DEPARTMENT OF STATISTICS

North Carolina State University

2501 Founders Drive, Campus Box 8203

Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2580

Variable Selection for Linear Transformation Models via Penalized Marginal Likelihood

Wenbin Lu

Hao Helen Zhang

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

lu@stat.ncsu.edu, hzhang2@stat.ncsu.edu

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

VARIABLE SELECTION FOR LINEAR TRANSFORMATION MODELS VIA PENALIZED MARGINAL LIKELIHOOD

WENBIN LU and HAO H. ZHANG

Department of Statistics, North Carolina State University

ABSTRACT. We study the problem of variable selection for linear transformation models, a class of general semiparametric models for censored survival data. The penalized marginal likelihood methods with shrinkage-type penalties are proposed to automate variable selection in linear transformation models; we consider the LASSO penalty and propose a new penalty called the adaptive-LASSO (ALASSO). Unlike the LASSO, the ALASSO imposes different penalties on different coefficients: unimportant covariates receive larger penalties than important ones. In this way, important variables can be protectively preserved in models while unimportant ones are more likely to be shrunk to zeros. An efficient iterative algorithm is proposed for optimization. The performance of both penalties is illustrated through simulated examples and one real data, the Veteran's Administration lung cancer data. In terms of both variable selection and coefficient estimation, we find that both shrinkage estimators outperform the maximum marginal likelihood estimator, and the ALASSO gives better performance than the LASSO.

Key words: Adaptive LASSO; Censored survival data; LASSO; Linear transformation model; Marginal likelihood; Variable selection.

1. Introduction

One main issue in survival analysis is to study the dependence of the survival time T of patients on various clinical covariates $\mathbf{Z} = (Z_1, \cdots, Z_p)$. Though the proportional hazards model (Cox, 1972) has been widely used in survival data analysis, it may not be an appropriate choice when homogeneity between different groups increases with time. For example, if the hazard functions for two treatment groups converge to the same limit, the proportional odds model is preferable to the proportional hazards model for such data (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 $\mathbf{Z}$ be the survival time, the censoring time, and the $p \times 1$ covariate vector. A linear transformation model assumes

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

where H is an unknown monotone increasing function, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the regression parameter vector, and ϵ has a completely known continuous distribution that is independent of C and $\mathbf{Z}$. Let $\Lambda(x)$ denote the cumulative hazard function of ϵ, so that $P(\epsilon > x) = \exp\{-\Lambda(x)\}$. If ϵ follows the extreme value distribution, i.e. $\Lambda(x) = \exp(x)$, then (1) becomes the proportional hazards model. If ϵ has the standard logistic distribution, i.e. $\Lambda(x) = \log\{1 + \exp(x)\}$, then (1) represents the proportional odds model. When there is no censoring and ϵ follows the standard normal distribution, (1) generalizes the usual Box-Cox transformation models.

Variable selection is fundamental to survival modeling, and it refers to the process of selecting from $Z_1, \cdots, Z_p$ those that are most associated or predictive of the survival time. An effective variable selection always leads to better risk assessment and model

interpretation. The nature of censoring makes variable selection challenging in survival data analysis. Classical methods like stepwise selection procedures can be expensive in computation and often 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, 1996, 1997) and the SCAD (Fan and Li, 2001, 2002). However, very little work has been done for variable selection in linear transformational models. For linear transformation models, the partial likelihood function is not available. And most estimation procedures for the regression coefficients are based on estimating equations (Cheng et al., 1995; Fine et al., 1998; Chen et al., 2002), which are not convenient to incorporate the shrinkage penalty. In this paper, we propose the penalized marginal likelihood methods for variable selection in linear transformation models. The marginal likelihood was also used by Lam and Kuk (1997) for the frailty model and by Lam and Leung (2001) for the proportional odds model.

In the proportional hazards model, the LASSO penalty tends to produce sparse solutions and has given good results (Tibshirani, 1997). Therefore we first extend the LASSO to linear transformation models and investigate its performance in that context. Then we propose a new modified LASSO penalty, called the adaptive LASSO (ALASSO) which minimizes the penalized marginal likelihood function with a weighted-L_1 penalty. The weights are data-driven such that different weights are imposed on coefficients according to their relative importance. The remainder of this article is organized as follows. Section 2 introduces the marginal likelihood function estimation for linear transformation models. In Section 3, we present the penalized likelihood methods with the LASSO and ALASSO penalties. An algorithm is proposed for optimizing the likelihood functions. In Section 4, we discuss the choice of tuning parameters with the

generalized cross validation (GCV) score. We also derive the sandwich-type formula to estimate the standard errors of our estimates. Section 5 is devoted to simulation studies and one application to a real data set. Some final remarks are given in Section 6.

