



Variable selection through adaptive elastic net for proportional odds model

Chunxiang Wang¹ · Nan Li² · Hongbin Diao¹ · Lanqing Lu¹

Received: 22 February 2023 / Accepted: 22 November 2023

© The Author(s) 2024

Abstract

In this paper, we propose a method for fitting the proportional odds model by maximizing the marginal likelihood while incorporating an elastic net penalty. We assign adaptive weights to different coefficients, allowing important variables to receive smaller penalties and be more protectively retained in the final model, while unimportant variables receive larger penalties and are more likely to be eliminated. This approach combines the strengths of adaptively weighted LASSO shrinkage and quadratic regularization, resulting in optimal large sample performance and the ability to effectively handle collinearity. We also present a computational algorithm for the proposed method and compare its performance to that of LASSO, elastic net, and adaptive LASSO through simulation studies and applications to real datasets. The results demonstrate that the proposed method tends to perform better than existing approaches.

Keywords Shrinkage · Penalization · Marginal likelihood · Survival analysis

1 Introduction

Variable selection, the process of identifying the covariates that are most associated or predictive of survival time from the set of available covariates, is crucial in survival modeling. Deciding which covariates to include in the final model is an important task for investigators, as it affects both accuracy and interpretability of the model. In ordinary linear regression settings, a variety of variable selection methods have been well established, such as all possible subsets method, forward selection, backward selection, and step-wise selection. These methods are typically evaluated using crite-

✉ Chunxiang Wang
chunxiangwang@live.com

¹ Shandong Academician Workstation of Big Data and AI, Shandong City Service Institute, Yantai 264000, Shandong, China

² U.S. Bank, Minneapolis, USA

ria such as Mallows' C_p (Mallows, 1973), Akaike's Information Criterion (Akaike, 1974) (AIC), Schwarz's Bayesian Information Criterion (Schwarz, 1978) (BIC), Copula Information Criterion (Grønneberg & Hjort, 2014) (CIC), Deviance Information Criterion (Spiegelhalter et al., 2014) (DIC), etc. However, these methods can suffer from high variability and long computation time for large datasets, as well as selection bias, which leads to overestimating the effects of the selected covariates.

In the last decades, shrinkage methods have become increasingly popular in model selection, such as LASSO (Tibshirani, 1996), Adaptive LASSO (Schneider & Wagner, 2012), and Elastic Net (Zou & Hastie, 2005). These methods use penalties to shrink the coefficients of less important variables to zero, thus reducing the risk of overfitting and improves the interpretability of the model. For survival analysis, variable selection is more challenging due to the nature of censored data. Recent methods, such as LASSO and Smoothly Clipped Absolute Deviation or SCAD (Fan & Li, 2001), have been proposed and applied for variable selection in survival analysis (Salerno & Li, 2022).

Cox's model is the most widely used model in survival analysis. However, proportional hazards, the most important assumption of Cox's model, are violated in some cases (e.g., when modeling prognostic factors in studies with long follow-up times). While Cox's model can be expanded to accommodate non-proportional hazards, such as by integrating time-varying regression coefficients (Hess, 1994; Hastie & Tibshirani, 1993), there isn't a universally endorsed or straightforward method to do so and developing a suitable model using these approaches can be a very complex and challenging task. In the cases where proportional hazards assumption is not satisfied, the proportional odds (PO) model is a useful alternative. The PO model was initially presented by Bennett (1983) within a semi-parametric framework. Under the PO model, which specifies that the covariate effect is a multiplicative factor on the baseline odds function, the hazard ratio between two sets of covariate values tends to approach unity instead of remaining constant as time progresses. Collett (1994) applied the PO model to the data on the survival times of women with breast tumors that were negatively or positively stained. Crowder et al. Crowder (1991) employed the PO model for the analysis of reliability data. Rossini and Tsiatis (1996) adapted the PO model for modeling current status data.

Regarding the variable selection problem for the PO model, Lu and Zhang (2007) suggested to fit it by maximizing the marginal likelihood subject to LASSO and adaptive LASSO penalties and their numerical study shown that adaptive LASSO outperforms LASSO. However, both of the two approaches have weakness when confront with high-dimensional data (Zou & Hastie, 2005; Zou & Zhang, 2009) or collinearity issues. The adaptive elastic net penalty (Zou & Zhang, 2009), using both l_2 and weighted l_1 constraints, inherits the oracle properties (Zou, 2006) from adaptive LASSO and has stronger capability of handling collinearity problem.

In this paper, we propose using the adaptive elastic net penalty to the marginal likelihood function of PO model and solve variable selection problems under proportional odds assumption. We compare its performance with LASSO, adaptive LASSO and elastic net methods in simulation studies as well as in applications to real datasets. Results show that the proposed method tends to work better than existing ones.

In Sect. 2, a brief review is given to common proportional odds model with right-censored data and its marginal likelihood function. In Sect. 3, we apply the adaptive

elastic net approach to marginal likelihood of proportional odds model. In Sect. 4, we develop a computational algorithm. We also discuss how to choose tuning parameters. In Sect. 5, we present the results of simulation study. Then, in Sect. 6, we apply this method to real data sets to compare its performance with other methods. Finally, summary and discussion are given in Sect. 7.

2 Proportional odds model and its marginal likelihood function

For a survival analysis problem involving right-censored data, the dataset comprises n independent observations, each denoted as $(\min\{T_i, C_i\}, \theta_i)$, where T_i represents the time until the occurrence of a specific event of interest, C_i denotes a censoring time, and θ_i stands for the censoring indicator $\mathbf{1}_{(T_i < C_i)}$. Additionally, we possess $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})'$, which denotes a p -dimensional vector of covariates for the i th observation. Our primary objective is to explore the relationship between the survival time T and the covariates \mathbf{Z} . In the realm of survival analysis, the most commonly employed model is the Proportional Hazards (PH) model, originally proposed by Cox (1972). However, in some certain scenarios, the underlying assumptions under the PH model may not hold. In such cases, the Proportional Odds model serves as a valuable alternative (Peterson et al., 1990). The Proportional Odds model is based on the assumption 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 $S(t | \mathbf{Z})$ denotes the conditional survival function of T given \mathbf{Z} and $S_0(t)$ is the baseline survival function with $\mathbf{Z} = \mathbf{0}$, which is completely unspecified. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the regression parameter vector.

Let $H(t) = \log[(1 - S_0(t))/S_0(t)]$, a regression model under proportional odds assumption (1) can be expressed as

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

where ε follows standard logistic distribution.

