7}$$

where $G(\hat{\boldsymbol{\beta}}_{n1})=(0.5\lambda\ \mathrm{sign}(\hat{\beta}_1)/|\tilde{\beta}_1|,\cdots,0.5\lambda\ \mathrm{sign}(\hat{\beta}_q)/|\tilde{\beta}_q|)'$. Using the theorem's condition $\sqrt{n}\lambda\to 0$, for each component in $\sqrt{n}G(\widehat{\beta}_{n1})$, we have

$$0.5\sqrt{n}\lambda\ \mathrm{sign}(\widehat{\beta}_j)/|\tilde{\beta}_j|=o_p(1),\ 1\le j\le q.$$

Then (A.7) implies that

$$\begin{aligned} \sqrt{n}(\widehat{\beta}_{n1}-\beta_{01}) &= \sqrt{n}(\tilde{\beta}_{n1}-\beta_{01})+\tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{12}(\sqrt{n}\tilde{\beta}_{n2})-\tilde{\Omega}_{11}^{-1}\sqrt{n}G(\tilde{\beta}_{n1})+o_p(1) \\ &= \sqrt{n}(\tilde{\beta}_{n1}-\beta_{01})+\tilde{\Omega}_{11}^{-1}\tilde{\Omega}_{12}(\sqrt{n}\tilde{\beta}_{n2})+o_p(1), \end{aligned} \tag{A.8}$$

which converges in distribution to normal with mean $\mathbf{0}$ and variance-covariance matrix

$$\Sigma_{11}+2\Omega_{11}^{-1}\Omega_{12}\Sigma_{21}+\Omega_{11}^{-1}\Omega_{12}\Sigma_{22}\Omega_{21}\Omega_{11}^{-1}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

since $\Sigma^{-1}=\Omega\equiv\begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}=\lim_{n\to\infty}\tilde{\Omega}_n$ and $\Omega_{11}^{-1}\Omega_{12}=-\Sigma_{12}\Sigma_{22}^{-1}$.

REFERENCES

Andersen, PK.; Borgan, O.; Gill, R.; Keiding, N. Statistical Models Based on Counting Processes. New York: Springer; 1993.

Bagdonavicius, V.; Nikulin, M. Monographs on Statistics and Applied Probability 94. Chapman and Hall/CRC; 2002. Accelerated life models: modeling and statistical analysis.

Bagdonavicius V, Hafdi MA, Himdi KE, Nikulin M. Statistical analysis of the generalized linear proportional hazards model. Journal of Mathematical Sciences 2003;127:1673–1681.

Bennett S. Analysis of survival data by the proportional odds model. Statistics in Medicine 1983;2:273–277. [PubMed: 6648142]

Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins University Press; 1993.

Breiman L. Heuristics of instability and stabilization in model selection. Annals of Statistics 1996;24:2350–2383.

Chen K, Jin Z, Ying Z. Semiparametric of transformation models with censored data. Biometrika 2002;89:659–668.

Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika 1995;82:835–845.

Cox DR. Regression models and life tables (with Discussion). Journal of the Royal Statistical Society, Series B 1972;34:187–220.

Cox DR. Partial Likelihood. Biometrika 1975;62:269–276.

Dabrowska DM. Quantile regression in transformation models. Sankhya 2005;67:153–187.

Dabrowska, DM. Estimation in a class of semiparametric transformation models; IMS Lecture Notes-Monograph Series, 2nd Lehmann Symposium-Optimality; 2006. p. 131-169.

Dabrowska DM, Doksum KA. Estimation and testing in the two-sample generalized odds rate model. Journal of American Statistical Association 1988;83:744–749.

Dickson E, Grambsch P, Fleming T, Fisher L, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. Hepatology 1989;10:1–7. [PubMed: 2737595]

Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. Annals of Statistics 2004;32:407–451.

Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Annals of Statistics 2002;30:74–99.

Fine J, Ying Z, Wei LJ. On the linear transformation model for censored data. Biometrika 1998;85:980–986.

Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. New York: Wiley; 1991.