2. Marginal Likelihood for Linear Transformation Models

Suppose a random sample of n individuals is chosen. Let T_i and C_i be respectively the failure time and censoring time of subject i ($i = 1, \cdots, n$). For the ith individual, we define $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, Z_{ij} the jth covariate value, and $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ip})^T$ the p-dimensional vector of covariates. Throughout the paper, we focus on the standard situation where $p < n$. The observations consist of $(\tilde{T}_i, \delta_i, \mathbf{Z}_i), i = 1, \ldots, n$, which are independent copies of $(\tilde{T}, \delta, \mathbf{Z})$. Assume that the distribution of the failure time T_i is specified by the linear transformation model (1). The likelihood function of the observed data is then given by

$$L_n(\boldsymbol{\beta}, H) = \prod_{i=1}^{n} [h(\tilde{T}_i)\lambda\{H(\tilde{T}_i) + \boldsymbol{\beta}'\mathbf{Z}_i\}]^{\delta_i} e^{-\Lambda\{H(\tilde{T}_i)+\boldsymbol{\beta}'\mathbf{Z}_i\}}, \tag{2}$$

where $\lambda(x) = d\Lambda(x)/dx$ and $h(x) = dH(x)/dx$. Since L_n involves the nonparametric function H which is infinite dimensional, it is difficult to penalize L_n directly for selecting important variables. To get rid of the nonparametric function H in the likelihood function (2), we choose to use the marginal likelihood function (Lam and Kuk, 1997; Lam and Leung, 2001). To be specific, let $T_{(1)} < \cdots < T_{(K)}$ denote the ordered uncensored failure times in the sample and define $T_{(0)} = 0$, $T_{(K+1)} = \infty$. For $0 \leq k \leq K$, let $\mathcal{L}_k$ denote the set of labels i corresponding to those observations censored in the interval $[T_{(k)}, T_{(k+1)})$. Due to the censoring scheme, the complete ranks of T_i's are not observed. Let $\mathbf{R}$ denote the unobserved rank vector of T_i's and let $\mathcal{C}$ denote the collection of all

possible rank vectors of T_i's consistent with the observed data $(\tilde{T}_i, \delta_i)$ $(i = 1, \cdots, n)$. The marginal likelihood is then defined by $L_{n,M}(\boldsymbol{\beta}) = P(\mathbf{R} \in \mathcal{C})$, where the probability is with respect to the underlying uncensored version of the study. Furthermore, the event $\{\mathbf{R} \in \mathcal{C}\}$ can be characterized by the following domain:

$$D = \{T_{(k)}, T_i > T_{(k)} | i \in \mathcal{L}_k,\ k = 0, \cdots, K\}.$$

It follows that

$$L_{n,M}(\boldsymbol{\beta}) = \int \cdots \int_D \prod_{i=1}^{n} \lambda\{H(T_i) + \boldsymbol{\beta}'\mathbf{Z}_i\} e^{-\Lambda\{H(T_i)+\boldsymbol{\beta}'\mathbf{Z}_i\}} \prod_{i=1}^{n} dH(T_i). \tag{4}$$

As noted by Clayton and Cuzick (1985), the integral in (4) can be expressed as an integral over only the uncensored failure times $T_{(1)}, \cdots, T_{(K)}$, i.e.

$$L_{n,M}(\boldsymbol{\beta}) = \int \cdots \int_{T_{(1)}<\cdots<T_{(K)}} \prod_{i=1}^{n} [\lambda\{H(T_{(k_i)})+\boldsymbol{\beta}'\mathbf{Z}_i\}]^{\delta_i} e^{-\Lambda\{H(T_{(k_i)})+\boldsymbol{\beta}'\mathbf{Z}_i\}} \prod_{k=1}^{K} dH(T_{(k)}), \tag{5}$$

where $k_i = \max\{k : T_{(k)} \le \tilde{T}_i,\ 0 \le k \le K\}, i = 1, \cdots, n$. For any i with $k_i = 0$, we have $[\lambda\{H(T_{(0)}) + \boldsymbol{\beta}' Z_i\}]^{\delta_i} e^{-\Lambda\{H(T_{(0)})+\boldsymbol{\beta}' Z_i\}} \equiv 1$.

Define $V_i = H(T_i)$, $i = 1, \cdots, n$. Since $H(T)$ is a strictly monotone increasing function of T, the rank vector of V_i's is the same as that of T_i's. Then (5) can be rewritten as

$$L_{n,M}(\boldsymbol{\beta}) = \int \cdots \int_{V_{(1)}<\cdots<V_{(K)}} \prod_{i=1}^{n} \{\lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)\}^{\delta_i} e^{-\Lambda(V_{(k_i)}+\boldsymbol{\beta}'\mathbf{Z}_i)} \prod_{k=1}^{K} dV_{(k)}, \tag{6}$$

where $V_{(k)} = H(T_{(k)})$, $k = 1, \cdots, K$. Note that (6) is independent of the nonparametric function H, or it is baseline-free. In addition, when the model is the proportional hazards model, i.e. $\Lambda(x) = \exp(x)$, (6) becomes the partial likelihood (Kalbfleisch and Prentice, 2002). However, in general the integral in (6) has no analytical form and Monte Carlo method is needed to approximate (6). Towards this, we multiply and

divide the integrand in (6) by

$$c\prod_{i=1}^{n}\{\lambda(V_{(k_i)})\}^{\delta_i}e^{-\Lambda(V_{(k_i)})}, \tag{7}$$

where the constant c is the total number of possible rank vectors in $\mathcal{C}$. It can be shown that (7) is the density function of $V_{(1)},\cdots,V_{(K)}$ under progressive type II censoring (Lawless, 1982) when the underline V_i $(i=1,\cdots,n)$ are independent and identically distributed according to the distribution function $F(x)=1-\exp\{-\Lambda(x)\}$. Here, progressive type II censoring means that we remove l_k (the number of observations censored in the interval $[T_{(k)},T_{(k+1)})$) observations at random from the risk set immediately after removing the kth uncensored observation $V_{(k)}$, $k=0,\cdots,K$, with $V_{(0)}\equiv 0$.