The partial likelihood function of $\boldsymbol{\beta}$ under the PO model is unavailable, we use Lam and Leung's (2001) method to estimate $\boldsymbol{\beta}$ by maximizing its marginal likelihood function: Let $T_{(1)} < \dots < T_{(K)}$ represent the ordered uncensored failure times in the sample and define $T_{(0)} = 0, T_{(K+1)} = \infty$. For $0 \leq k \leq K$, let L_k denote the set of labels i corresponding to those observations censored in the interval $[T_{(k)}, T_{(k+1)})$. The complete ranks of T_i 's are unknown as the result of the censoring scheme. Let \mathbf{R} denote the unobserved rank vector of T_i 's and let \mathbf{G} denote the collection of all possible rank vectors of T_i 's consistent with the observed data (\tilde{T}_i, θ_i) ($i = 1, \dots, n$). The marginal likelihood is then defined by $L_{n,M}(\boldsymbol{\beta}) = P(\mathbf{R} \in \mathbf{G})$, where the probability is with respect to the underlying uncensored version of the study. It can be shown that $L_{n,M}(\boldsymbol{\beta})$

can be represented as

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

where $V_{(k)} = H(T_{(k)})$, $k = 1, \dots, K$. $\Lambda(x)$ denote the cumulative hazard function of ε , i.e., $\Lambda(x) = \log\{1 + \exp(x)\}$ and $\lambda(x) = d\Lambda(x)/dx$.

Because there is no explicit solution to the maximization of (3), importance sampling method is used to approximate it. Following Lu and Zhang (2007), (3) can be estimated by performing a multiplication followed by a division by:

$$c \sum_{i=1}^n \{\lambda(V_{(k_i)})\}^{\theta_i} e^{-\Lambda(V_{(k_i)})}, \tag{4}$$

where the constant c is the total number of possible rank vectors in \mathbf{G} . When $V_i \equiv H(T_i)$ ($i = 1, \dots, n$) are independent and identically distributed according to the distribution function $F(x)$, and then it can be shown that (4) is the density function of $V_{(1)}, \dots, V_{(K)}$ under progressive type II censoring scheme. Then, the marginal likelihood (3) can be expressed as

$$L_{n,M}(\boldsymbol{\beta}) = E\{Q(V_{(1)}, \dots, V_{(K)}; \boldsymbol{\beta})\}, \tag{5}$$

where the expectation is with respect to the density (4) and

$$Q(V_{(1)}, \dots, V_{(K)}; \boldsymbol{\beta}) = \frac{1}{c} \prod_{i=1}^n \frac{\{\lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)\}^{\theta_i} e^{-\Lambda(V_{(k_i)} + \boldsymbol{\beta}'\mathbf{Z}_i)}}{\{\lambda(V_{(k_i)})\}^{\theta_i} e^{-\Lambda(V_{(k_i)})}}. \tag{6}$$

Then, (5) can be estimated by

$$\hat{L}_{n,M}(\boldsymbol{\beta}) = \frac{1}{B} \sum_{b=1}^B Q\{F^{-1}(U_{(1)}^b), \dots, F^{-1}(U_{(K)}^b); \boldsymbol{\beta}\}, \tag{7}$$

where $F^{-1}(\cdot)$ is the inverse of $F(\cdot)$ and $U_{(1)}^b, \dots, U_{(K)}^b$, $b = 1, \dots, B$, denote B independent instances of the uncensored order statistics derived from a random sample of size n , taken from a uniform distribution subject to the progressive type II censoring scheme.

3 Adaptive elastic net approach for proportional odds model

To address the challenges posed by the inherent instability of high-dimensional data and the limitation in handling collinearity, as originally encountered with the LASSO

and adaptive LASSO methods proposed by Lu and Zhang (2007), we have chosen to employ the adaptive elastic net penalty, as outlined in the work of Zou and Zhang (2009). This approach encompassed both the l_2 and weighted l_1 penalties and is applied to enhance the robustness of the marginal likelihood estimation

$$\min_{\beta} \left(-\frac{1}{n} \hat{l}_{n,M}(\beta) + \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \lambda_2 \sum_{j=1}^p w_j \beta_j^2 \right), \tag{8}$$

where $\hat{l}_{n,M}(\beta) = \log[\hat{L}_{n,M}(\beta)]$.

The tuning parameters, denoted as λ_1 and λ_2 , exercise control over the magnitudes of the l_1 (LASSO) and l_2 (ridge) penalty, respectively. Analogous to conventional regression scenarios, the distinguishing feature between l_1 and l_2 penalty lies in that l_2 penalty tends to yield small but non-zero coefficient estimates across all variables, whereas l_1 penalty is inclined to make some regression coefficients shrunk to exactly 0 and some other coefficients with comparatively little shrinkage. Combining l_1 and l_2 penalties is likely to give a result in between and an intermediate outcome, with fewer coefficient estimates set to 0 than in a pure LASSO setting, and more shrinkage for other coefficients.

The larger the tuning parameters λ_1 and λ_2 are, the greater degree of penalty or shrinkage is imposed upon the coefficients. In practice, it can be challenging to determine the appropriate values for these parameters. We propose to use BIC criterion to find the optimal values. The choice of tuning parameters is further discussed at the end of next section.

$w = (w_1, \dots, w_p)'$ is a non-negative weight vector, which adjusts penalties applied to the coefficients. As the weight value increase, the corresponding penalties are augmented accordingly. For important covariates, we take larger value weights; for unimportant covariates, we take small value weights. In practice, the weights are chosen adaptively by data. For example, any consistent estimator of β could be used as a good candidate (Zou, 2006).

Here, we denote the maximum marginal likelihood estimate (MMLE) of β as

$$\tilde{\beta} = \arg \max_{\beta} (\hat{L}_{n,M}(\beta)).$$

Lam and Leung (2001) have shown that $\tilde{\beta}$ is a consistent estimator of β . The absolute values of the elements in $\tilde{\beta}$ reflect the relative importance of the covariates. Hence, we set $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|}$.

We define our adaptive elastic net estimate for proportional odds model as

$$\hat{\beta} = \arg \min_{\beta} \left(-\frac{1}{n} \hat{l}_{n,M}(\beta) + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right), \tag{9}$$

where $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|}$.

If $\tilde{\beta}_j = 0$, then we assign $\hat{\beta}_j = 0$. When equal weights are used in (9), the adaptive elastic net estimate simplifies and transforms to elastic net estimate. If we set $\lambda_2 = 0$, then it leads to the simplification of the adaptive elastic net estimate into the adaptive LASSO estimate.