Fu WJ. Penalized estimating equations. Biometrics 2003;59:126–132. [PubMed: 12762449]

Fu WJ. Penalized regression: the bridge versus the lasso. Journal of Computational and Graphical Statistics 1998;7:397–416.

Lam KF, Kuk YC. A marginal likelihood approach to estimation in frailty models. Journal of American Statistical Association 1997;92:985–990.

Le Cam, L. On the asymptotic theory of estimation and testing hypotheses; Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability; 1956. p. 129-156.

Le Cam, L.; Yang, GL. Asymptotics in Statistics. New York: Springer Verlag; 1990.

Li H, Luan Y. Boosting proportional hazards models using smoothing spline, with application to high-dimensional microarray data. Bioinformatics 2005;21:2403–2409. [PubMed: 15713732]

Lu W, Zhang HH. Variable selection for proportional odds model. Statistics in Medicine 2007;26:3771–3781. [PubMed: 17266170]

Murphy SA, Rossini AJ, van der Vaart AW. Maximum likelihood estimation in the proportional odds model. Journal of American Statistical Association 1997;92:968–976.

Osborne MR, Presnell B, Turlach BA. On the LASSO and its dual. Journal of Computational and Graphical Statistics 2000;9:319–337.

Pettitt AN. Inference for the linear model using a likelihood based on ranks. Journal of the Royal Statistical Society, Series B 1982;44:234–243.

Pettitt AN. Proportional odds model for survival data and estimates using ranks. Applied Statistics 1984;33:169–175.

Prentice RL. Exponential survivals with censoring and explanatory variables. Biometrika 1973;60:279–288.

Qu A, Li R. Quadratic inference functions for varying-coefficient models with longitudinal data. Biometrics 2006;62:379–391. [PubMed: 16918902]

Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RR, Muller-Hermelink HK, Smeland EB, Staudt LM, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. New England Journal of Medicine 2002;346:1937–1947. [PubMed: 12075054]

Therneau, TM.; Grambsch, PM. Modeling survival data: extending the Cox model. New York: Springer; 2000.

Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine 1997;16:385–395. [PubMed: 9044528]

van der Vaart, AW. Asymptotic statistics. New York: Cambridge University Press; 1998.

Wang H, Leng C. Unified LASSO estimation with least squares approximation. Journal of American Statistical Association 2007;102:1039–1048.

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

Wang H, Li G, Tsai CL. Regression coefficients and autoregressive order shrinkage and selection via lasso. Journal of Royal Statistical Society, B 2007;69:63–78.

Zhang HH, Lu W. Adaptive-LASSO for Cox's proportional hazards model. Biometrika 2007;94:691–703.

Zeng D, Lin DY. Maximum likelihood estimation in semiparametric transformation models for counting processes. Biometrika 2006;93:627–640.

