STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:3771–3781
Published online 1 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2833

Variable selection for proportional odds model

Wenbin Lu*,† and Hao H. Zhang

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

SUMMARY

In this paper we study the problem of variable selection for the proportional odds model, which is a useful alternative to the proportional hazards model and might be appropriate when the proportional hazards assumption is not satisfied. We propose to fit the proportional odds model by maximizing the marginal likelihood subject to a shrinkage-type penalty, which encourages sparse solutions and hence facilitates the process of variable selection. Two types of shrinkage penalties are considered: the LASSO and the adaptive-LASSO (ALASSO) penalty. In the ALASSO penalty, different weights are imposed on different coefficients such that important variables are more protectively retained in the final model while unimportant ones are more likely to be shrunk to zeros. We further provide an efficient computation algorithm to implement the proposed methods, and demonstrate their performance through simulation studies and an application to real data. Numerical results indicate that both methods can produce accurate and interpretable models, and the ALASSO tends to work better than the usual LASSO. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: marginal likelihood; proportional odds model; variable selection; shrinkage estimate

1. INTRODUCTION

One main issue in survival analysis is to study the dependence of the survival time T of patients on various clinical covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)$. Though the proportional hazards model [1] 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 [2–6]. In this paper, we focus on the proportional odds model and develop new methods for joint model estimation and variable selection. Variable selection refers

*Correspondence to: Wenbin Lu, Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.
†E-mail: lu@stat.ncsu.edu

Contract/grant sponsor: National Science Foundation; contract/grant numbers: DMS-0504269, DMS-0405913

Received 9 July 2006
Accepted 15 December 2006

Copyright © 2007 John Wiley & Sons, Ltd.

to the process of selecting from $Z_1, \ldots, Z_p$ those that are most associated or predictive of the survival time. Variable selection is fundamental to survival modelling, and an effective variable selection can lead to better risk assessment and model interpretation. In ordinary linear regression, there are a variety of classical techniques for variable selection such as forward selection, backward elimination, and stepwise selection; see References [7, 8] for a thorough review on these methods. To gauge the number of variables included in the model, Mallows' C_p [9, 10], Akaike's information criterion [11, 12] and Schwarz's Bayesian information criterion [13] are widely used. However, classical methods like stepwise selection procedures can be expensive in computation for large data sets and often suffer from high variability. The nature of censored data makes variable selection a challenging problem in survival data analysis. Recently, shrinkage methods have been proposed for Cox's proportional hazards model based on the partial or pseudo-partial likelihoods. Among them, the LASSO [14, 15] and the SCAD [16, 17] are popular and have shown good performance in practice. More recently, the adaptive-LASSO (ALASSO) [18, 19] was proposed for linear models and shown to produce parsimonious models more effectively than the LASSO; Zhang and Lu [20] studied the ALASSO estimator for Cox's proportional hazards models. However, very little work has been done for variable selection in the proportional odds model. One main difficulty in dealing with the proportional odds model is that the partial likelihood function is not available. In this paper, we adopt the marginal likelihood procedure [21] and propose the penalized marginal likelihood method for variable selection in the proportional odds model. Both the LASSO and ALASSO penalties are investigated in the context of maximizing the constrained marginal likelihoods.

The next section gives the marginal likelihood function for the proportional odds model. In Section 3, we present the penalized marginal likelihood methods with the LASSO and ALASSO penalties. An efficient algorithm is proposed for optimizing the penalized log marginal 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 (SEs) of our estimates. Section 5 is devoted to simulation studies and one application to a real data set from the Veteran's Administration lung cancer study. Some final remarks are given in Section 6.