Thus, the marginal likelihood (6) can be expressed as

$$L_{n,M}(\boldsymbol{\beta})=E\{Q(V_{(1)},\cdots,V_{(K)};\boldsymbol{\beta})\},$$

where the expectation is with respect to the density (7) and

$$Q(V_{(1)},\cdots,V_{(K)};\boldsymbol{\beta})=\frac{1}{c}\prod_{i=1}^{n}\frac{\{\lambda(V_{(k_i)}+\boldsymbol{\beta}'\mathbf{Z}_i)\}^{\delta_i}e^{-\Lambda(V_{(k_i)}+\boldsymbol{\beta}'\mathbf{Z}_i)}}{\{\lambda(V_{(k_i)})\}^{\delta_i}e^{-\Lambda(V_{(k_i)})}}. \tag{8}$$

Now we can use the important sampling technique to approximate $L_{n,M}$. Let $F^{-1}(x)$ denote the inverse function of $F(x)$. Then $L_{n,M}$ can be approximated by

$$\hat{L}_{n,M}(\boldsymbol{\beta})=\frac{1}{b}\sum_{b=1}^{B}Q\{F^{-1}(U^b_{(1)}),\cdots,F^{-1}(U^b_{(K)});\boldsymbol{\beta}\}, \tag{9}$$

where $U^b_{(1)},\cdots,U^b_{(K)}$, $b=1,\cdots,B$, represent B independent realizations of the uncensored order statistics of a random sample of size n from the uniform distribution under the above progressive type II censoring scheme.

As noted by Lam and Kuk (1997), the same set of $(U^b_{(1)},\cdots,U^b_{(K)})$ is used in (9) regardless of the values of $\boldsymbol{\beta}$, which saves computing time and implies that $\hat{l}_{n,M}(\boldsymbol{\beta})\equiv$

$\log\{\hat{L}_{n,M}(\boldsymbol{\beta})\}$ is a bona fide function of $\boldsymbol{\beta}$, whose first- and second-order derivatives can be obtained analytically. Thus, $\hat{l}_{n,M}(\boldsymbol{\beta})$ can be maximized by the usual Newton-Raphson algorithm; the estimator is called maximum marginal likelihood estimator (MMLE). In the next section, we propose a modified LASSO-type estimate for $\boldsymbol{\beta}$ based on $\hat{l}_{n,M}(\boldsymbol{\beta})$.

3. Adaptive LASSO for Linear Transformation Models

Recall that $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ip})^T$ represents the covariate vector for the ith individual, $i = 1, \cdots, n$. For $j = 1, \cdots, p$, we define $\mathbf{Z}^j = (Z_{1j}, \cdots, Z_{nj})^T$, the vector of n observations for the jth covariate. Throughout this paper, we assume that $\mathbf{Z}^j$'s are standardized, i.e.

$$\sum_{i=1}^{n} Z_{ij} = 0, \quad \sum_{i=1}^{n} Z_{ij}^2 = 1, \quad \text{for} \quad j = 1, \cdots, p.$$

The LASSO, proposed by Tibshirani (1996), is the penalized least squares estimates with the L_1 penalty in linear regression settings. Overall the LASSO estimates achieve a smaller mean squared error (MSE) than the ordinary least squares estimates. The LASSO penalty was also studied for Cox's proportional hazard model by Tibshirani (1997). For linear transformation models, we first consider the penalized marginal likelihood with the LASSO penalty

$$\min_{\boldsymbol{\beta}} -\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{10}$$

The nature of the L_1 penalty shrinks small coefficients to be exactly zeros and hence results in a sparse representation of the solution. Thus the LASSO does a kind of continuous subset selection. Here $\lambda \geq 0$ is a tuning parameter which controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. In Section 4, we suggest the use of generalized cross validation (GCV) to tune λ.

One drawback of the LASSO penalty is that the estimates for important covariates may suffer from substantial bias (Fan and Li, 2001). The reason is that the same penalty is applied to all the coefficients: larger values of λ give sparser solutions at the price of causing larger bias to nonzero coefficients. Therefore we propose a modified L_1 penalty, called the adaptive LASSO,

$$\min_{\boldsymbol{\beta}} -\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tau_j, \tag{11}$$

where the non-negative weights $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_p)^T$ are chosen adaptively by data. The τ's are regarded as leverage factors, which adjust penalties on the coefficients by taking large values for unimportant covariates and small values for important ones.