4 Computational algorithm and choice of tuning parameters

In this section, we discuss the computational algorithm to solve the adaptive elastic net challenges within the context of the proportional odds model. The algorithm involves 3 transformations. First, leveraging Taylor’s expansion, we convert the adaptive elastic net penalized likelihood problem into an adaptive elastic net penalized least square problem. Subsequently, we transform the adaptive elastic net penalized least square problem into an adaptive LASSO penalized least square problem. Finally, we transform the adaptive LASSO problem into an ordinary LASSO problem. Following these successive transformations, the ordinary LASSO problem can be easily solved using existing methods as described as LARS (Efron et al., 2004).

We define the gradient vector of $l(\beta)$ as $\nabla l(\beta) = -\partial \hat{l}_{n,M}(\beta) / \partial \beta$ and the Hessian matrix $\nabla^2 l(\beta) = -\partial^2 \hat{l}_{n,M}(\beta) / \partial \beta \beta'$. Let X be $\frac{1}{\sqrt{2n}}$ of the Cholesky decomposition of $\nabla^2 l(\beta)$, such that $X'X = \frac{1}{2n} \nabla^2 l(\beta)$. A pseudo-response vector Y is set as $Y = \frac{1}{2n} (X')^{-1} (\nabla^2 l(\beta) \beta - \nabla l(\beta))$.

Applying similar argument by Lu and Zhang (2007), we can show that $-\frac{1}{n} \hat{l}_{n,M}(\beta)$ can be approximated using the second-order Taylor expansion by $(Y - X\beta)'(Y - X\beta)$.

Therefore, we can solve (9) iteratively. First, we solve maximum marginal likelihood estimate $\tilde{\beta}$. Afterward, we compute $\nabla l(\tilde{\beta})$, $\nabla^2 l(\tilde{\beta})$, X and Y based on $\tilde{\beta}$. Then, we update β by minimizing

$$(Y - X\beta)'(Y - X\beta) + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \tag{10}$$

until it converges.

Next, we show that this adaptive elastic net problem can be transformed into an adaptive LASSO type problem in some augmented space. Building upon the approach outlined by Zou and Hastie (2005), we construct an artificial augmented data set (X^A, Y^A) , where

$$X^A_{(n+p) \times p} = (1 + \lambda_2)^{-\frac{1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \tag{11}$$

$$Y^A_{(n+p) \times 1} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix}. \tag{12}$$

Let $\gamma = \lambda_1 / \sqrt{1 + \lambda_2}$, $\beta^A = \sqrt{1 + \lambda_2} \beta$. For the adaptive elastic net solution, we have

$$\begin{aligned}
 & \arg \min_{\beta} (Y - X\beta)'(Y - X\beta) + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \\
 &= \arg \min_{\beta} \left\{ \left(\begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix} - \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \sqrt{1 + \lambda_2} \beta \right)' \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix} \right. \\
 & \quad \left. - \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \sqrt{1 + \lambda_2} \beta + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| \right\} \\
 &= \arg \min_{\beta} \left\{ \left(\begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix} - \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \sqrt{1 + \lambda_2} \beta \right)' \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix} \right. \\
 & \quad \left. - \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \sqrt{1 + \lambda_2} \beta + \frac{\lambda_1}{\sqrt{1 + \lambda_2}} \sum_{j=1}^p \hat{w}_j |\sqrt{1 + \lambda_2} \beta_j| \right\} \\
 &= \frac{1}{\sqrt{1 + \lambda_2}} \arg \min_{\beta^A} \left[(Y^A - X^A \beta^A)'(Y^A - X^A \beta^A) + \gamma \sum_{j=1}^p \hat{w}_j |\beta_j^A| \right]. \tag{13}
 \end{aligned}$$

At this juncture, the problem is reconfigured as an adaptive LASSO problem, featuring a tuning parameter denoted as $\gamma = \lambda_1/\sqrt{1 + \lambda_2}$, which is a convex optimization problem and does not suffer from the multiple local minima issue (Zou, 2006).

According to Zou (2006), the solution to an adaptive LASSO problem

$$\arg \min_{\beta} \left((y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right)$$

is $\hat{\beta}_j^{\text{alasso}} = \text{sign}(\hat{\beta}_j^{\text{ols}}) (|\hat{\beta}_j^{\text{ols}}| - \frac{1}{2} \hat{w}_j \lambda)_+$, for $j = 1, \dots, p$, where $\hat{\beta}_j^{\text{ols}}$ is the ordinary least square estimate and z_+ denotes the positive part of z , i.e., $z_+ = z$ if $z > 0$ and 0 otherwise.

By multiplying the LASSO solution by its respective weight, we derive the solution to $\hat{\beta}_j^{\text{ols}}$. Notably, established computational techniques are available for addressing LASSO problems, such as least angle regression (LARS) introduced by Efron et al. (2004) and path-wise coordinate descent algorithm proposed by Wu and Lange (2008).

Here, we use a modified shooting algorithm (Fu, 1998; Lu & Zhang, 2007) to solve the adaptive LASSO problem to avoid additional transformation and make computation more efficient. We define $G(\beta) = \sum_{i=1}^n (y_i - \beta'x_i)^2$, $\dot{G}_j(\beta) = \frac{\partial G(\beta)}{\partial \beta_j}$, $j = 1, \dots, p$, and denote β by $(\beta_j, \beta^{-j})'$ where β^{-j} is the $(p - 1)$ -dimensional vector consisting of the β_i 's other than β_j .

The complete algorithm to compute adaptive elastic net solution for proportional odds model when $p < n$ is given as follows:

$$\arg \min_{\beta} \left(-\frac{1}{n} \hat{l}_{n,M}(\beta) + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right). \tag{14}$$

1. Solve $\tilde{\beta}$ by maximizing $\hat{l}_{n,M}(\beta)$. Set $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|}$ for $j = 1, \dots, p$.
2. Let $k = 0$, and $\beta_j^{(0)} = 0$ for $j = 1, \dots, p$.
3. Compute $\nabla l, \nabla^2 l, X$ and Y based on current value of $\beta^{(k)}$.
4. Solve

$$\beta^{(k+1)} = \arg \min_{\beta} [(Y - X\beta)'(Y - X\beta) + \lambda_1 \sum_{j=1}^p \hat{w}_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2].$$

(a) Let

$$X_{(n+p) \times p}^A = (1 + \lambda_2)^{-\frac{1}{2}} \left(\frac{X}{\sqrt{\lambda_2 I}} \right), Y_{(n+p) \times 1}^A = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix} \text{ and } \gamma = \lambda_1 / \sqrt{1 + \lambda_2}.$$

(b) Solve $\hat{\beta}^A = \arg \min_{\beta^A} [(Y^A - X^A \beta^A)'(Y^A - X^A \beta^A) + \gamma \sum_{j=1}^p \frac{|\beta_j^A|}{|\tilde{\beta}_j|}]$.