2. MARGINAL LIKELIHOOD FOR THE PROPORTIONAL ODDS MODEL

In a survival analysis setting with right censored data, we observe n independent copies of $\tilde{T}_i = \min(T_i, C_i)$, where T_i is the time until the occurrence of some event of interest, C_i is a censoring time, and $\delta_i = I(T_i \leqslant C_i)$ is the censoring indicator. For the ith individual, let Z_{ij} be the jth covariate value and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$ be the p-dimensional vector of covariates. Let $S(t|\mathbf{Z})$ denote the conditional survival function of T given $\mathbf{Z}$ and $S_0(t)$ denote the completely unspecified baseline survival function for an individual with $\mathbf{Z} = \mathbf{0}$. The proportional odds model assumes that

$$\frac{1 - S(t|\mathbf{Z})}{S(t|\mathbf{Z})} = \frac{1 - S_0(t)}{S_0(t)} \exp(\boldsymbol{\beta}'\mathbf{Z}) \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the regression parameter vector. The proportional odds model (1) can be equivalently written as

$$H(T) = -\boldsymbol{\beta}'\mathbf{Z} + \varepsilon \tag{2}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $H(t) = \log[\{1 - S_0(t)\}/S_0(t)]$ and ε follows the standard logistic distribution, i.e. $F(x) = P(\varepsilon \leqslant x) = \exp(x)/\{1 + \exp(x)\}$. Let $\Lambda(x)$ denote the cumulative hazard function of ε, i.e. $\Lambda(x) = \log\{1 + \exp(x)\}$, and $\lambda(x) = \mathrm{d}\Lambda(x)/\mathrm{d}x$.

For the proportional odds model, the usual partial likelihood function of $\boldsymbol{\beta}$ is not available. Therefore, we choose to estimate the model by maximizing the marginal likelihood function [21]. 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 $0 \leqslant k \leqslant K$, let $\mathscr{L}_k$ 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 $\mathscr{C}$ denote the collection of all possible rank vectors of T_i's consistent with the observed data $(\tilde{T}_i, \delta_i)$ $(i = 1, \ldots, n)$. The marginal likelihood is then defined by $L_{n,M}(\boldsymbol{\beta}) = P(\mathbf{R} \in \mathscr{C})$, where the probability is with respect to the underlying uncensored version of the study. Following Lam and Leung [21], $L_{n,M}(\boldsymbol{\beta})$ can be represented as

$$L_{n,M}(\boldsymbol{\beta}) = \int \cdots \int_{V_{(1)} < \cdots < V_{(K)}} \prod_{i=1}^{n} \{\lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)\}^{\delta_i} \mathrm{e}^{-\Lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)} \prod_{k=1}^{K} \mathrm{d}V_{(k)} \tag{3}$$

where $V_{(k)} = H(T_{(k)})$, $k = 1, \ldots, K$. Note that (3) is independent of the non-parametric function H, or it is baseline-free. Since the integral in (3) has no analytical form, we use the importance sampling method [22] to approximate (3). Towards this, we multiply and divide the integrand in (3) by

$$c \prod_{i=1}^{n} \{\lambda(V_{(k_i)})\}^{\delta_i} \mathrm{e}^{-\Lambda(V_{(k_i)})} \tag{4}$$

where the constant c is the total number of possible rank vectors in $\mathscr{C}$. It can be shown that (4) is the density function of $V_{(1)}, \ldots, V_{(K)}$ under progressive type II censoring [23] when the underline $V_i \equiv H(T_i)$ $(i = 1, \ldots, n)$ are independent and identically distributed according to the distribution function $F(x)$. Then the marginal likelihood (3) can be expressed as $L_{n,M}(\boldsymbol{\beta}) = E\{Q(V_{(1)}, \ldots, V_{(K)}; \boldsymbol{\beta})\}$, where the expectation is with respect to the density (4) and

$$Q(V_{(1)}, \ldots, V_{(K)}; \boldsymbol{\beta}) = \frac{1}{c} \prod_{i=1}^{n} \frac{\{\lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)\}^{\delta_i} \mathrm{e}^{-\Lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)}}{\{\lambda(V_{(k_i)})\}^{\delta_i} \mathrm{e}^{-\Lambda(V_{(k_i)})}} \tag{5}$$

Now we can estimate $L_{n,M}$ by

$$\hat{L}_{n,M}(\boldsymbol{\beta}) = \frac{1}{b} \sum_{b=1}^{B} Q\{F^{-1}(U^b_{(1)}), \ldots, F^{-1}(U^b_{(K)}); \boldsymbol{\beta}\} \tag{6}$$

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

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3. VARIABLE SELECTION FOR PROPORTIONAL ODDS MODEL

3.1. Penalized marginal likelihood methods

Recall that $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$ represents the covariate vector for the ith individual, $i = 1, \ldots, n$. For each $j = 1, \ldots, p$, we assume that the vector $(Z_{1j}, \ldots, Z_{nj})^{\mathrm{T}}$ is standardized, i.e.