The choice of $\boldsymbol{\tau}$ in the ALASSO is important to assure good solutions. Let $\tilde{\boldsymbol{\beta}}$ be any consistent estimator of $\boldsymbol{\beta}$. Our empirical experience shows that $1/|\tilde{\beta}_j|$'s are proper choices for weights since their magnitudes reflect the relative importance of covariates. In Zhang and Lu (2006), we studied the theoretical properties of the ALASSO under Cox's proportional hazard models with the maximum partial likelihood estimates chosen as weights, and the findings justify their use in practice. Furthermore, due to the consistency of $\tilde{\beta}_j$, the term $|\beta_j|/|\tilde{\beta}_j|$ converges to $I(\beta_j \neq 0)$ in probability as n goes to infinity. This shows the ALASSO penalty is closely related to the L_0 penalty $\sum_{j=1}^{d} I(|\beta_j| \neq 0)$, also called the entropy penalty in wavelet literature (Donoho and Johnstone 1998; Antoniadis and Fan 2001). Therefore the ALASSO can be regarded as an automatic implementation of the best subset selection in some asymptotic sense.

In practice, consistent estimates are readily available in standard statistical models, such as the least squared estimates for ordinary linear models and the maximum likelihood estimates for parametric families. When the design matrix is ill-posed due to the collinearity among covariates, other robust estimates such as ridge estimates can

be used for weights. Denote the maximizer of the marginal likelihood function $\hat{l}_{n,M}(\boldsymbol{\beta})$ as $\tilde{\boldsymbol{\beta}}$. Because $\tilde{\boldsymbol{\beta}}$ are consistent estimates (Lam and Kuk, 1997), their absolute values reflect the relative importance of covariates. Therefore we use $\tau_j^{-1} = |\tilde{\beta}_j|$ and solve

$$\min_{\boldsymbol{\beta}} -\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j|. \tag{12}$$

If $\tilde{\beta}_j = 0$, we set the solution $\hat{\beta}_j = 0$. To solve (12), we use the Newton-Raphson procedure through the iterative least squares subject to the weighted L_1 penalty. Similar algorithms were also used by Tibshirani (1997) and Fan and Li (2002). Define the gradient $\nabla l(\boldsymbol{\beta}) = -\partial \hat{l}_{n,M}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and the Hessian matrix $\nabla^2 l(\boldsymbol{\beta}) = -\partial^2 \hat{l}_{n,M}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\boldsymbol{\beta}'$. Let X denote the Cholesky decomposition of $\nabla^2 l(\boldsymbol{\beta})$, i.e. $\nabla^2 l(\boldsymbol{\beta}) = X'X$, and set the pseudo response vector $\mathbf{Y} = (X')^{-1}(\nabla^2 l(\boldsymbol{\beta})\boldsymbol{\beta} - \nabla l(\boldsymbol{\beta}))$. By the second-order Taylor expansion, $-\hat{l}_{n,M}(\beta)$ can be approximated by the quadratic form $\frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta})$, and in each iterative step we minimize

$$\frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j|. \tag{13}$$

The problem (13) is different from the LASSO in that different weights are imposed on β_j's. For solving the standard LASSO, Tibshirani (1996) suggested an algorithm based on quadratic programming techniques, and Fu (1998) proposed the shooting algorithm. Recently Efron et al. (2004) showed that, under the least squares setting, the whole solution path of LASSO can be obtained by a modified Lars algorithm. In this paper, we have modified Fu's shooting algorithm to take into account different weights in the ALASSO. For any fixed λ, we suggest the following algorithm to solve (12).

1. Solve $\tilde{\boldsymbol{\beta}}$ by minimizing the unpenalized marginal likelihood $-\hat{l}_{n,M}(\boldsymbol{\beta})$.

2. Initialization: $k = 1$ and $\beta_j^{(1)} = 0$ for $j = 1, ..., p$.

3. Compute $\nabla l, \nabla^2 l, X, \mathbf{Y}$ based on the current value $\boldsymbol{\beta}^{(k)}$.

4. Minimize (13) using the modified shooting algorithm. Denote the solution as $\boldsymbol{\beta}^{(k+1)}$.

5. Let $k = k + 1$. Go back to step 3 until the convergence criterion meets.

4. Standard Errors and Parameter Tuning

4.1. Standard Errors

Assume the true parameter $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{02})'$, where $\boldsymbol{\beta}_{01}$ consists of nonzero components and $\boldsymbol{\beta}_{02}$ consists of zero components. Since the ALASSO estimators are non-linear and non-differentiable functions of the response values for any fixed λ, it is difficult to obtain an accurate estimate of their standard errors. We use the conventional technique in the likelihood setting and approximate the covariance matrix of the ALASSO estimators using the corresponding sandwich formula. Similar methods were used in Fan and Li (2002).