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

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

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

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

(c) Set $\beta^{k+1} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}^A$.

5. If $\|\beta^{k+1} - \beta^k\|^2 < 0.0001$ (or other given small $\varepsilon > 0$), then stop, else set $k = k + 1$ and go to 3.

For $p \geq n$ cases, the MMLE of β is not available. Following Zou and Zhang (2009), we use elastic net estimates to construct the adaptive weight \hat{w}_j . We first apply the algorithm with the initial adaptive weight $\hat{w}_j^{(0)} = 1$ for $j = 1, \dots, p$ to get the elastic net estimates $\hat{\beta}_j^{\text{enet}}$, then set $\hat{w}_j = (|\hat{\beta}_j^{\text{enet}}| + \frac{1}{n})^{-1}$ and run step 2 through 5 to get the adaptive elastic net solution.

Tuning is a very important aspect of model fitting. For adaptive elastic net approaches, we need to find the optimal value of λ_1 and λ_2 . We use Bayesian Information Criterion (BIC) (Schwarz, 1978) to choose the best combination of λ_1 and λ_2 .

We define the BIC for proportional odds model as

$$\text{BIC} = -2\hat{L}_{n,M}(\hat{\boldsymbol{\beta}}^{\text{aenet}} \mid \lambda_1, \lambda_2) + k \log(n), \quad (15)$$

where k is the total number of non-zero parameters and n is the number of observations.

The typical way to deal with two tuning parameters in adaptive elastic net problem is to pick a relatively small grid of values for λ_2 , for example (0, 0.001, 0.01, 0.1, 1, 10). Then, for each value of λ_2 , we get the BIC scores for a sequence of λ_1 . The chosen (λ_1, λ_2) is the pair that gives the smallest BIC score.

5 Simulation studies

We generate data from proportional odds models and apply adaptive elastic net procedure to do variable selection. For each model, we generate 100 simulated datasets and gauge the variable selection performance using $(C0, IC0)$, where $C0$ is the number of unimportant covariates that the procedure correctly estimates as zero and $IC0$ is the number of important covariates that the procedure incorrectly estimates as zero. To measure prediction accuracy, we follow Tibshirani (1996) to summarize the average mean square error (MSE) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{V} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ over the 100 runs, where \mathbf{V} is the population covariance matrix of the covariates. BIC method is used to choose tuning parameters. The simulation is run in no censoring, 20% censoring rate and 40% censoring rate settings, respectively. Also, 3 sample sizes, $n = 100$, $n = 200$, and $n = 500$ are used for each model. The results are then compared with LASSO, adaptive LASSO, and elastic net. In our implementation, we set $\lambda_2 = 0$ in the adaptive elastic net to get the adaptive LASSO fit. To get the elastic net fit, we set $w_j = 1$ for $j = 1, 2, \dots, p$. For these 3 methods, BIC method is also used to select the tuning parameter. Five models with different $\boldsymbol{\beta}$ and Pearson's correlation coefficient $\rho_{i,j}$ are used for our simulation studies. The results are as follows:

Model 1: The design contains ten covariates: $(Z_1, Z_2, \dots, Z_{10})$. The covariates are marginally standard normal distributed and $\rho_{i,j} = 0.2$ for $i, j = 1, 2, \dots, 10$ and $i \neq j$. $\boldsymbol{\beta}' = (-0.8, 0, 0, -0.8, 0, 0, -0.7, 0, 0, -0.7)$. Therefore, Z_1, Z_4, Z_7 and Z_{10} are important variables. This model is used to compare the performance of adaptive elastic net and other 3 procedures in a scenario that important covariates all have large effects and that the pairwise correlations between the covariates are weak. The simulation result is summarized and shown in Fig. 1.

For model 1, when sample size is small ($n = 100$), the performances of two procedures with oracle properties (Zou, 2006), adaptive LASSO and adaptive elastic net, are comparable in terms of accuracy. LASSO and elastic net are outperformed by their counterparts. Regarding mean square error, the adaptive elastic net approach is better than any of the other 3 approaches (around 24.5% smaller MSE on average compared to the next best approach). Adaptive LASSO and elastic net have similar MSE. LASSO is the one that have the largest MSE. When the sample size is increased to 200, the difference in MSE for adaptive LASSO and adaptive elastic net becomes small (around 10% on average). When sample size is increased to 500, adaptive LASSO

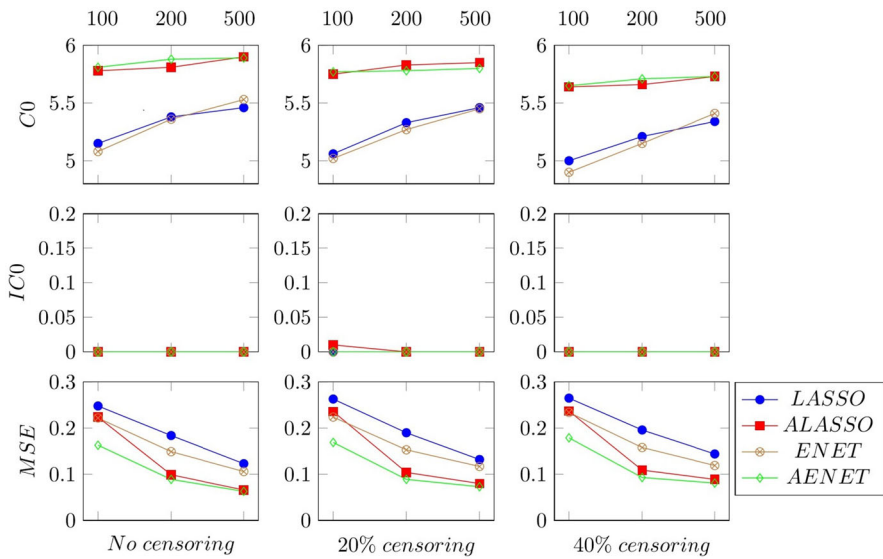


Fig. 1 Variable selection results and MSE for $\rho = 0.2$ and $\beta' = (-0.8, 0, 0, -0.8, 0, 0, -0.7, 0, 0, -0.7)$. The MSE is the averaged value over 100 replicates

and adaptive elastic net perform equally good in terms of accuracy. However, adaptive elastic net is still slightly better than adaptive LASSO in terms of MSE (around 6% on average). For this model, we conclude that adaptive LASSO and adaptive elastic net are the two best approaches in terms of selection accuracy. Also, the adaptive elastic net does better than adaptive LASSO in terms of MSE when sample size is small.

