Stat Comput (2016) 26:725–741
DOI 10.1007/s11222-015-9555-8

Variable selection for survival data with a class of adaptive elastic net techniques

Md Hasinur Rahaman Khan · J. Ewart H. Shaw

Received: 17 April 2014 / Accepted: 2 March 2015 / Published online: 17 March 2015

Abstract The accelerated failure time (AFT) models have proved useful in many contexts, though heavy censoring (as for example in cancer survival) and high dimensionality (as for example in microarray data) cause difficulties for model fitting and model selection. We propose new approaches to variable selection for censored data, based on AFT models optimized using regularized weighted least squares. The regularized technique uses a mixture of ℓ_1 and ℓ_2 norm penalties under two proposed elastic net type approaches. One is the adaptive elastic net and the other is weighted elastic net. The approaches extend the original approaches proposed by Ghosh (Adaptive elastic net: an improvement of elastic net to achieve oracle properties, Technical Reports 2007) and Hong and Zhang (Math Model Nat Phenom 5(3):115–133 2010), respectively. We also extend the two proposed approaches by adding censoring observations as constraints into their model optimization frameworks. The approaches are evaluated on microarray and by simulation. We compare the performance of these approaches with six other variable selection techniques-three are generally used for censored data and the other three are correlation-based greedy methods used for high-dimensional data.

Keywords Adaptive elastic net · AFT · Variable selection · Stute's weighted least squares · Weighted elastic net

M. H. R. Khan (✉)
Applied Statistics, Institute of Statistical Research and Training, University of Dhaka, Dhaka 1000, Bangladesh
e-mail: hasinur@isrt.ac.bd

J. E. H. Shaw
Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

1 Introduction

The practical importance of variable selection is huge and well recognized in many disciplines, and has been the focus of much research. Several variable selection techniques have been developed for linear regression models; some of these have been extended to censored survival data. The methods include stepwise selection (Peduzzi et al. 1980) and penalized likelihood-based techniques, such as Akaike's information criterion (AIC) (Akaike 1973), bridge regression (Frank and Friedman 1993), least absolute shrinkage and selection operator (lasso) (Tibshirani 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), least angle regression selection (LARS) (Efron et al. 2004), the elastic net (Zou and Hastie 2005), MM algorithms (Hunter and Li 2005) that are based on extensions of the well-known class of EM algorithms, group lasso (Yuan and Lin 2006), the Dantzig selector (Candes and Tao 2007) that is based on a selector that minimizes the ℓ_1 norm of the coefficients subject to a constraint on the error terms, and MC+ (Zhang 2010) that is based on a minimax concave penalty and penalized linear unbiased selection. Stability selection as proposed in Meinshausen and Bühlmann (2010) is a variable selection technique that is based on subsampling in combination with (high-dimensional) selection algorithms. It is also used as a technique to improve variable selection performance for a range of selection methods.

Recently, there has been a surge of interest in variable selection with ultra-high dimensional data. By ultra-high dimension, Fan and Lv (2008) meant that the dimensionality grows exponentially in the sample size, i.e., $\log(p) = O(n^a)$ for some $a \in (0, 1/2)$. The issue of high correlations among the variables for variable selection with ultra-high dimensional data has been dealt with using various greedy approaches. For linear regression with ultra-high dimensional datasets, Fan and Lv (2008) proposed sure independence screening (SIS) based on marginal correlation ranking. Bühlmann et al. (2010) proposed the PC-simple algorithm that uses partial correlation to infer the association between each variable and the response conditional on other variables. Radchenko and James (2011) proposed the forward-lasso adaptive shrinkage (FLASH) that includes the lasso and forward selection as special cases at two extreme ends. Cho and Fryzlewicz (2012) proposed a *tilting* procedure that provides an adaptive choice between the use of marginal correlation and tilted correlation for each variable, where the choice is made depending on the values of the hard-thresholded sample correlation of the design matrix.

The Cox model (Cox 1972) with high-dimensional data has been the focus of many variable selection studies. For example, Tibshirani (1997) developed a regularized Cox regression by minimizing an ℓ_1 lasso penalty to the partial likelihood, Faraggi and Simon (1998) proposed a Bayesian variable selection method, Fan and Li (2002) developed a non-concave penalized likelihood approach, Li and Luan (2003) used kernel transformations, Gui and Li (2005) introduced a threshold gradient descent regularization estimation method, and Antoniadis et al. (2010) developed a variable selection approach for the Cox model based on the Dantzig selector.