Since the ALASSO penalty function is singular at the origin, it does not have continuous second order derivatives. However, it can be approximated by a quadratic function. Given an initial value $\boldsymbol{\beta}^{(1)}$ that is close to the ALASSO minimizer, if $\beta_j^{(1)}$ is very close to zero, then set $\hat{\beta}_j = 0$. Otherwise the penalty function can be locally approximated by a quadratic function

$$|\beta_j| = \frac{1}{2}|\beta_j^{(1)}| + \frac{1}{2|\beta_j^{(1)}|}\beta_j^2.$$

We use the second-order Taylor expansion for the log marginal likelihood function

$$-\hat{l}_{n,M}(\boldsymbol{\beta}) = -\hat{l}_{n,M}(\boldsymbol{\beta}^{(1)}) + \nabla l(\boldsymbol{\beta}^{(1)})'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)})'\nabla^2 l(\boldsymbol{\beta}^{(1)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}).$$

Then (12) can be locally approximated (except for a constant term) by

$$-\hat{l}_{n,M}(\boldsymbol{\beta}^{(1)}) + \nabla l(\boldsymbol{\beta}^{(1)})'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)})'\nabla^2 l(\boldsymbol{\beta}^{(1)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}) + \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{2|\beta_j^{(1)}||\tilde{\beta}_j|}. \tag{14}$$

At the $(k+1)$th step, the solution in the Newton-Raphson algorithm is updated by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \left[\nabla^2 l(\hat{\boldsymbol{\beta}}^{(k)}) + \lambda A(\hat{\boldsymbol{\beta}}^{(k)})\right]^{-1} \left[\nabla l(\hat{\boldsymbol{\beta}}^{(k)}) + \lambda b(\hat{\boldsymbol{\beta}}^{(k)})\right], \tag{15}$$

where

$$\begin{aligned} A(\boldsymbol{\beta}) &= \operatorname{diag}\left\{\frac{1}{\beta_1^2}, ..., \frac{1}{\beta_p^2}\right\}, \\ b(\boldsymbol{\beta}) &= \left(\frac{\operatorname{sign}(|\beta_1|)}{|\tilde{\beta}_1|}, ..., \frac{\operatorname{sign}(|\beta_p|)}{|\tilde{\beta}_p|}\right)'. \end{aligned}$$

Then the corresponding sandwich formula for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1} \widehat{\operatorname{cov}}\left(\nabla l(\boldsymbol{\beta}_0) + \lambda b(\boldsymbol{\beta}_0)\right) \left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1}. \tag{15}$$

Since the MMLE estimate $\tilde{\boldsymbol{\beta}}$ is consistent for the true parameter $\boldsymbol{\beta}_0$, we have the linear approximation $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = -\{\nabla^2 l(\boldsymbol{\beta}_0)\}^{-1}\nabla l(\boldsymbol{\beta}_0)$. In order to estimate the covariance of $b(\boldsymbol{\beta}_0)$, we use the approximation

$$\frac{1}{|\tilde{\beta}_j|} = \frac{1}{|\beta_{j0}|} - \frac{\operatorname{sign}(\beta_{j0})}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0}),$$

which leads to

$$\begin{aligned} b_j(\boldsymbol{\beta}_0) &= \frac{\operatorname{sign}(\beta_{j0})}{|\tilde{\beta}_j|} = \operatorname{sign}(\beta_{j0}) \left[\frac{1}{|\beta_{j0}|} - \frac{\operatorname{sign}(\beta_{j0})}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0})\right] \\ &= \frac{\operatorname{sign}(\beta_{j0})}{|\beta_{j0}|} - \frac{I(\beta_{j0} \neq 0)}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0}). \end{aligned}$$

Define $D(\boldsymbol{\beta}) = \operatorname{diag}\{\frac{I(\beta_1 \neq 0)}{\beta_1^2}, ..., \frac{I(\beta_p \neq 0)}{\beta_p^2}\}$ and the vector $g(\boldsymbol{\beta}) = \left(\frac{\operatorname{sign}(|\beta_1|)}{|\beta_1|}, ..., \frac{\operatorname{sign}(|\beta_p|)}{|\beta_p|}\right)'$. We then have $b(\boldsymbol{\beta}_0) = g(\boldsymbol{\beta}_0) + D(\boldsymbol{\beta}_0)\{\nabla^2 l(\boldsymbol{\beta}_0)\}^{-1}\nabla l(\boldsymbol{\beta}_0)$ and

$$\nabla l(\boldsymbol{\beta}_0) + \lambda b(\boldsymbol{\beta}_0) = \left[I + \lambda D(\boldsymbol{\beta}_0)\{\nabla^2 l(\boldsymbol{\beta}_0)\}^{-1}\right] \nabla l(\boldsymbol{\beta}_0) + \lambda g(\boldsymbol{\beta}_0).$$

Therefore

$$\begin{aligned}\widehat{\text{cov}}\left(\nabla l(\boldsymbol{\beta}_0) + \lambda b(\boldsymbol{\beta}_0)\right) &= \left[I + \lambda D(\hat{\boldsymbol{\beta}})\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}\right] \nabla^2 l(\hat{\boldsymbol{\beta}}) \left[I + \lambda D(\hat{\boldsymbol{\beta}})\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}\right]' \\ &= [\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})]\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})].\end{aligned}$$