Model 2: The design contains ten covariates: $(Z_1, Z_2, \dots, Z_{10})$. The covariates are marginally standard normal distributed and $\rho_{i,j} = 0.8$ for $i, j = 1, 2, \dots, 10$ and $i \neq j$. $\beta' = (-0.8, 0, 0, -0.8, 0, 0, -0.7, 0, 0, -0.7)$. This model is used to compare the performance of the 4 procedures in a scenario that important covariates all have large effects and that the pairwise correlations between the covariates are strong. The simulation result is summarized and shown in Fig. 2.

For model 2, when sample size is 100, adaptive elastic net and adaptive LASSO seem to have the largest accuracy rate and adaptive elastic net has the smallest MSE (around 3% smaller on average than the elastic net, which has the second smallest MSE). At high censoring rate, adaptive elastic net and adaptive LASSO tend to shrink more important variables to 0 than LASSO and adaptive LASSO do. When sample size is increased to 500, none of the 4 procedures misses any important variables and adaptive LASSO and adaptive elastic net have comparable performance in terms of accuracy. Adaptive elastic net is still the best in terms of MSE (around 7% smaller on average). We conclude that adaptive elastic net method has the best overall performance for model 2.

Model 3: The design contains ten covariates: $(Z_1, Z_2, \dots, Z_{10})$. The covariates are marginally standard normal distributed and $\rho_{i,j} = 0.2$ for $i, j = 1, 2, \dots, 10$ and $i \neq j$. $\beta' = (-0.3, 0, 0, -0.3, 0, 0, -0.2, 0, 0, -0.2)$. This model is used to compare

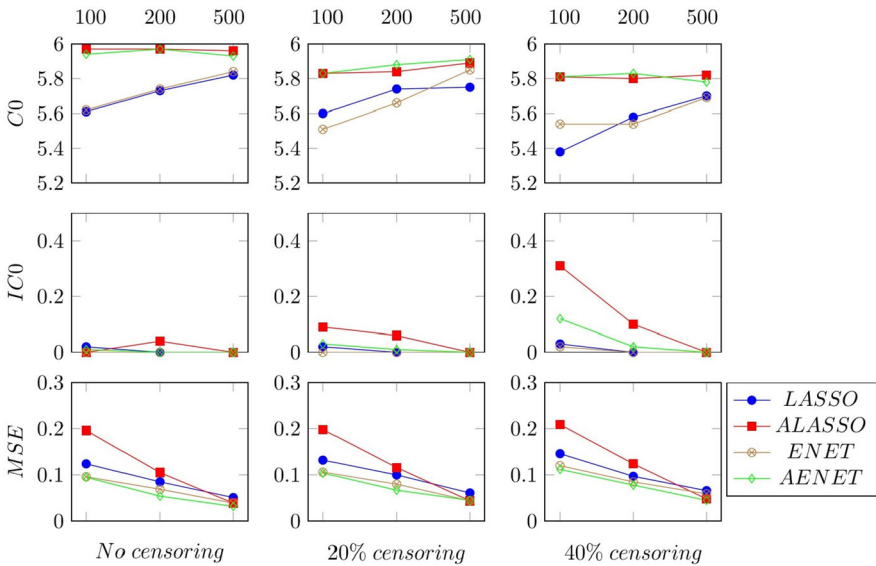


Fig. 2 Variable selection results and MSE for $\rho = 0.8$ and $\beta' = (-0.8, 0, 0, -0.8, 0, 0, -0.7, 0, 0, -0.7)$. The MSE is the averaged value over 100 replicates

the performance of the 4 procedures in a scenario that important covariates all have small effects and that the pairwise correlations between the covariates are low. The simulation result is summarized and shown in Fig. 3.

When sample size is 100, adaptive LASSO seems to have the best accuracy in terms with dropping unimportant variables, but it also tends to drop important variables more than the other 3 methods. On the contrary, elastic net selects the least number of unimportant variables, but it does the worst job in keeping important variables. The adaptive elastic net is very closed to adaptive LASSO in keeping no-zero variables (7% less accurate rate) and is almost as good as elastic net in eliminating zero variables. Also, adaptive elastic net is consistently the best among the 4 approaches in terms of MSE (around 24% less than elastic net, the approach that has the second smallest MSE). As sample size increases to 200, the difference in correct 0s as well as MSE gets closer between adaptive elastic net and adaptive LASSO. When sample size gets to 500, adaptive elastic net outperforms adaptive LASSO in dropping unimportant variables. Elastic net is still the best in keeping important variables. Adaptive elastic net still has the least MSE. Considering all 3 factors, we conclude that adaptive elastic net is the best approach for this scenario.

Model 4: The design contains ten covariates: $(Z_1, Z_2, \dots, Z_{10})$. The covariates are marginally standard normal distributed and $\rho_{i,j} = 0.8$ for $i, j = 1, 2, \dots, 10$ and $i \neq j$. $\beta' = (-0.3, 0, 0, -0.3, 0, 0, -0.2, 0, 0, -0.2)$. This model is used to compare the performance of the 4 procedures in a scenario that important covariates all have small effects and that the pairwise correlations between the covariates are strong. The simulation result is summarized and shown in Fig. 4.

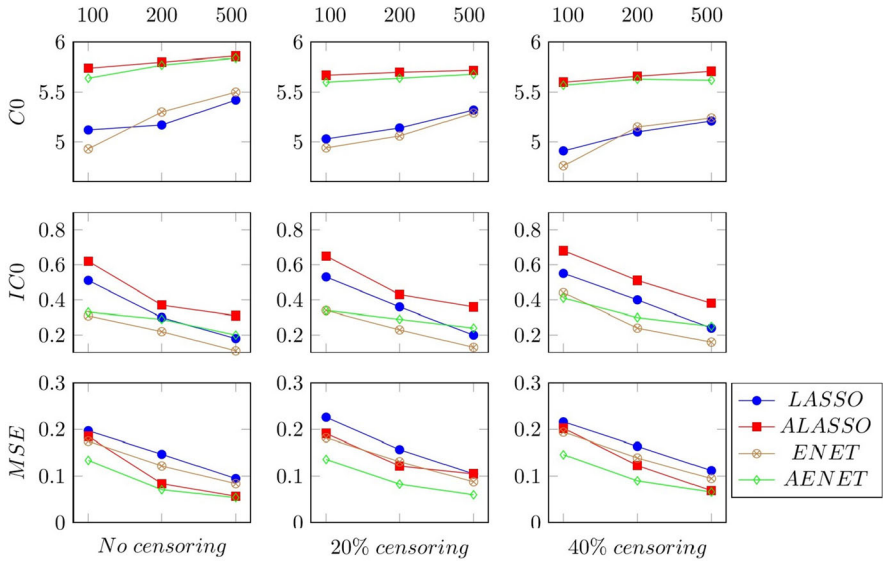


Fig. 3 Variable selection results and MSE for $\rho = 0.2$ and $\beta' = (-0.3, 0, 0, -0.3, 0, 0, -0.2, 0, 0, -0.2)$. The MSE is the averaged value over 100 replicates