There are also some variable selection studies for AFT models. For example, Huang et al. (2006) used the lasso regularization for estimation and variable selection in the AFT model based on the inverse probability of censoring. The lasso-regularized Buckley–James method for the AFT model is investigated by Huang and Harrington (2005) and Datta et al. (2007). Sha et al. (2006) developed a Bayesian variable selection approach. Variable selection using the elastic net is investigated in Wang et al. (2008). Engler and Li (2009) and Cai et al. (2009) proposed variable selection using the lasso-regularized rank-based estimator. Huang and Ma (2010) used a bridge method for variable selection. Hu and Rao (2010) proposed a sparse penalization technique with censoring constraints. Recently, Khan and Shaw (2013c) proposed a variable selection technique for AFT model that is based on the synthesis of the Buckley–James method and the Dantzig selector.

In this paper, we consider variable selection methods for the AFT modeling of censored data and propose new regularized Stute's weighted least squares (SWLS) approaches. We introduce classes of elastic net-type regularized variable selection techniques based on SWLS. The classes include an adaptive elastic net, a weighted elastic net, and two extended versions that are carried out by introducing censoring constraints into the optimization function.

The rest of the paper is structured as follows. Section 2 provides the regularized framework of SWLS. Section 3 provides proposed variable selection methods including variable selection criteria and prediction formula. All the methods are demonstrated with two simulated examples in Sect. 4 and with one microarray real data example in Sect. 5. In Sects. 4 and 5, we also compare the performance of the proposed approaches with three other variable selection approaches: the typical elastic net implemented for weighted data, the adaptive elastic net for censored data (Engler and Li 2009), and a Bayesian approach (Sha et al. 2006) and three other correlation-based greedy variable selection methods generally used for high-dimensional data: sure independence screening (Fan and Lv 2008), tilted correlation screening (Cho and Fryzlewicz 2012), and PC-simple (Bühlmann et al. 2010).

2 Methodology

2.1 Regularized Stute's weighted least squares (SWLS) for AFT models

The objective function of Stute's weighted least squares Stute (1993, 1996) for a typical AFT model is given by

$$\underset{(\alpha,\beta)}{\arg\min}\left[\sum_{i=1}^{n} w_i\,(Y_{(i)} - \alpha - X_{(i)}^T\beta)^2\right], \tag{1}$$

where Y_i is the log survival time for the i-th observation, $\mathbf{X}$ is the covariate vector, α is the intercept term, β is the unknown $p \times 1$ vector of true regression coefficients, and w_i are the Kaplan–Meier (K–M) weights that are obtained by

$$w_1 = \frac{\delta_{(1)}}{n},\ w_i = \frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}},\ i = 2,\cdots,n. \tag{2}$$

Under some regularity conditions, the WLS estimator with K−M weights (1) is consistent and asymptotically normal Stute (1993, 1996) for fixed p. As discussed in Huang et al. (2006), the SWLS method is computationally more amenable to high-dimensional covariates than the B–J estimator (Buckley and James 1979) or a rank-based estimator e.g., (Ying 1993). This is because the OLS structure makes it computationally efficient to apply a regularized method in the AFT model. The method has rigorous theoretical justifications under reasonable assumptions.

In matrix notation, the objective function of SWLS (1) is given by

$$(Y - \alpha - X\beta)^T w\,(Y - \alpha - X\beta), \tag{3}$$

where w is the $n \times n$ diagonal weight matrix.

2.2 Censoring constraints

Let the uncensored and censored data be subscripted by u and $\bar{u}$, respectively. Thus the number of uncensored and censored observations is denoted by n_u and $n_{\bar{u}}$, the predictor and response observations for censored data by $X_{\bar{u}}$ and $Y_{\bar{u}}$, and the unobserved true failure time for censored observation by $T_{\bar{u}}$. Since under right censoring, $Y_{\bar{u}} < \log(T_{\bar{u}})$ that is equivalent to $Y_{\bar{u}} \leq \alpha + X_{\bar{u}}\,\beta$ can be added to the SWLS objective function (3). These constraints are called censoring constraints (Hu and Rao 2010) which may be too stringent due to the random noise. This might suggest modifying the constraints to $Y_{\bar{u}} \leq \alpha + X_{\bar{u}}\,\beta + \xi$, where ξ is a vector of non-negative values that measure the severities of violations of the constraints. The SWLS objective function now can be defined by

$$L(\alpha,\ \beta) = (Y_u - \alpha - X_u\,\beta)^T\, w_u\,(Y_u - \alpha - X_u\,\beta) + \frac{\lambda_0}{n}\,\xi^T\xi,$$
$$\text{subject to } Y_{\bar{u}} \leq \alpha + X_{\bar{u}}\,\beta + \xi, \tag{4}$$

where λ_0 is a positive value that accounts for the penalties of violations of constraints, and n is included for scaling to match the w_u. An alternative form of (4) is given by

$$L(\alpha,\ \beta) = (Y_u - \alpha - X_u\,\beta)^T\, w_u\,(Y_u - \alpha - X_u\,\beta) + \lambda_0\,(Y_{\bar{u}} - \alpha - X_{\bar{u}}\,\beta)^T\, w_{\bar{u}}\,(Y_{\bar{u}} - \alpha - X_{\bar{u}}\,\beta), \tag{5}$$