Combining (15) and (16), we get the covariance of the ALASSO estimator $\hat{\boldsymbol{\beta}}$

$$\left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1} [\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})]\{\nabla^2 l(\hat{\boldsymbol{\beta}})\}^{-1}[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda D(\hat{\boldsymbol{\beta}})] \left[\nabla^2 l(\hat{\boldsymbol{\beta}}) + \lambda A(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$, where $\hat{\boldsymbol{\beta}}_1$ consists of the r non-zero components. Correspondingly, we decompose the Hessian matrix as

$$G = \nabla^2 l(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} denotes the first $r \times r$ submatrix of G. Let A_{11} be the first $r \times r$ submatrix of $A \equiv A(\hat{\boldsymbol{\beta}})$, $E = G_{22} - G_{21}G_{11}^{-1}G_{12}$, and $\tilde{G}_{11} = G_{11} + \lambda A_{11}$. Then the covariance estimate of the nonzero estimates $\hat{\boldsymbol{\beta}}_1$ is

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}_1) = G_{11}^{-1} + \left[G_{11}^{-1} - \tilde{G}_{11}^{-1}\right] G_{12}E^{-1}G_{21} \left[G_{11}^{-1} - \tilde{G}_{11}^{-1}\right]. \tag{17}$$

REMARK. Since $\hat{l}_{n,M}(\boldsymbol{\beta})$ is obtained by importance sampling, the Monte Carlo simulation introduces additional variations in the estimation of $\boldsymbol{\beta}$. In fact, the variance of $\tilde{\boldsymbol{\beta}}$ (when $\lambda = 0$) should be $G^{-1} + S$, where S represents the share of the variability due to Monte Carlo simulations. However, as noted by Lam and Kuk (1997) and Lam and Leung (2001), S is relatively small compare to G^{-1} and can be ignored for estimating the variance of $\tilde{\boldsymbol{\beta}}$. Therefore, in this paper, we ignore the variation due to the Monte Carlo simulations when computing the variances of the ALASSO and LASSO estimates.

4.2. Parameter Tuning

To estimate the tuning parameter λ, we use the generalized cross validation (GCV) criterion (Wahba, 1990), which was also suggested by Tibshirani (1997), and Fan and Li (2002). At convergence, the ALASSO solution $\hat{\boldsymbol{\beta}}$ can be approximated by a ridge regression estimator $(G + \lambda A)^{-1} X'\mathbf{Y}$. Therefore the number of effective parameters in the ALASSO estimator can be approximated by $p(\lambda) = \text{tr}[(G + \lambda A)^{-1} G]$. The GCV-type statistic is constructed as

$$GCV(\lambda) = \frac{-\hat{l}_{n,M}(\hat{\boldsymbol{\beta}})}{n\left[1 - p(\lambda)/n\right]^2}.$$

5. Numerical Studies

5.1. Simulation Study

We compare the ALASSO, the LASSO, and the maximum marginal likelihood estimates (MMLE) in terms of the mean squared error, model size, and variable selection accuracy. Following Tibshirani (1997), we report the median of the mean squared error (MSE) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ over 50 simulations for each method. Here Σ is the population covariance matrix of the covariates. We also show the average numbers of correct and incorrect zero coefficients in the final models. Generalized cross validation is used to estimate the tuning parameter λ for the ALASSO and LASSO. All simulations are done with R codes.

Both the proportional hazards and proportional odds model are considered for the survival times. The base design contains eight covariates $(Z_1, ..., Z_8)$, which are marginally standard normal and the correlation between Z_j and Z_k is $\rho^{|j-k|}$ for $j \neq k$, with $\rho = 0.2$. The regression coefficients are chosen as $\boldsymbol{\beta} = (-0.7, 0, 0, -0.7, 0, 0, -0.7, 0)^T$,

and thus only Z_1, Z_4 and Z_7 are important covariates. The transformation function takes the form $H(t) = \log(t)$ for the proportional hazards model and $H(t) = 3\log(t)$ for the proportional odds model. 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 the sample size $n = 100$. Table 1 summarizes the mean square errors and variable selection results for three methods under four different settings. Overall, the ALASSO works the best and the LASSO is second best in terms of median MSE and the correct number of zero coefficients appearing in the estimates. For example, under the proportional hazards model with 25% censoring rate, the ALASSO solution selects important covariates most accurately (the true model size 3, ALASSO 3.8, LASSO 5, MMLE 8), and gives the smallest mean squared error (ALASSO 0.078, LASSO 0.104, MMLE 0.135).

(Insert Table 1 here)

In Figures 1 and 2 are the box-plots of the estimated coefficients for the proportional hazards and odds models with 25% censoring rate, respectively. We can see that the ALASSO effectively shrinks small coefficients to zero and estimates non-zero coefficients with little bias. Similar results also hold for the case of 40% censoring rate, which are not presented here.

To test the accuracy of the proposed standard error formula given in Section 4, we compare the sample standard errors with their estimates obtained using (17). For the LASSO estimates, we use the formula in Tibshirani (1997) to compute their standard errors. Under each scenario, we generate 50 new data sets of sample size 100 and fix $\lambda = \hat{\lambda}$, the optimal value chosen by the GCV in the previous simulation study. In Table 2, we summarize the mean of the estimated standard errors and the sample standard

errors from Monte Carlo simulations. The estimated standard errors of MMLEs are close to the sample standard errors for both the proportional hazards and odds models. The estimated standard errors of the LASSO and ALASSO estimators are in agreement with the sample standard errors under the proportional hazards model, but a little bit underestimate the sample standard errors under the proportional odds model.