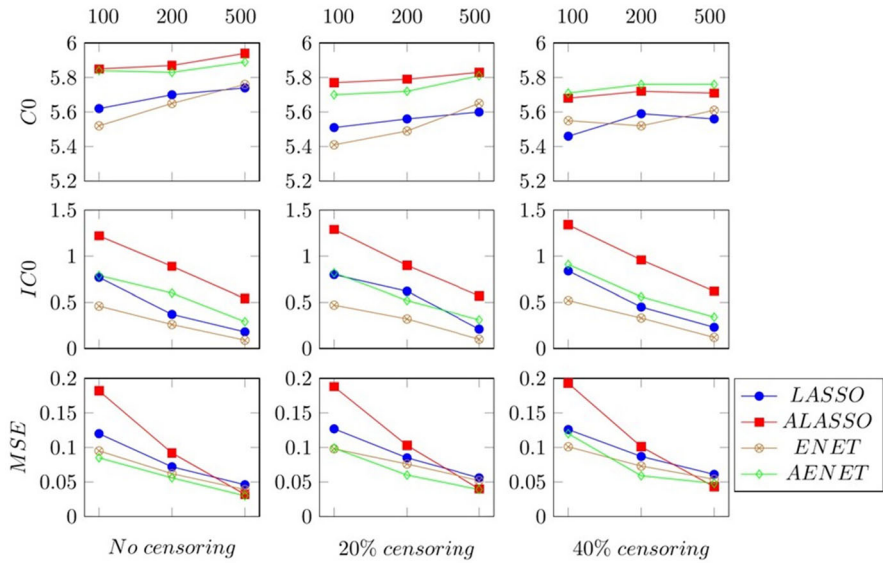


Fig. 4 Variable selection results and MSE for $\rho = 0.8$ and $\beta' = (-0.3, 0, 0, -0.3, 0, 0, -0.2, 0, 0, -0.2)$. The MSE is the averaged value over 100 replicates

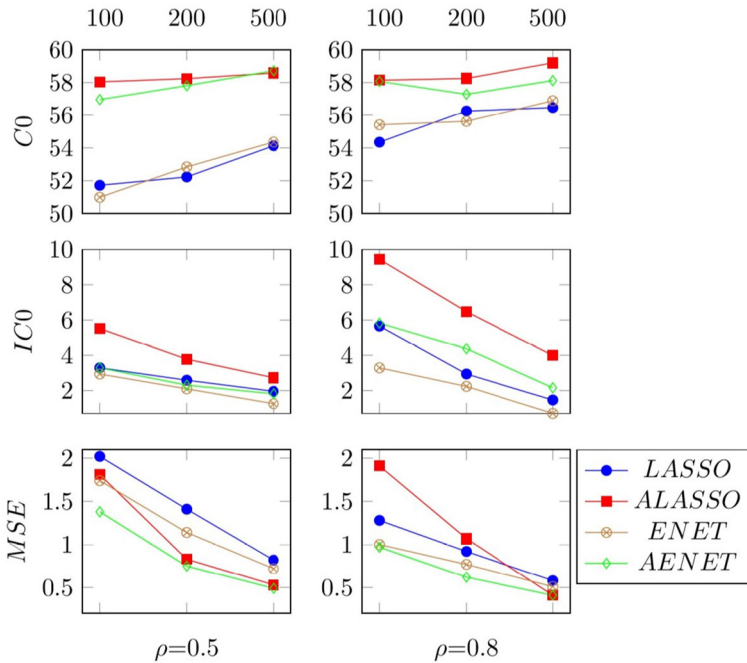


Fig. 5 Variable selection results and MSE for $\rho = 0.5, 0.8$ and $\beta' = (-0.8\mathbf{I}_{10}, -0.7\mathbf{I}_{10}, -0.3\mathbf{I}_{10}, -0.2\mathbf{I}_{10}, 0_{60})$ censoring rate 20% The MSE is the averaged value over 100 replicates

For model 4, elastic net is still the best in retaining the none-zero variables in the model and adaptive LASSO and adaptive elastic net is the best in eliminating zero variables out of the model. For small sample size, adaptive elastic net has the smallest MSE (around 3% difference on average). At large sample size, adaptive elastic net and adaptive LASSO have the smallest MSE. For this model, we conclude that adaptive elastic net has overall the best performance when sample size is small. When sample size is large, elastic net is the best procedure in terms of selection accuracy and adaptive elastic net is the best procedure in terms of MSE.

Model 5: The design contains 100 marginally standard normal distributed covariates: $(Z_1, Z_2, \dots, Z_{100})$. $\beta' = (-0.8\mathbf{I}_{10}, -0.7\mathbf{I}_{10}, -0.3\mathbf{I}_{10}, -0.2\mathbf{I}_{10}, 0_{60})$. So Z_1, Z_2, \dots, Z_{20} are important variables having large effects, $Z_{21}, Z_{22}, \dots, Z_{40}$ are important variables having small effects, and $Z_{41}, Z_{42}, \dots, Z_{100}$ are unimportant variables. We use this model to compare the performance of the 4 procedures in a complicated case that have large dimension of covariates and the important covariates have both small and large effects. We run this model with 20% and 40% censoring rate and consider pairwise correlation coefficient in 0.5 and 0.8 settings. Result is summarized and shown in Figs. 5 and 6.