where $w_{\bar{u}}$ is a diagonal matrix with diagonal element $w_{\bar{u}i,i} = \frac{1}{n}$ if $Y_{\bar{u}i} > \alpha + X_{\bar{u}i}\,\beta$ and 0 otherwise. So, for a censored observation $(X_{\bar{u}i},\ Y_{\bar{u}i})$, when its estimated survival time is not less than the censoring time, the difference between the estimated survival time and the censoring time is ignored. Otherwise, the difference will be added into the objective function just like the residuals for uncensored observations. Since the objective function in (5) is not continuously differentiable, we choose to use (4) in which we expand the dimension of the model space to include additional parameters, ξ, and thus to make the objective function continuously differentiable in a constrained (β, ξ) space.

An intercept term α typically is included in the AFT model. However, for notational convenience, we can remove α by (weighted) standardization of the predictors and response. The weighted means are defined by

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i X_{(i)}}{\sum_{i=1}^n w_i}, \qquad \bar{Y}_w = \frac{\sum_{i=1}^n w_i Y_{(i)}}{\sum_{i=1}^n w_i}.$$

Then the adjusted predictors and responses are defined by

$$X_{(i)}^w = (w_i)^{1/2}(X_{(i)} - \bar{X}_w), \qquad Y_{(i)}^w = (w_i)^{1/2}(Y_{(i)} - \bar{Y}_w).$$

For simplicity, we still use $X_{(i)}$ and $Y_{(i)}$ to denote the weighted and centered values and $(Y_{(i)},\ \delta_{(i)},\ X_{(i)})$ to denote the weighted data.

The objective function of SWLS (3) therefore becomes

$$L(\beta) = (Y_u - X_u\,\beta)^T\,(Y_u - X_u\,\beta). \tag{6}$$

So, it is easy to show that the SWLS in Eq. (1) is equivalent to the OLS estimator without intercept on the weighted data with K−M weights. Unfortunately, OLS estimation does not perform well with variable selection and is simply infeasible when $p > n$. Hence, the need to introduce various regularized methods improves OLS, such as lasso (Tibshirani 1996), the elastic net (Zou and Hastie 2005), and the Dantzig selector (Candes and Tao 2007). Many of these regularized methods are developed for data where $p > n$ and the coefficients vector is sparse.

2.3 Penalized SWLS with censoring constraints

The general frame of regularized WLS objective function is therefore defined by

$$L(\beta,\ \lambda) = (Y_u - X_u\,\beta)^T\,(Y_u - X_u\,\beta) + \lambda\,\mathrm{pen}(\beta), \tag{7}$$

where λ is the (scalar or vector) penalty parameter and the penalty quantity $\mathrm{pen}(\beta)$ is set typically in a way so that it controls the complexity of the model. For example, the penalty $\mathrm{pen}(\beta)$ for ridge, lasso, and elastic net are defined as

$$\sum_{j=1}^p \beta_j^2, \qquad \sum_{j=1}^p \mid \beta_j \mid, \qquad \left(\sum_{j=1}^p \mid \beta_j \mid,\ \sum_{j=1}^p \beta_j^2\right),$$

respectively. This type of regularized WLS with lasso penalty is studied recently by Huang et al. (2006). A regularized WLS method called CCLASSO which combines ridge and lasso penalty is used in Hu and Rao (2010). So, the objective function of the regularized WLS method with censoring constraints becomes

$$L(\beta,\ \lambda,\ \lambda_0) = (Y_u - X_u\,\beta)^T\,(Y_u - X_u\,\beta) + \lambda\,\mathrm{pen}(\beta) + \frac{\lambda_0}{n}\,\xi^T\xi,\ \text{subject to } Y_{\bar{u}} \leq X_{\bar{u}}\,\beta + \xi. \tag{8}$$

3 Proposed model framework

3.1 Adaptive elastic net (AEnet)

The elastic net (Zou and Hastie 2005) has proved useful when analyzing data with very many correlated covariates. The ℓ_1 part of the penalty for elastic net generates a sparse model. On the other hand, the quadratic part of the penalty removes the limitation on the number of selected variables when $p \gg n$. The quadratic part of the penalty also stabilizes the ℓ_1 regularization path and shrinks the coefficients of correlated predictors toward each other, allowing them to borrow strength from each other. The elastic net cannot be applied directly to the AFT models because of censoring, but the regularized

WLS (7) with the elastic net penalty overcomes this problem. The naive elastic net estimator $\hat{\beta}$ for censored data is obtained as

$$\underset{\beta}{\arg\min}\, (Y_u - X_u\,\beta)^T (Y_u - X_u\,\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2\, \beta^T \beta. \tag{9}$$