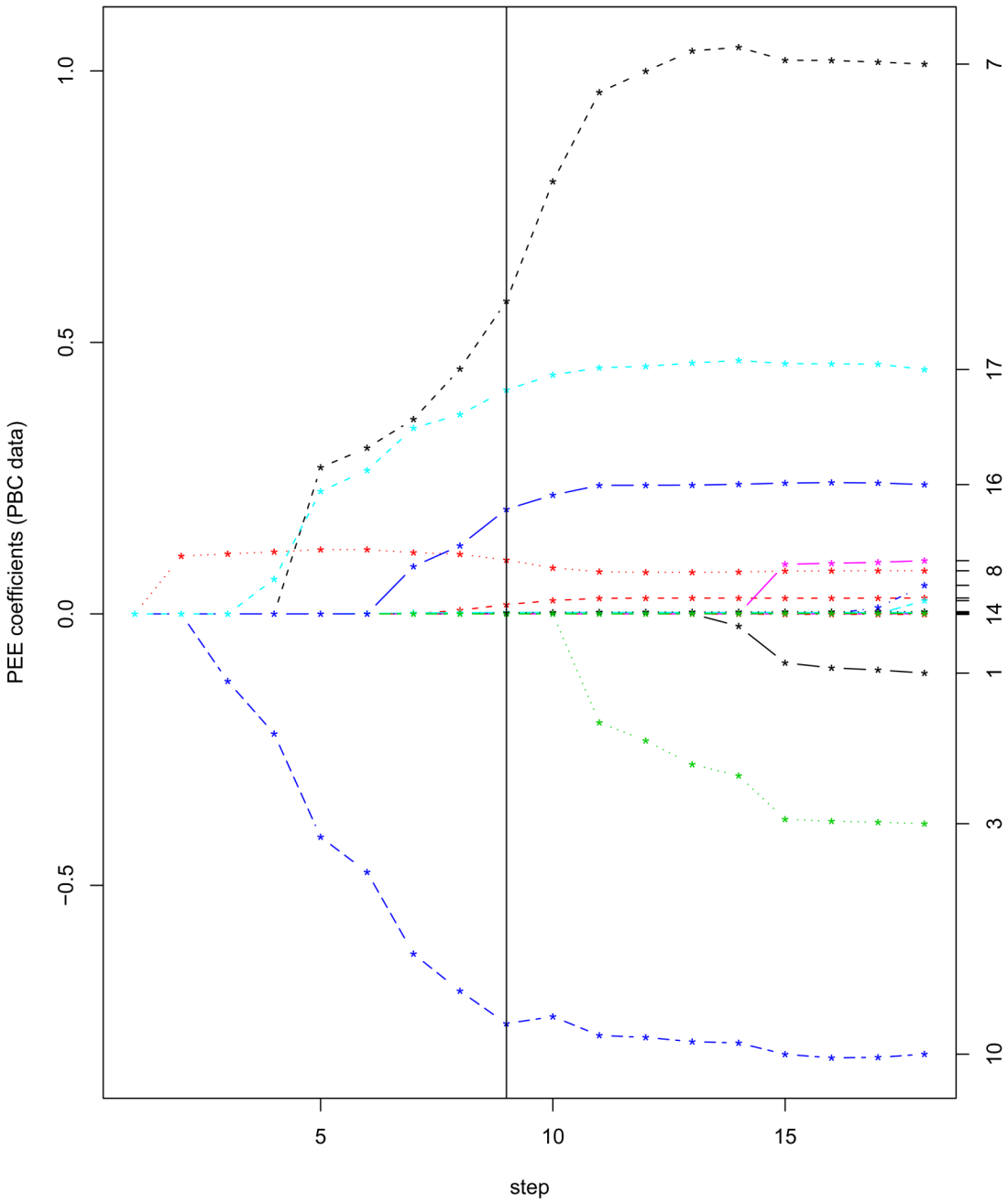
Zeng D, Lin DY. Semiparametric transformation models with random effects for recurrent events. Journal of American Statistical Association 2007;102:167–180.

Zou H. The adaptive lasso and its oracle properties. Journal of American Statistical Association 2006;101:1418–1429.

Zou H. A note on path-based variable selection in the penalized proportional hazards model. Biometrika 2008;95:241–247.

Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. Annals of Statistics 2007;35:2173–2192.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Annals of Statistics 2008;36:1509–1533. [PubMed: 19823597]