(Insert Table 2 here)

5.2. Application to Lung Cancer Data

This 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). The data set has been analyzed by many authors such as Tibshirani (1997), Lam and Kuk (2001), and Chen, Jin and Ying (2002). Tibshirani (1997) fitted the Cox's proportional hazards model with the LASSO penalty, and found that Karnofsky score shows a dominant effect, while Treatment and Cell type have moderate influence to the survival time. Therein, Cell type was treated as a continuous variable. Lam and Kuk (2001) fitted the proportional odds model to a subset of the data containing only 97 patients with no prior therapy based on the marginal likelihood approach. Chen, Jin and Ying (2002) fitted the linear transformation models to the same subset of the data using estimating equations. They only considered two variables Cell type and Karnofsky score and found both of them significant, where Cell type was treated as categorical.

(Insert Table 3 here)

We consider all the covariates in the linear transformation models. The MMLE,

LASSO, ALASSO estimates are computed for both the proportional hazards and odds models. Tables 3 and 4 summarize the estimated coefficients and their standard errors. The MMLEs are in good agreement with those reported in literature (Lam and Kuk, 2001; Chen et al., 2002). For the proportional hazards model, both the LASSO and ALASSO select Cell type (squamous vs large, small vs large, adeno vs large) and Karnofsky score as important variables. For the proportional odds model, the ALASSO selects Cell type (small vs large, adeno vs large) and Karnofsky score, while the LASSO selects the same set of variables as in the proportional hazards model.

6. Discussion

The class of semiparametric linear transformation models has received much attention recently due to its high flexibility. In this paper, we have studied the penalized marginal likelihood method with the classical LASSO penalty and proposed a new adaptive L_1 penalty ALASSO for variable selection under linear transformation models. Based on the numerical study results, the ALASSO shows better performance in terms of variable selection and model estimation compared to the maximum marginal likelihood estimate and its LASSO variant. The theoretical properties, such as root-n consistency and oracle property, of the ALASSO have been studied under Cox's proportional hazards model by the authors in another paper (Zhang and Lu, 2006) using the partial likelihood. However, for linear transformation models, the marginal likelihood usually does not have a closed form and the numerical method is needed to approximate it. Therefore, the derivation of the theoretical properties of the ALASSO under linear transformation models based on the marginal likelihood is quite complicated and needs further investigation.

Acknowledgment

Wenbin Lu's research was partially supported by National Science Foundation Grant DMS-0504269. Hao Helen Zhang's research was partially supported by National Science Foundation Grant DMS-0405913.

Reference

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations. *Journal of the American Statistical Association* **96**, 939-963.

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.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200-1224.

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.

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.

Zhang, H. H. and Lu, W. (2006). Adaptive-LASSO for Cox's proportional hazards model. *Submitted*.

Table 1: Mean squared errors and model selection results.

Model setting	Method	Median MSE	Ave. number of zero coefficients	
			correct (5)	incorrect (0)
PH	MMLE	0.135	0.0	0.0
25%	LASSO	0.104	3.0	0.0
	ALASSO	**0.078**	4.2	0.0
PH	MMLE	0.169	0.0	0.0
40%	LASSO	0.111	2.6	0.0
	ALASSO	**0.079**	3.9	0.0
PO	MMLE	0.337	0.0	0.0
25%	LASSO	0.233	4.0	0.0
	ALASSO	**0.229**	4.6	0.1
PO	MMLE	0.439	0.0	0.0
40%	LASSO	0.336	3.9	0.1
	ALASSO	**0.303**	4.4	0.2

Note: PH stands for the proportional hazards model and PO for the proportional odds model.

Table 2: Estimated and actual standard errors for the coefficients.

		$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
Model setting	Method	SE	$\widehat{\text{SE}}$	SE	$\widehat{\text{SE}}$	SE	$\widehat{\text{SE}}$
PH	MMLE	0.119	0.147	0.159	0.142	0.125	0.110
25%	LASSO	0.088	0.115	0.136	0.111	0.110	0.111
	ALASSO	0.098	0.139	0.154	0.134	0.122	0.135
PH	MMLE	0.136	0.159	0.167	0.155	0.148	0.159
40%	LASSO	0.118	0.125	0.147	0.121	0.131	0.121
	ALASSO	0.126	0.151	0.163	0.148	0.144	0.148
PO	MMLE	0.173	0.213	0.236	0.208	0.239	0.212
25%	LASSO	0.151	0.125	0.193	0.119	0.194	0.120
	ALASSO	0.193	0.204	0.262	0.190	0.256	0.185
PO	MMLE	0.200	0.231	0.251	0.225	0.256	0.231
40%	LASSO	0.171	0.142	0.210	0.132	0.208	0.136
	ALASSO	0.213	0.221	0.271	0.203	0.271	0.214

Note: PH and PO are defined the same as in Table 1. SE stands for the sample standard errors of the estimated coefficients and $\widehat{\text{SE}}$ stands for the mean of estimated standard errors.

Table 3: Estimated coefficients and standard errors for lung cancer data under PH model.