For this complex model, the adaptive elastic net demonstrates the lowest MSE in 10 out of 12 combinations of sample sizes, correlation coefficients, and censoring rates. Adaptive LASSO and adaptive elastic net are the two best approaches in terms of dropping correct zero variables. Elastic net outperforms the other 3 methods in

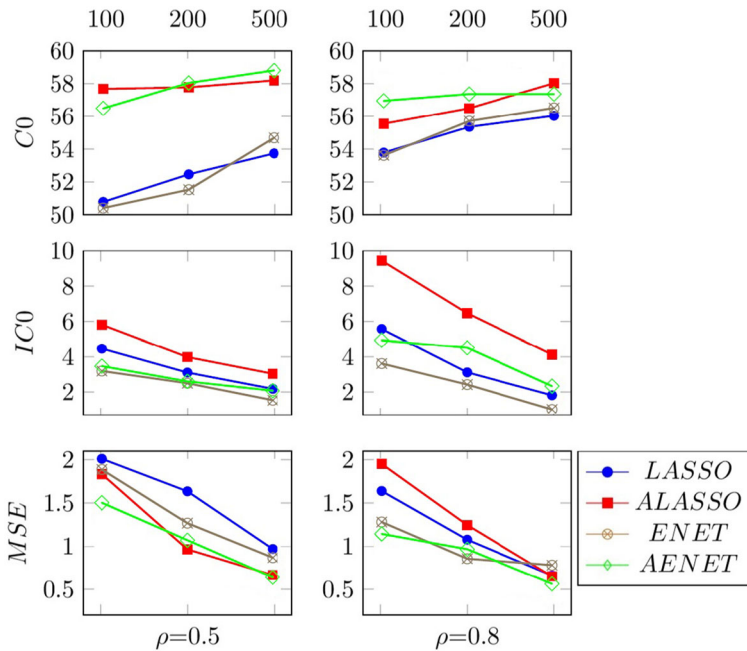


Fig. 6 Variable selection results and MSE for $\rho = 0.5, 0.8$ and $\beta' = (-0.8\mathbf{I}_{10}, -0.7\mathbf{I}_{10}, -0.3\mathbf{I}_{10}, -0.2\mathbf{I}_{10}, 0_{60})$ censoring rate 40% The MSE is the averaged value over 100 replicates

keeping non-zero variables. If penalizing the selection of unimportant variables and the omission of important ones carry the same weight, then the adaptive elastic net is considered the most favorable approach for this complex model.

6 Application in real data

6.1 Veteran cancer data

We apply the adaptive elastic net method to the data from the Veteran's Administration lung cancer trial (Prentice & Kalbfleisch, 2002). 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; prior therapy (0 = no, 10 = yes).

We include all the covariates and all the patients in our analysis and compute the adaptive elastic net estimates under the proportional odds model. Maximum marginal likelihood, LASSO, adaptive LASSO, and elastic net estimates are also computed.

Table 1 summarizes the estimated coefficients estimated by the 4 approaches. The maximum marginal likelihood estimates are in good agreement with those reported in Lam and Leung (2001) and Lu and Zhang (2007). The LASSO selects cell type

Table 1 Estimated coefficients for lung cancer data

Covariate	MMLE	LASSO	ALASSO	ENET	AENET
Treatment	0.132	0	0	0	0
Squamous versus large	0.023	-0.059	0	0	0
Small versus large	0.927	0.612	0.608	0.503	0.516
Adeno versus large	1.122	0.732	0.811	0.682	0.713
Karnofsky	-0.056	-0.053	-0.051	-0.049	-0.049
Month from diagnosis	0.002	0	0	0	0
Age	-0.011	0	0	0	0
Prior therapy	0.004	0	0	0	0

(squamous versus large, small versus large, and adeno versus large) and Karnofsky score as important variables, while all other three methods eliminate one more variable (squamous versus large).

We use k -fold cross-validation, a standard resampling procedure to evaluate and compare the models in general machine learning and survival analysis. In the cross-validation, the dataset is randomly partitioned into k blocks, which are used, respectively, as test set in each cross-validation iteration. The remaining $k - 1$ blocks are combined and used as the training set to fit the model. Concordance index proposed by Zheng and Heagerty (2005) is used to evaluate the performance of the models in the cross-validation procedure. The concordance index is defined as

$$CI = 1 - \frac{1}{|\varepsilon|} \sum_{\{i:\delta_i=1\}} \sum_{t_i < t_j} \left(\mathbf{1}_{f(x_i) < f(x_j)} + \frac{1}{2} \mathbf{1}_{f(x_i) = f(x_j)} \right), \tag{16}$$

where ε is defined as the set of all pairs (t_i, t_j) with $i, j = 1, \dots, n$ where it can be concluded that $t_i < t_j, t_i = t_j,$ or $t_i > t_j,$ despite censoring; $f(x)$ denotes the predicted survival time of the event given covariate vector x .

The concordance index is the probability of concordance of observed and predicted survival time and can be interpreted as the portion of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered. Values of concordance index that close to 0 mean nearly perfect prediction; values of concordance index that close to 0.5 are signs of random predictions.

Here, we apply threefold cross-validation, the reason why we use a low number of folds is because we need a large enough number of samples in the test set to get a meaningful concordance index. In each iteration of the cross-validation, the 4 selection procedures are used to build the models and the concordance indices (CI) are calculated using the linear predictor $X_{\text{test}} \hat{\beta}_{\text{train}}$. The CI for the cross-validation is the average of the CIs of the 3 iterations. We run the cross-validation procedure 3 times and take the average of concordance indices to reduce the performance error due to randomization.

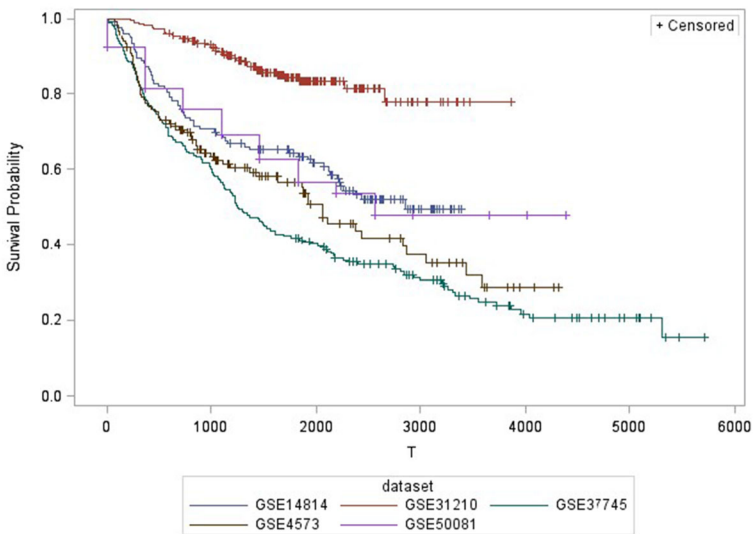
The results of cross-validation are shown in Table 2; we can see that the adaptive elastic net has the lowest mean concordance index, meaning that it has the best performance among the 4 methods in the cross-validation.