With some algebra, this naive elastic net can be transformed into a lasso-type problem in an augmented space as below

$$\underset{\beta}{\arg\min}\, \left(Y_u^* - X_u^*\,\beta\right)^T \left(Y_u^* - X_u^*\,\beta\right) + \lambda_1 \sum_{j=1}^{p} |\beta_j|,$$

where

$$X_u^* = \begin{pmatrix} X_u \\ \sqrt{\lambda_2}\, I \end{pmatrix} \quad \text{and} \quad Y_u^* = \begin{pmatrix} Y_u \\ 0 \end{pmatrix}.$$

The original elastic net estimator is now defined by

$$\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net}).$$

It is established in Zou (2006) that the lasso does not exhibit the oracle properties. These properties include that the method selects the correct subset of predictors with probability tending to one, and estimates the non-zero parameters as efficiently as would be possible if we knew which variables were uninformative ahead of time. A modification, the adaptive elastic net that does satisfy the oracle properties, was studied in Zou (2006), Ghosh (2007), and Zou and Zhang (2009). The adaptive elastic net is a convex combination of the adaptive lasso penalty and the ridge penalty. Here, we present an adaptive elastic net approach designed for censored data that is referred to as the AEnet approach throughout the paper. We introduce the adaptive elastic net penalty terms including coefficients to the regularized WLS objective function (7).

$$\underset{\beta}{\arg\min}\, (Y_u - X_u\,\beta)^T (Y_u - X_u\,\beta) + \lambda_1 \sum_{j=1}^{p} \hat{\kappa}_j\, |\beta_j| + \lambda_2\, \beta^T \beta, \tag{10}$$

where $\hat{\kappa} = 1/|\hat{\beta}_0|^\gamma$ is the adaptive weight based on the initial estimator $\hat{\beta}_0$ for some γ. For the rest of this paper, κ will denote weights obtained from an initial estimator. For the initial estimator $\hat{\beta}_0$, the OLS estimator as suggested in Ghosh (2007) or the elastic net estimator as suggested in Zou and Zhang (2009) can be used. For this study, we use $\gamma = 1$ and the elastic net estimator on the weighted data as given by Eq. (9) as the initial estimator $\hat{\beta}_0$.

The adaptive elastic net can be transformed into an adaptive lasso-type problem in an augmented space in a similar way as for the naive elastic net.

$$\hat{\beta}^*_{a-nenet} = \underset{\beta}{\arg\min}\, \left(Y_u^* - X_u^*\,\beta\right)^T \left(Y_u^* - X_u^*\,\beta\right) + \lambda_1 \sum_{j=1}^{p} \hat{\kappa}_j\, |\beta_j|, \tag{11}$$

where

$$X^*_{u\,(n_u+p)\times p} = \begin{pmatrix} X_u \\ \sqrt{\lambda_2}\, I \end{pmatrix} \quad \text{and} \quad Y^*_{u\,(n_u+p)\times p} = \begin{pmatrix} Y_u \\ 0 \end{pmatrix}.$$

So, for fixed λ_2, the adaptive elastic net is equivalent to an adaptive lasso in augmented space. The adaptive lasso estimates in (11) can be solved for fixed λ_2 by the LARS algorithm (Efron et al. 2004). Then the adaptive elastic net estimate can be obtained by rescaling the estimate found in Eq. (11).

$$\hat{\beta}^*_{a-enet} = (1+\lambda_2)\,\hat{\beta}^*_{a-nenet}.$$

3.1.1 AEnet algorithm

The algorithm for the proposed adaptive elastic net approach as shown below is referred to as the AEnet algorithm.
Input: Design matrix X_u^*, response Y_u^*, a fixed set for λ_2, and $\hat{\kappa}$.

1. Define $X^{**}_{j(u)} = X^*_{j(u)}/\hat{\kappa}_j, \;\; j = 1, \cdots, p.$
2. Solve the lasso problem for all λ_1 and a fixed λ_2,

$$\hat{\beta}^{**}_{a-nenet} = \underset{\beta}{\arg\min}\, \left(Y_u^* - X_u^{**}\,\beta\right)^T \left(Y_u^* - X_u^{**}\,\beta\right) + \lambda_1 \sum_{j=1}^{p} |\beta_j|.$$

3. Calculate $\hat{\beta}_{j\,a-enet}^{\;*} = (1+\lambda_2)\,\hat{\beta}_{j\,a-nenet}^{\;**}/\hat{\kappa}_j$.

To find the optimal value for the tuning parameters (λ_1, λ_2), λ_2 is typically assumed to take values in a relatively small grid, say (0, 0.5, 1.0, 1.5, 2.0,$\cdots$,5). For each λ_2, the LARS algorithm produces the entire solution path. This gives the optimal equivalent specification for lasso in terms of fraction of the ℓ_1 norm (t_1). Then the optimal pair of (t_1, λ_2) is obtained using k-fold cross-validation.