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

The LASSO, proposed by Tibshirani [14] in linear regression settings, is the penalized least squares estimates with the L_1 penalty. It was also studied for Cox's proportional hazard model by Tibshirani [15]. Overall the LASSO estimates achieve a smaller mean-squared error (MSE) than the ordinary least squares estimates. For the proportional odds model, we first consider the penalized marginal likelihood with the LASSO penalty

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j| \tag{7}$$

where $\hat{l}_{n,M}(\boldsymbol{\beta}) = \log\{\hat{L}_{n,M}(\boldsymbol{\beta})\}$. 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 $\lambda \geqslant 0$ 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 GCV to choose λ.

One drawback of the LASSO penalty is that the estimates for important covariates may suffer from substantial bias [24]. 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 non-zero coefficients. Alternatively, we can impose the ALASSO penalty on the marginal likelihood

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|\tau_j \tag{8}$$

where the non-negative weights $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^{\mathrm{T}}$ 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. As for the choice of $\boldsymbol{\tau}$, any consistent estimate of $\boldsymbol{\beta}$ can be good candidates [18–20]. Denote the maximum marginal likelihood estimate (MMLE) of $\hat{l}_{n,M}$ by $\tilde{\boldsymbol{\beta}}$. Since $\tilde{\boldsymbol{\beta}}$ are shown to be consistent [21], their absolute values reflect the relative importance of covariates. Therefore, we choose $\tau_j^{-1} = |\tilde{\beta}_j|$ and solve

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\hat{l}_{n,M}(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j| \tag{9}$$

If $\tilde{\beta}_j = 0$, we set the solution $\hat{\beta}_j = 0$. When equal weights are used in (8), the ALASSO reduces to the standard LASSO.

3.2. Computation algorithm

We provide an iterative computation algorithm to solve (9). By slightly modifying the weights, we can solve the LASSO with the same algorithm. The proposed algorithm uses the Newton–Raphson

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

updates and sequentially solves a least squares problem subject to the weighted L_1 penalty. At each step, we suggest a modified shooting algorithm [25] to solve the L_1-constrained quadratic programming exactly.

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 $(1/2n)(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta})$. In each iterative step, we need to minimize

$$\frac{1}{2n}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j| \tag{10}$$

The modified the shooting algorithm [25] to minimize (10) is presented in the Appendix. For any fixed λ, the following is a complete algorithm to compute the ALASSO solution.

1. Solve $\tilde{\boldsymbol{\beta}}$ by maximizing the marginal likelihood function $\hat{l}_{n,M}(\boldsymbol{\beta})$.
2. Initialization: $k = 1$ and $\beta_j^{(1)} = 0$ for $j = 1, \ldots, p$.
3. Compute ∇l, $\nabla^2 l$, X, $\mathbf{Y}$ based on the current $\boldsymbol{\beta}^{(k)}$.
4. Minimize (10) *via* the modified shooting algorithm. Denote the solution as $\boldsymbol{\beta}^{(k+1)}$.
5. Let $k = k + 1$. Go to step 3 until convergence.

4. STANDARD ERRORS AND PARAMETER TUNING

4.1. Standard errors of ALASSO estimates

Assume $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. Since the ALASSO estimates 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 SEs. We use the conventional technique in the likelihood setting and approximate the covariance matrix of the ALASSO estimates using the corresponding sandwich formula.