Table 2 Mean concordance indices comparison for VC data

Method	LASSO	ALASSO	ENET	AENET
Mean CI	0.363	0.352	0.327	0.305

Table 3 Characteristics of lung cancer datasets

Dataset	Chip	Patients	Events	Gene covariates
GSE4573	HG-U133-A	129	67	22,283
GSE14814	HG-U133-A	90	38	22,283
GSE31210	HG-U133-Plus2	226	35	54,675
GSE37745	HG-U133-Plus2	194	143	54,675
GSE50081	HG-U133-Plus2	181	75	54,675

**Fig. 7** Kaplan–Meier estimator of survival for 5 GSE datasets

6.2 GSE data

We select 5 lung cancer data sets with genome wide gene expression measurements and additional clinical information from Gene Expression Omnibus: GSE4573 (Beer, 2006), GSE14814 (Tsao, 2010), GSE31210 (Gotoh, 2012), GSE37745 (Micke, 2013), and GSE50081 (Tsao, 2014). Table 3 and Fig. 7 show the characteristics of the datasets and the Kaplan–Meier estimates of survival times, respectively. As we can see, GSE14814, GSE50081, and GSE4573 have comparable estimated survival curves, GSE37745 has slightly lower survival times over time, and GSE31210 has significantly higher survival rate than the other 4 datasets.

In addition to the gene covariates, important clinical variables, including sex, age, stage, and histology, are also used in the analysis. We remove all the incomplete

Table 4 Mean concordance indices for 5 GSE datasets

Dataset	LASSO	ALASSO	ENET	AENET
GSE4573	0.59	0.53	0.45	0.51
GSE14814	0.48	0.51	0.45	0.47
GSE31210	0.31	0.26	0.28	0.22
GSE37745	0.52	0.54	0.48	0.47
GSE50081	0.44	0.37	0.34	0.30
Mean	0.476	0.448	0.400	0.388

The bold indicates the smallest mean concordance index for the GSE dataset

observations before we apply the model selection methods. For all the 5 datasets, we evaluate the LASSO, adaptive LASSO, elastic net, and adaptive elastic net methods through threefold cross-validation. The CI for the cross-validation is the average of the CIs of the three iterations. Again, we run the cross-validation procedure 3 times and take the average of concordance indices to reduce the performance error due to randomization. The results are shown in Table 4.

We can see that though the 5 datasets are taken from the same field, to fit a predictive model on them is unequally difficult. This is quite obvious by comparing the concordance indices across the 5 datasets. For datasets GSE4573, GSE14814, and GSE37745, the concordance indices of cross-validation are around 0.5 for all model selection procedures that we are comparing, which means that the prediction is nearly random. Adaptive elastic net does not seem to perform better in terms of prediction than other methods. However, for datasets GSE31210 and GSE50081, it is much easier to build a model to predict the survival. In these two datasets and the adaptive elastic net method does perform better in prediction than other methods. For GSE31210, the concordance index for adaptive elastic net method is 0.04 (15.3%) smaller than the next best method, and for GSE50081, the concordance index is also 0.04 (11.8%) smaller than the next best method. The mean concordance index for adaptive elastic net, across all the 5 datasets, is 0.012 (3%) smaller than the next best approach.

7 Summary and discussion

In this paper, we have studied the application of adaptive elastic net for variable selection problem under the proportional odds model and compared its performance with LASSO, adaptive LASSO, and elastic net. Our simulation results show that the adaptive elastic net method has superior result in terms of accuracy of variable selection and MSE in most of the cases. The simulation also indicates that as the censoring rate increases, all the approaches tend to have higher error rate in selection and higher MSE, but the relative ranks of their performance do not change. Our proposed method naturally exhibits greater complexity compared to LASSO, adaptive LASSO, and elastic net. As a result, when conducting a comparison with those three methods, it should be anticipated that our method will demand more computation time, with the percentage varying across different scenarios. Nevertheless, taking into account its

superior performance in the majority of scenarios, our method continues to be an efficient approach. Moreover, because adaptive elastic net has the oracle properties (Zou, 2006), the bias of its coefficient estimates tends to be zero when the number of samples goes to infinity. For finite samples, because of the nature of shrinkage, the adaptive elastic net estimator may have obvious bias. Therefore, in real applications, it may be a good choice to do variable selection and estimation separately: We first eliminate unimportant variables using the adaptive elastic net procedure, and then fit the model using classical method such as maximum marginal likelihood estimation to get the coefficient estimates.

Data Availability The data that support the findings of this study are openly available in R project [<https://search.r-project.org/CRAN/refmans/ncvreg/html/Lung.html>] and NCBI Gene Expression Omnibus [<https://www.ncbi.nlm.nih.gov/geo/>].

Declarations

Conflicts of interest There is no conflicts of interest in this research. It has no involving Human Participants and/or Animals.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Beer, D. G. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research*, 66(1), 7466–7472.
- Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2), 273–277.
- Collett, D. (1994). *Modelling survival data in medical research* (2nd ed.). Chapman and Hall.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2), 187–220.
- Crowder, M.J. (1991). *Statistical analysis of reliability data* (1st ed.). Routledge. <https://doi.org/10.1201/9780203738726>.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–451.
- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.
- Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.
- Gotoh, N. (2012). Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS One*, 7, 43923.
- Grønneberg, S., & Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics*, 41(2), 436–459.

- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4), 757–796.
- Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in medicine*, 13(10), 1045–62.
- Lam, K. F., & Leung, T. L. (2001). Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis*, 7(1), 39–54.
- Lu, W., & Zhang, H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, 26, 3771–81.
- Mallows, C. L. (1973). Some comments on c_p . *Technometrics*, 15(4), 661–675.
- Micke, P. (2013). Biomarker discovery in non-small cell lung cancer: Integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical Cancer Research*, 19, 194–204.
- Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(2), 205–217.
- Prentice, R. L., & Kalbfleisch, J. D. (2002). *The statistical analysis of failure time data* (2nd ed.). Wiley.
- Rossini, A. J., & Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(434), 713–721.
- Salerno, S., & Li, Y. (2022). High-dimensional survival analysis: Methods and applications. *Annual Review of Statistics and Its Application*, 10(1), 25–49.
- Schneider, U., & Wagner, M. (2012). Catching growth determinants with the adaptive lasso. *German Economic Review*, 13(1), 71–85.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3), 485–493.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
- Tsao, M.-S. (2010). Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *Journal of Clinical Oncology*, 28(10), 4417–4424.
- Tsao, M.-S. (2014). Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *Journal of Thoracic Oncology*, 9, 59–64.
- Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2, 224–244.
- Zheng, Y., & Heagerty, P. J. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), 92–105.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
- Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4), 1733–1751.