3.2 Adaptive elastic net with censoring constraints (AEnetCC)

Here, we present an extension of the above adaptive elastic approach that allows the censoring constraints to be implemented into the optimization framework. The adaptive elastic net estimator for censored data given by Eq. (10) can be rewritten with censoring constraints as

$$\tilde{\beta}^*_{a-enet} = \underset{\beta,\, \xi}{\arg\min}\, (Y_u - X_u\,\beta)^T (Y_u - X_u\,\beta) + \lambda_2\,\beta^T\beta + \frac{\lambda_0}{n}\,\xi^T\xi,$$
$$\text{subject to } \sum_{j=1}^{p} \hat{\kappa}_j\,|\beta_j| \le t_1 \text{ and } Y_{\bar{u}} \le X_{\bar{u}}\,\beta + \xi, \tag{12}$$

where t_1 is the lasso tuning parameter. We use a quadratic programming (QP) approach to solve the minimization problem of (12). The QP cannot handle $|\beta_j|$ because the lasso constraint ($\sum_{j=1}^{p} |\beta_j| \le t_1$) makes the QP solutions nonlinear in the Y_i. Further modification is needed to use $|\beta_j|$ in the QP framework. Following Tibshirani (1996), we use a modified design matrix $\tilde{X} = [X, -X]$ and represent coefficients β as the difference between two non-negative coefficients β^+ and β^-. Although the technique doubles the number of variables in the problem, it requires only a (known and bounded) linear number of constraints and only requires the solution to one QP problem. Now Eq. (12) becomes

$$\tilde{\beta}^*_{a-enet} = \underset{\beta^+,\, \beta^-,\, \xi}{\arg\min}\, [Y_u - X_u\,(\beta^+ - \beta^-)]^T [Y_u - X_u\,(\beta^+ - \beta^-)] + \lambda_2\,\beta^{+^T}\beta^+ + \lambda_2\,\beta^{-^T}\beta^- + \frac{\lambda_0}{n}\,\xi^T\xi,$$
$$\text{subject to } \sum_{j=1}^{p} \hat{\kappa}_j\,(\beta_j^+ + \beta_j^-) \le t_1$$
$$\text{and } Y_{\bar{u}} \le X_{\bar{u}}\,(\beta^+ - \beta^-) + \xi,\ \beta^+ \ge 0,\ \beta^- \ge 0. \tag{13}$$

According to Ghosh (2007), the estimator $\tilde{\beta}^*_{a-enet}$ is asymptotically normal.

3.2.1 AEnetCC algorithm

The algorithm for the proposed adaptive elastic net with censoring constraints approach is referred to as the AEnetCC algorithm.

Input: $\hat{\kappa}$.

1. Define $X^{**}_{j(u)} = X_{j(u)}/\hat{\kappa}_j,\ j = 1, \cdots, p$.
2. Solve the elastic net problem,

$$\tilde{\beta}^{**}_{a-nenet} = \underset{\beta^+,\, \beta^-,\, \xi}{\arg\min}\, [Y_u - X^{**}_u\,(\beta^+ - \beta^-)]^T [Y_u - X^{**}_u\,(\beta^+ - \beta^-)] + \lambda_2\,\beta^{+^T}\beta^+ + \lambda_2\,\beta^{-^T}\beta^- + \frac{\lambda_0}{n}\,\xi^T\xi,$$
$$\text{subject to } \sum_{j=1}^{p} \left(\beta_j^+ + \beta_j^-\right) \le t_1$$
$$\text{and } Y_{\bar{u}} \le X_{\bar{u}}\,(\beta^+ - \beta^-) + \xi,\ \beta^+ \ge 0,\ \beta^- \ge 0.$$

3. Calculate $\tilde{\beta}^{\,*}_{j\,a-enet} = (1 + \lambda_2)\,\tilde{\beta}^{\,**}_{j\,a-nenet}/\hat{\kappa}_j$.

The AEnetCC has three tuning parameters $(\lambda_0, \lambda_1, \lambda_2)$. For this method, we use the same optimal pair of (λ_1, λ_2) as found in AEnet. Then λ_0 is typically allowed to take values in a grid such as $(0, 0.5, 1, 1.5, \cdots, 10)$, and the optimal value for λ_0 obtained by 5-fold cross-validation. Here, the value of λ_0 typically depends upon how stringently one wants the model to satisfy the censoring constraints compared to how good is the prediction for uncensored data.