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, following the LQA algorithm proposed by Fan and Li [24] and recently studied by Hunter and Li [26], we can approximate the penalty function locally 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

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

Then (9) can be locally approximated (except for a constant term) by

$$-\hat{l}_{n,M}(\boldsymbol{\beta}^{(1)}) + \nabla l(\boldsymbol{\beta}^{(1)})'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)})'\nabla^2 l(\boldsymbol{\beta}^{(1)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(1)}) + n\lambda \sum_{j=1}^{p} \frac{\beta_j^2}{2|\beta_j^{(1)}||\tilde{\beta}_j|} \tag{11}$$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - [\nabla^2 l(\hat{\boldsymbol{\beta}}^{(k)}) + n\lambda A(\hat{\boldsymbol{\beta}}^{(k)})]^{-1}[\nabla l(\hat{\boldsymbol{\beta}}^{(k)}) + n\lambda b(\hat{\boldsymbol{\beta}}^{(k)})] \tag{12}$$

where $A(\boldsymbol{\beta}) = \text{diag}\{1/\beta_1^2, \ldots, 1/\beta_p^2\}$ and $b(\boldsymbol{\beta}) = (\text{sign}(|\beta_1|)/|\tilde{\beta}_1|, \ldots, \text{sign}(|\beta_p|)/|\tilde{\beta}_p|)'$. Then the corresponding sandwich formula for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$[\nabla^2 l(\hat{\boldsymbol{\beta}}) + n\lambda A(\hat{\boldsymbol{\beta}})]^{-1}\, \widehat{\text{cov}}(\nabla l(\boldsymbol{\beta}_0) + n\lambda b(\boldsymbol{\beta}_0))[\nabla^2 l(\hat{\boldsymbol{\beta}}) + n\lambda A(\hat{\boldsymbol{\beta}})]^{-1} \tag{13}$$

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 $1/|\tilde{\beta}_j| = 1/|\beta_{j0}| - (\tilde{\beta}_j - \beta_{j0}) \cdot \text{sign}(\beta_{j0})/\beta_{j0}^2$, which leads to

$$b_j(\boldsymbol{\beta}_0) = \frac{\text{sign}(\beta_{j0})}{|\tilde{\beta}_j|} = \frac{\text{sign}(\beta_{j0})}{|\beta_{j0}|} - \frac{I(\beta_{j0} \neq 0)}{\beta_{j0}^2}(\tilde{\beta}_j - \beta_{j0})$$

Define $D(\boldsymbol{\beta}) = \text{diag}\{I(\beta_1 \neq 0)/\beta_1^2, \ldots, I(\beta_p \neq 0)/\beta_p^2\}$ and the vector $g(\boldsymbol{\beta}) = (\text{sign}(|\beta_1|)/|\beta_1|, \ldots, \text{sign}(|\beta_p|)/|\beta_p|)'$. 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

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

Therefore,

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

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

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

Since $\hat{l}_{n,M}(\boldsymbol{\beta})$ is obtained by importance sampling [22], 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 $\{\nabla^2 l(\tilde{\beta})\}^{-1} + S$, where S represents the share of the variability due to Monte Carlo simulations. However, as noted by Lam and Leung [21], S is relatively small compare to $\{\nabla^2 l(\tilde{\boldsymbol{\beta}})\}^{-1}$ and can be ignored for estimating the variance of $\tilde{\boldsymbol{\beta}}$. Therefore, 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 GCV criterion [27], which was also used for tuning in Cox's proportional hazards model [15, 16, 20]. At convergence, the ALASSO solution $\hat{\boldsymbol{\beta}}$ can be approximated by a ridge regression estimator $\{\nabla^2 l(\hat{\boldsymbol{\beta}}) + n\lambda A\}^{-1}X'\mathbf{Y}$. Therefore, the number of effective parameters in the ALASSO estimate can be approximated by

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$p(\lambda) = \mathrm{tr}[\{\nabla^2 l(\hat{\boldsymbol{\beta}}) + n\lambda A\}^{-1} \nabla^2 l(\hat{\boldsymbol{\beta}})]$. The GCV-type statistic is then constructed as

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

5. NUMERICAL STUDIES

5.1. Simulation study

We compare the MMLEs, the LASSO, and the ALASSO estimates in terms of their MSE, model size, and variable selection accuracy. Following Tibshirani [15], we report the median of the MSE $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ 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 and available upon request.