Covariate	Proportional Hazards Model		
	MMLE	LASSO	ALASSO
Treatment	0.246 (0.211)	0 (-)	0 (-)
squamous vs large	-0.311 (0.288)	-0.270 (0.107)	-0.078 (0.265)
small vs large	0.416 (0.277)	0.037 (0.023)	0.145 (0.272)
adeno vs large	0.688 (0.312)	0.295 (0.118)	0.497 (0.304)
Karnofsky	-0.030 (0.005)	-0.025 (0.004)	-0.029 (0.005)
Months from Diagnosis	-0.0004 (0.009)	0 (-)	0 (-)
Age	-0.008 (0.010)	0 (-)	0 (-)
Prior therapy	0.009 (0.024)	0 (-)	0 (-)

Table 4: Estimated coefficients and standard errors for lung cancer data under PO model.

Covariate	Proportional Odds Model		
	MMLE	LASSO	ALASSO
Treatment	0.144 (0.302)	0 (-)	0 (-)
squamous vs large	-0.040 (0.458)	-0.061 (0.048)	0 (-)
small vs large	1.085 (0.418)	0.620 (0.214)	0.706 (0.356)
adeno vs large	1.202 (0.447)	0.732 (0.251)	0.841 (0.397)
Karnofsky	-0.054 (0.008)	-0.049 (0.007)	-0.053 (0.008)
Months from Diagnosis	-0.001 (0.017)	0 (-)	0 (-)
Age	-0.013 (0.015)	0 (-)	0 (-)
Prior therapy	0.013 (0.036)	0 (-)	0 (-)

Figure 1: Box-plot for the proportional hazards model under 25% censoring.

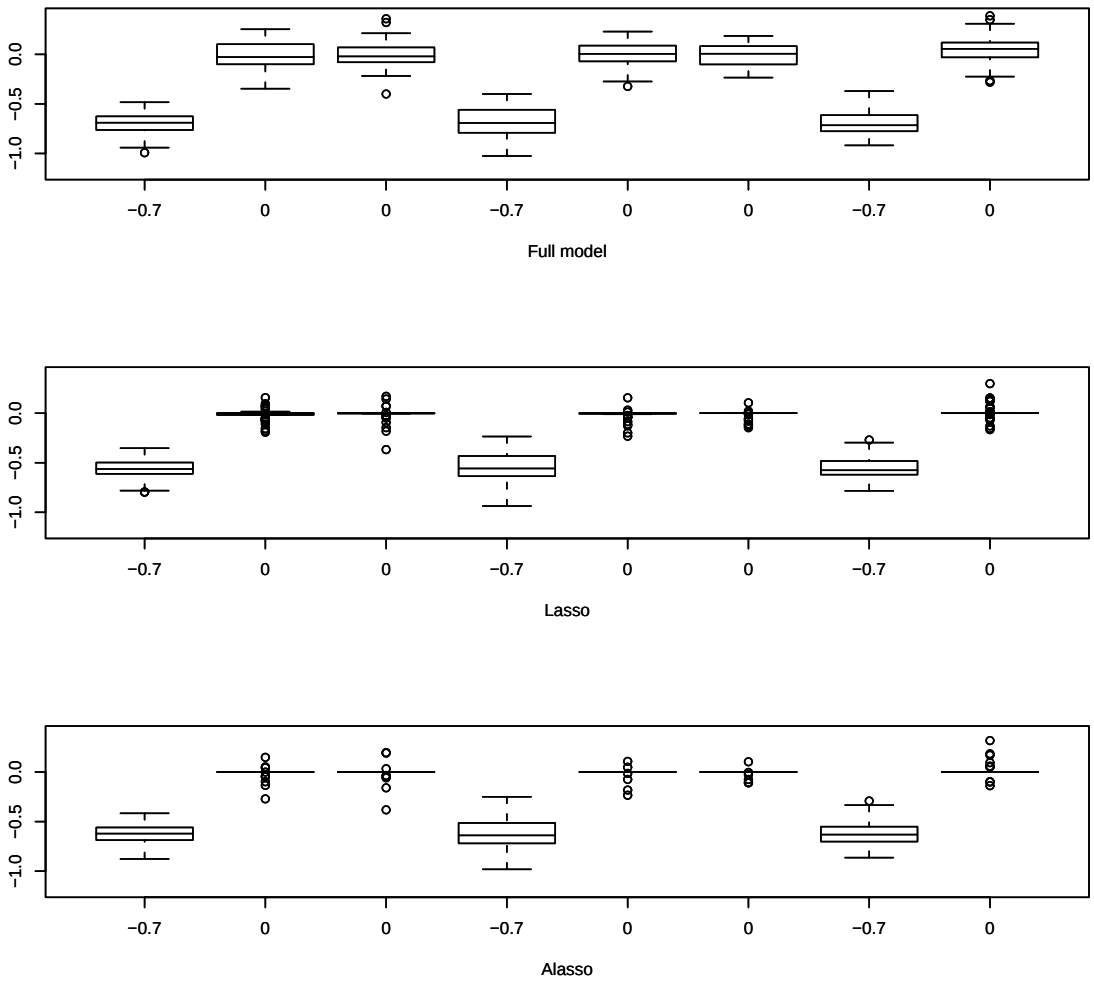

Figure 2: Box-plot for the proportional odds model under 25% censoring.