3.3 Weighted elastic net (WEnet)

In this section, we present a *weighted elastic net* for censored data. This is an extension of the adaptive elastic net where for suitable weight κ the ridge penalty term is expressed as $\sum_{j=1}^{p} (\kappa_j\,\beta_j)^2$, instead of $\sum_{j=1}^{p} \beta_j^2$. This is a doubly adaptive-type model. This type of regularized technique for uncensored data was first studied in Hong and Zhang (2010). They established the model consistency and its oracle property under some regularity conditions. Following the regularized WLS given in (7) and the weighted elastic net as proposed in Hong and Zhang (2010), the weighted elastic net for censored data can be defined by

$$\underset{\beta}{\arg\min}\, (Y_u - X_u\,\beta)^T (Y_u - X_u\,\beta) + n_u\,\lambda_1 \sum_{j=1}^{p} \kappa_j\,|\beta_j| + n_u\,\lambda_2 \sum_{j=1}^{p} (\kappa_j\,\beta_j)^2, \tag{14}$$

where $\kappa_j > 0,\ j = 1, \cdots, p$ are the weighted penalty coefficients and n_u are the number of uncensored observations. The weight is typically chosen as the standard deviations of the associated estimators (Hong and Zhang 2010). Since standard deviations are unknown in practice, we use the standard error of an initial consistent estimator. For estimating standard error under high-dimensional data, we use a bootstrap procedure based on the elastic net model on the weighted data (9). For data where $n > p$, as in Jin et al. (2003) and Jin et al. (2006), we choose the Gehan-type rank estimator as an initial estimator. This is defined (Gehan 1965) as the solution to the system of estimating equations, $0 = U_G(\beta)$, where

$$U_G(\beta) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \delta_i\,(X_i - X_{i'})\,I\{\xi_i(\beta) \le \xi_{i'}(\beta)\}, \tag{15}$$

and $\xi_i(\beta) = Y_i - X_i^T\,\beta$. Note that Eq. (15) can be expressed as the p-dimensional gradient of the convex loss function, $L_G(\beta)$, where

$$L_G(\beta) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \delta_i \{\xi_i(\beta) - \xi_{i'}(\beta)\}^-,$$

$$a^- = 1_{\{a<0\}}|a|.$$

Similar to the adaptive elastic net, the weighted elastic net can be transformed into a weighted lasso-type problem on an augmented data set. We rewrite Eq. (14) with a scaled coefficient difference as

$$\begin{aligned}\hat{\beta}_{w-enet} &= \underset{\beta}{\arg\min}\,(Y_u - X_u\,\beta)^T(Y_u - X_u\,\beta)\\&\quad + \lambda_1 \sum_{j=1}^{p} \kappa_j\,|\beta_j| + \times\lambda_2 \sum_{j=1}^{p} (\kappa_j\,\beta_j)^2 \qquad (16)\\&= \underset{\beta}{\arg\min}\left[\begin{pmatrix}Y_u\\0\end{pmatrix} - \begin{pmatrix}X_u\\\sqrt{\lambda_2}\,K\end{pmatrix}\right.\\&\quad \left.\times\, \frac{1}{\sqrt{1+\lambda_2}}\sqrt{1+\lambda_2}\,\beta\right]^T\\&\quad \times\left[\begin{pmatrix}Y_u\\0\end{pmatrix} - \begin{pmatrix}X_u\\\sqrt{\lambda_2}\,K\end{pmatrix}\frac{1}{\sqrt{1+\lambda_2}}\sqrt{1+\lambda_2}\,\beta\right]\\&\quad + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\,\kappa_j \times \sqrt{1+\lambda_2}\,|\beta_j|, \qquad (17)\end{aligned}$$

where $K = \mathrm{diag}[\kappa_1, \cdots, \kappa_p]$. Now assume that

$$\begin{aligned}X^*_{u\,(n_u+p)\times p} &= (1+\lambda_2)^{-\frac{1}{2}}\begin{pmatrix}X_u\\\sqrt{\lambda_2}\,K\end{pmatrix},\\Y^*_{u\,(n_u+p)\times p} &= \begin{pmatrix}Y_u\\0\end{pmatrix},\\\tilde{\lambda} &= \frac{\lambda_1}{\sqrt{1+\lambda_2}},\\\beta^* &= \sqrt{1+\lambda_2}\,\beta.\end{aligned}$$

Then the estimator in (17) with new notation becomes

$$\begin{aligned}\hat{\beta}^*_{w-enet} &= \underset{\beta}{\arg\min}\,(Y^*_u - X^*_u\,\beta^*)^T(Y^*_u - X^*_u\,\beta^*)\\&\quad + \tilde{\lambda}\sum_{j=1}^{p}\kappa_j\,|\beta^*_j|. \qquad (18)\end{aligned}$$

So, for fixed λ_2, the weighted elastic net can be transformed into an adaptive lasso problem (18) in some augmented space. The naive weighted elastic net estimator $\hat{\beta}_{w-nenet}$ can be obtained by

$$\hat{\beta}_{j\,w-nenet} = \hat{\beta}^{\,*}_{j\,w-enet}/\sqrt{1+\lambda_2}. \qquad (19)$$

The original weighted elastic net estimator $\hat{\beta}_{j\,w-enet}$ can therefore be obtained by rescaling the estimate found in (19)

$$(1+\lambda_2)\,\hat{\beta}_{j\,w-nenet} = \sqrt{1+\lambda_2}\,\hat{\beta}^{\,*}_{j\,w-enet}. \qquad (20)$$

3.3.1 WEnet algorithm

The algorithm for the proposed weighted elastic net approach is referred to as the WEnet algorithm.