The proportional odds model (2) is considered for the survival times. The base design contains eight covariates $(Z_1, \ldots, Z_8)$, 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 $\boldsymbol{\beta} = (-0.7, 0, 0, -0.7, 0, 0, -0.7, 0)^{\mathrm{T}}$, and thus only Z_1, Z_4, and Z_7 are important covariates. We choose $H(t) = 3 \log(t)$. Censoring times are generated from the uniform distribution over $[0, c_0]$, where c_0 is chosen to obtain the desired censoring rate. We consider two types of censoring rate: 25 and 40 per cent, and two sample size settings: $n = 100$ and 200. Table I summarizes the MSEs 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, with 25 per cent censoring

Table I. Mean-squared errors and model selection results.

				Average number of zero coefficients	
n	Censoring (per cent)	Method	Median MSE	Correct (5)	Incorrect (0)
100	25	MMLE	0.337	0.0	0.0
		LASSO	0.233	4.0	0.0
		ALASSO	**0.229**	4.6	0.1
	40	MMLE	0.439	0.0	0.0
		LASSO	0.336	3.9	0.1
		ALASSO	**0.303**	4.4	0.2
200	25	MMLE	0.123	0.0	0.0
		LASSO	0.168	3.8	0.0
		ALASSO	**0.115**	4.6	0.0
	40	MMLE	0.163	0.0	0.0
		LASSO	0.158	3.6	0.0
		ALASSO	**0.099**	4.7	0.0

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table II. Estimated and actual standard errors for the non-zero coefficients.

			$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
n	Censor (per cent)	Method	Bias	SE	$\widehat{\text{SE}}$	Bias	SE	$\widehat{\text{SE}}$	Bias	SE	$\widehat{\text{SE}}$
100	25	LASSO	0.244	0.161	0.126	0.247	0.218	0.119	0.237	0.204	0.122
		ALASSO	0.176	0.198	0.206	0.186	0.268	0.188	0.166	0.254	0.195
	40	LASSO	0.266	0.182	0.125	0.287	0.232	0.115	0.271	0.207	0.120
		ALASSO	0.199	0.236	0.212	0.228	0.292	0.195	0.194	0.262	0.203
200	25	LASSO	0.177	0.137	0.103	0.186	0.131	0.104	0.200	0.133	0.103
		ALASSO	0.111	0.152	0.136	0.123	0.149	0.138	0.134	0.146	0.138
	40	LASSO	0.184	0.146	0.109	0.191	0.148	0.109	0.202	0.149	0.108
		ALASSO	0.119	0.169	0.145	0.128	0.162	0.149	0.135	0.163	0.147

Note: SE stands for the sample standard errors of the estimated coefficients and $\widehat{\text{SE}}$ stands for the mean of estimated standard errors.

rate, the ALASSO solution selects important covariates most accurately (the true model size 3, ALASSO 3.4, LASSO 4.0, MMLE 8.0), and gives the smallest MSE (ALASSO 0.229, LASSO 0.233, MMLE 0.337).

In Table II, we compare the biases of the LASSO and ALASSO estimates in various settings. For ordinary linear models, it is well-known that, although both procedures can produce consistent regression coefficients asymptotically, their shrinkage nature tends to produce biased estimates for finite samples. Table II shows that the ALASSO solutions consistently have smaller bias than the LASSO in all the settings and the differences are substantial. In addition, when the sample size increases, both procedures improve in term of bias reduction. This can be regarded as empirical evidence for the asymptotic consistency of the estimates.

To test the accuracy of the proposed SE formula given in Section 4, we compare the sample SEs with their estimates obtained using (13). For the LASSO estimates, we use the formula in Tibshirani [15] to compute their SEs. In Table II, we summarize the mean of the estimated SEs and the sample SEs from Monte Carlo simulations. The estimated SEs for both LASSO and ALASSO estimates are reasonably close to their sample SEs. Overall, the ALASSO estimated SEs for all important effects are closer to the sample SEs than the LASSO. The estimated SEs of the LASSO tend to underestimate the sample SEs, which was also observed by Tibshirani [15] in the Cox's proportional hazards model. When n increases from 100 to 200, the performance of SE estimation formula for both procedures shows significant improvement.