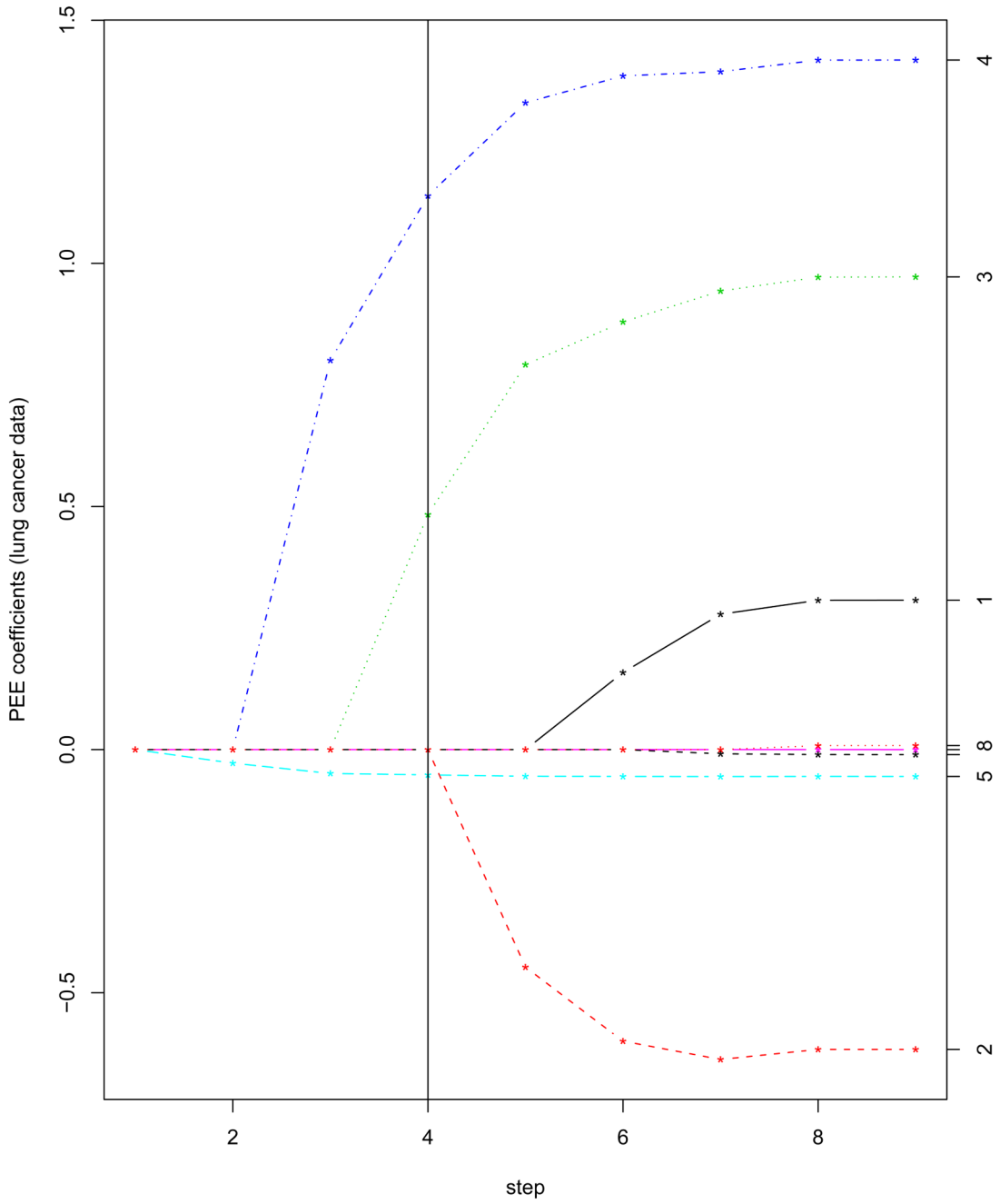


Figure 6.1.
The first plot shows the PEE solution path for PBC data fitted with PH model, and the second plot for lung cancer data fitted with PO model. The solid vertical line denotes the PEE estimates tuned with the BIC criterion.