Input: Design matrix X^*_u, response Y^*_u, $\tilde{\lambda}$, and $\hat{\kappa}_j$ for $j = 1, \cdots, p$.

1. Define $X^{**}_{j(u)} = X^*_{j(u)}/\hat{\kappa}_j,\ \ j = 1, \cdots, p$.
2. Solve the lasso problem for all λ_1 and a fixed λ_2,

$$\begin{aligned}\hat{\beta}^{**}_{w-nenet} &= \underset{\beta}{\arg\min}\,\left(Y^*_u - X^{**}_u\,\beta\right)^T\left(Y^*_u - X^{**}_u\,\beta\right)\\&\quad + \tilde{\lambda}\sum_{j=1}^{p}|\beta_j|.\end{aligned}$$

3. Calculate $\hat{\beta}^*_{w-enet} = \hat{\beta}^{**}_{w-nenet}/\hat{\kappa}_j$.

Output: $\hat{\beta}_{w-enet} = \sqrt{(1+\lambda_2)}\,\hat{\beta}^*_{w-enet}$.

To find the optimal value for the tuning parameters $(\lambda_1,\ \lambda_2)$, λ_2 is typically assumed to take values in a relatively small grid, similar to the grid used for AEnet algorithm. To optimize the tuning parameters $(\lambda_1,\ \lambda_2)$ for WEnet, we follow exactly the same procedure as described in the previous section for the AEnet.

3.4 Weighted elastic net with censoring constraints (WEnetCC)

Following the adaptive elastic net with censoring constraints, the weighted elastic net model with censoring constraints can be defined by

$$\begin{aligned}\tilde{\beta}^*_{w-enet} &= \underset{\beta,\ \xi}{\arg\min}\,(Y_u - X_u\,\beta)^T(Y_u - X_u\,\beta)\\&\quad + \lambda_2\,\beta^T\,K^T\,K\,\beta + \frac{\lambda_0}{n}\,\xi^T\xi,\\&\quad \text{subject to } \sum_{j=1}^{p}\hat{\kappa}_j\,|\beta_j| \le t_1 \text{ and } Y_{\bar{u}} \le X_{\bar{u}}\,\beta + \xi.\end{aligned}$$

Now after representing β^* as the difference between two non-negative coefficients β^{*+} and β^{*-}, Eq. (22) becomes

$$\begin{aligned}\tilde{\beta}^{**}_{w-enet} &= \underset{\beta^{*+},\ \beta^{*-},\ \xi}{\arg\min}\,\left(Y^*_u - X^*_u\left(\beta^{*+} - \beta^{*-}\right)\right)^T\\&\quad \times\left(Y^*_u - X^*_u\left(\beta^{*+} - \beta^{*-}\right)\right) + \frac{\lambda_0}{n}\,\xi^T\xi,\\&\quad \text{subject to } \sum_{j=1}^{p}\kappa_j\,(\beta^{*+} - \beta^{*-}) \le t_1 \text{ and}\\&\quad Y_{\bar{u}} \le X_{\bar{u}}\,(\beta^{*+} - \beta^{*-}) + \xi,\ \ \beta^{*+} \ge 0,\ \ \beta^{*-} \ge 0. \qquad (21)\end{aligned}$$

3.4.1 WEnetCC algorithm

Below we present the algorithm for the proposed weighted elastic net with censoring constraints approach which is referred to as the WEnetCC algorithm.

Input: Design matrix X_u^*, response Y_u^*, $\tilde{\lambda}$, and $\hat{\kappa}_j$ for $j = 1, \cdots, p$.

1. Define $X_{j(u)}^{**} = X_{j(u)}^*/\hat{\kappa}_j, \;\; j = 1, \cdots, p$.
2. Solve the lasso problem,

$$\tilde{\beta}^{**}_{w-enet} = \underset{\beta^{*+},\, \beta^{*-},\, \xi}{\arg\min} \left(Y_u^* - X_u^{**}\left(\beta^{*+} - \beta^{*-}\right)\right)^T \left(Y_u^* - X_u^{**} \times \left(\beta^{*+} - \beta^{*-}\right)\right) + \frac{\lambda_0}{n}\xi^T\xi,$$
$$\text{subject to } \sum_{j=1}^{p}(\beta^{*+} - \beta^{*-}) \le t_1 \text{ and}$$
$$Y_{\bar{u}} \le X_{\bar{u}}\,(\beta^{*+} - \beta^{*-}) + \xi,\;\; \beta^{*+} \ge 0,\;\; \beta^{*-} \ge 0.$$

3. Calculate $\tilde{\beta}^*_{w-enet} = \hat{\beta}^{**}_{w-enet}/\hat{\kappa}_j$.

Output: $\tilde{\beta}_{w-enet} = \sqrt{(1+\lambda_2)}\,\tilde{\beta}^*_{w-enet}$.