5.2. *Application to lung cancer data*

We apply all the three methods to the data from the Veteran's Administration lung cancer trial [28]. 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 analysed by many authors such as Tibshirani [15] and Lam and Leung [21]. Tibshirani [15] fitted the Cox's proportional hazards model

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table III. Estimated coefficients and standard errors for lung cancer data.

Covariate	Proportional odds model		
	MMLE	LASSO	ALASSO
Treatment	0.144 (0.302)	0 (—)	0 (—)
Squamous *versus* large	−0.040 (0.458)	−0.061 (0.048)	0 (—)
Small *versus* large	1.085 (0.418)	0.620 (0.214)	0.706 (0.356)
Adeno *versus* 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 (—)

with the LASSO penalty, and found that Karnofsky score shows a dominant effect, and treatment and cell type have moderate influence on the survival time. Therein, cell type was treated as a continuous variable. Lam and Leung [21] 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. They only considered two variables cell type and Karnofsky score and found both of them significant, where cell type was treated as categorical.

We include all the covariates in the proportional odds model, and compute the MMLE, LASSO, ALASSO estimates. Table III summarizes the estimated coefficients and their SEs. The MMLEs are in good agreement with those reported in literature [21]. The ALASSO selects cell type (small *versus* large, adeno *versus* large) and Karnofsky score as important variables, while the LASSO selects one more cell type (squamous *versus* large). The optimal tuning parameter selected by GCV for the ALASSO is $\lambda = 0.034$. It is easy to notice that the models obtained by MMLE, LASSO, and ALASSO are nested. Based on a likelihood ratio-type test (suggested by one referee), we conclude that the model selected by LASSO is better than MMLE, and the model obtained by ALASSO is better than LASSO for this example.

We also fit an augmented model by including the interaction effects between treatment and other covariates. The MMLE analysis shows that none of the interaction effects is significant, the ALASSO procedure does not select any interaction terms either, while the LASSO selects two interaction terms (treatment with squamous *versus* large, and treatment with small *versus* large).

6. DISCUSSION

In this paper, we have studied the penalized marginal likelihood method with the LASSO and ALASSO penalties for variable selection under the proportional odds model. Based on the numerical study results, the ALASSO shows better performance in terms of variable selection and model estimation than the MMLE and its LASSO variant. Simulation results also show that lower the censoring rate, better the performance given by both procedures in terms of estimation and variable selection. Furthermore, the ALASSO has been shown to own oracle properties for linear models [19] and Cox's proportional hazards model [20], which suggests that its bias tends to zero when n goes to infinity. For finite samples, like all other shrinkage estimators the ALASSO may have

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

noticeable bias, as observed in our simulations. A possible way to correct the biases in the estimates is to use a two-step procedure: firstly, we select important variables with the ALASSO procedure; secondly, fit the MMLE with the selected variables at the first step.

The choice of the weights τ_j's is very important for the ALASSO. In the paper, we suggest using the inverse of the absolute MMLEs. But other consistent estimates can also be used. Like most other variable selection methods, the proposed method should work for any data set with $n > p$. In some complicated situations, for example, in high-dimensional gene expression data the number of covariates p is much large than the sample size n, consistent estimates $\tilde{\beta}_j$'s may not be available. We suggest using some robust estimates such as ridge or the LASSO estimates to construct the weights.

We implemented the proposed computational algorithm in R. The codes are robust and converge quickly in all of our examples. The R code can be downloaded from www4.stat.ncsu.edu/∼lu, and www4.stat.ncsu.edu/∼hzhang.