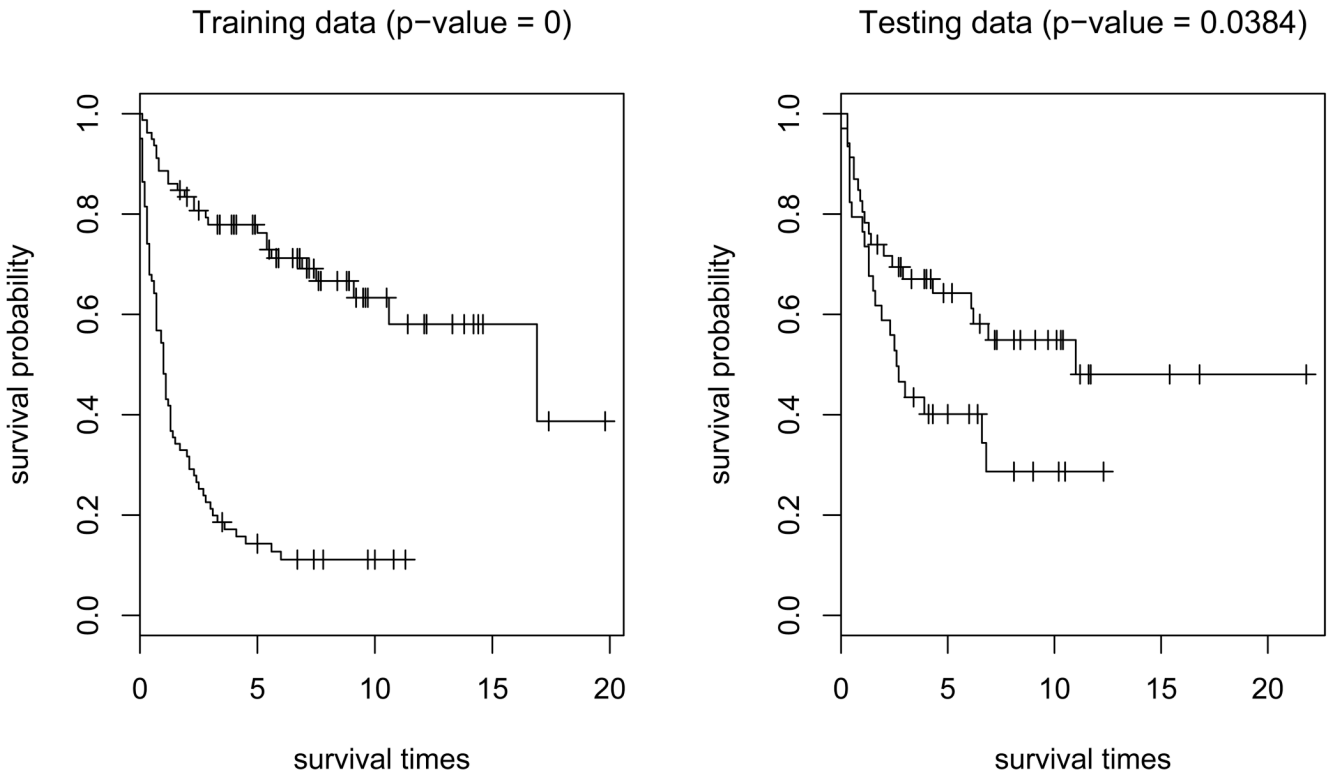


Figure 6.2.
Kaplan-Meier estimates of survival curves for high-risk and low-risk groups of patients using the selected genes by the PEE.

Table 6.1

Model estimation and variable selection 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)
		PEE	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)
		PEE	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)
		PEE	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)
		PEE	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)

PH stands for proportional hazards model.

EE stands for the estimation equation estimate.

PEE stands for the PEE estimate obtained with BIC.

PPL stands for the penalized partial likelihood with ALASSO penalty (Zhang and Lu, 2007).

Table 6.2

Model estimation and variable selection 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)
		PEE	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)
		PEE	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)
		PEE	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)
		PEE	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)

PO stands for proportional odds model.
EE stands for the estimation equation estimate.
PEE stands for the PEE estimate obtained with BIC.
PML stands for the penalized partial likelihood with ALASSO penalty (Lu and Zhang, 2007).

Table 6.3

Estimated and MC standard errors for the PEE nonzero estimates (n = 200).

		β_1			β_2			β_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

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

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

$\widehat{\text{SE}}$ stands for the average of estimated standard error based on (4.1).

$\widehat{\text{SE}}_S$ stands for the average of estimated standard error based on the sandwich formula (4.4).

Table 6.4

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

Covariate	EE	PEE	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)

PH stands for proportional hazards model.
EE stands for the estimation equation estimate.
PEE stands for the PEE estimate obtained with BIC.
PPL stands for the penalized partial likelihood with ALASSO penalty (Zhang and Lu, 2007).

Table 6.5

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

Covariate	EE	PEE	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 (−)

PO stands for proportional odds model.
EE stands for the estimation equation estimate.
PEE stands for the PEE estimate obtained with BIC.
PML stands for the penalized marginal likelihood with ALASSO penalty (Lu and Zhang, 2007).