APPENDIX

We have modified the shooting algorithm [25] to solve the penalized least squares

$$\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|/|\tilde{\beta}_j|$$

Define $G(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2$, $\dot{G}_j(\boldsymbol{\beta}) = \partial G(\boldsymbol{\beta})/\partial\beta_j$, $j = 1, \ldots, p$, and denote $\boldsymbol{\beta}$ by $(\beta_j, \boldsymbol{\beta}^{-j})'$, where $\boldsymbol{\beta}^{-j}$ is the $(p-1)$-dimensional vector consisting of the β_i's other than β_j. The modified shooting algorithm is then given as follows:

(i) Start with $\hat{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)'$ and let $\lambda_j = \lambda/|\tilde{\beta}_j|$ for $j = 1, \ldots, p$.

(ii) At step m, for each $j = 1, \ldots, p$, let $G_0 = \dot{G}_j(0, \hat{\boldsymbol{\beta}}_{m-1}^{-j})$ and set

$$\hat{\beta}_j = \begin{cases} \dfrac{\lambda_j - G_0}{2(\mathbf{x}^j)'\mathbf{x}^j} & \text{if } G_0 > \lambda_j \\ \dfrac{-\lambda_j - G_0}{2(\mathbf{x}^j)'\mathbf{x}^j} & \text{if } G_0 < -\lambda_j \\ 0 & \text{if } |G_0| \leqslant \lambda_j \end{cases}$$

where $\mathbf{x}^j = (x_{1j}, \ldots, x_{nj})'$. Form a new estimator $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ after updating all the $\hat{\beta}_j$'s.

(iii) Repeat (ii) until $\hat{\boldsymbol{\beta}}_m$ converges.

ACKNOWLEDGEMENTS

The authors are grateful to the referees and the editor for their constructive comments. 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.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

REFERENCES

1. Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
2. 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.
3. Pettitt AN. Proportional odds model for survival data and estimates using ranks. *Applied Statistics* 1984; **33**: 169–175.
4. Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine* 1983; **2**:273–277.
5. Dabrowska DM, Doksum KA. Estimation and testing in the two-sample generalized odds rate model. *Journal of the American Statistical Association* 1988; **83**:744–749.
6. Murphy SA, Rossini AJ, van der Vaart AW. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* 1997; **92**:968–976.
7. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics* 1976; **32**:1–49.
8. Miller AJ. *Subset Selection in Regression*. Chapman & Hall: London, 1990.
9. Mallows CL. Some comments on C_p. *Technometrics* 1973; **15**:661–675.
10. Mallows CL. More comments on C_p. *Technometrics* 1995; **37**:362–372.
11. Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 1973; **60**:255–265.
12. Akaike H. On entropy maximization principle. In *Application of Statistics*, Krishnaiah PR (ed.). North-Holland: Amsterdam, 1977; 27–41.
13. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464.
14. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *Series B* 1996; **58**:267–288.
15. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**:385–395.
16. Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* 2002; **30**:74–99.
17. Cai J, Fan J, Li R, Zhou H. Variable selection for multivariate failure time data. *Biometrika* 2005; **92**:303–316.
18. Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business and Economics Statistics* 2007, in press.
19. Zou H. The adaptive-LASSO and its oracle properties. *Journal of American Statistical Association* 2006; **101**:1418–1429.
20. Zhang HH, Lu W. Adaptive-LASSO for Cox's proportional hazards model. *Biometrika* 2007, in press.
21. Lam KF, Leung TL. Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis* 2001; **7**:39–54.
22. Monahan JF. *Numerical Methods of Statistics*. Cambridge University Press: Cambridge, UK, 2001.
23. Lawless JF. *Statistical Models and Methods for Lifetime Data*. Wiley: New York, 1982.
24. Fan J, Li R. Variable selection via penalized likelihood. *Journal of American Statistical Association* 2001; **99**:1348–1360.
25. Fu WJ. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 1998; **7**:397–416.
26. Hunter D, Li R. Variable selection using MM algorithms. *Annals of Statistics* 2005; **33**:1617–1642.
27. Wahba G. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. SIAM: Philadelphia, PA, 1990.
28. Kalbfleish JD, Prentice RL. *The Statistical Analysis of Failure Time Data* (2nd edn). Wiley: New Jersey, 2002.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2007, 20, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.2833 by Chungnam National University, Wiley Online Library on [26/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License