The estimator in the second step is obtained by optimizing the QP problem. To obtain optimal tuning parameter $(\lambda_0, \lambda_1, \lambda_2)$ of the WEnetCC, we follow exactly the same procedure as for the AEnetCC algorithm. According to *Theorem 2* in Zou (2006), and *Theorem 3.2* in Ghosh (2007), both estimators $\hat{\beta}_{j\,w-enet}$ and $\tilde{\beta}_{j\,w-enet}$ are asymptotically normal.

Remark 1 Under some regularity conditions, the WLS estimator with K−M weights is consistent and asymptotically normal Stute (1993, 1996). The proof is not directly applicable to the lasso penalty since the lasso penalty is not differentiable. In Huang et al. (2006), it is shown that the regularized WLS estimator with lasso penalty has the asymptotic normality property. In their proof, they added two more conditions additional to the regularity conditions mentioned in Stute (1993, 1996). The two additional conditions are (i) the regularized WLS lasso estimator has finite variance, and (ii) the bias of the K−M integrals is of the order $o(n^{1/2})$, which is related to the level of censoring and to the tail behavior of the K−M estimator.

3.5 Variable selection criteria

For AEnet and WEnet, we use LARS that produces exact zero coefficients in solution paths and hence does parsimonious variable selection (Khan 2013; Wu 2012; Ghosh 2011; Wang et al. 2008). For the remaining two methods AEnetCC and WEnetCC, we use an approach that simultaneously removes several variables that have very small coefficients.

We use an AIC$_c$-type score based on the weighted k-fold cross-validation error CV−S (which is the sum of squared residuals of uncensored data multiplied by the K−M weights i.e., $(Y_u - X_u\,\hat{\beta})^T\, w_u\,(Y_u - X_u\,\hat{\beta})$). The AIC$_c$ score is defined by

$$\text{AIC}_c \text{ score} = n_u \log(\text{CV} - \text{S}) + 2k\left(\frac{n_u}{n_u - k - 1}\right), \tag{22}$$

where k is the number of estimable parameters in the model.

3.5.1 Variable selection algorithm for AEnetCC and WEnetCC

The above AIC$_c$ score is used as variable selection criteria.

1. Get the optimal pair (λ_1, λ_2) from fitting AEnet or WEnet.
2. Fix a set of λ_0. Then for each λ_0,
   (a) Fit AFT model by the computational procedure (3.2.1) or (3.4.1). Find the predictor set PS using $|\hat{\beta}| > \varsigma$.
   (b) Use the PS and divide the dataset into M parts. Leave out one part at a time and fit AFT model by the computational procedure (3.2.1) or (3.4.1).
   (c) Combine the M-fitted models built in step 2(b) by averaging their coefficients. Compute the CV–S and then AIC$_c$ score.
3. Repeat step 2 until all λ_0 are exhausted. Return the model with the lowest AIC$_c$ score and corresponding λ_0.

We choose a very small value for precision parameter, say $\varsigma = 1e^{-5}$, as a default value but any other suitable value can be chosen. Alternatively, ς should be considered as an additional tuning parameter in the above variable selection algorithm. It is also possible either to adapt existing efficient optimization algorithm such as SCAD (Fan and Li 2001) or LARS-EN [a modified LARS developed for adaptive elastic net by Ghosh (2007)] or to develop a new algorithm that avoids the ς parameter completely.

3.6 Measures of fit and measures of prediction

The following MSE is computed to measure the fit in the training data:

$$MSE_{TR} = \frac{1}{n_u}\sum_{i=1}^{n}\delta_i\,(\hat{Y}_i - Y_i)^2, \tag{23}$$

where n_u is the number of uncensored observations. This measure compares the fitted values with the true values corresponding to the uncensored observations. We first generate a training dataset, such as (Y_i, δ_i) in (23), and then a test dataset $(Y_{i,\,\text{new}}, \delta_{i,\,\text{new}})$ (say) of the same size using the same design parameters. All the methods are fitted using the training data. Then in order to get predicted values $\hat{Y}_{\text{new}}$, the fitted model is used with the **X** matrix of the test data. We measure the prediction accuracy by

$$MSE_{TE} = \frac{1}{n_u} \sum_{i=1}^{n} \delta_{i,\text{ new}} (\hat{Y}_{i,\text{ new}} - Y_{i,\text{ new}})^2. \tag{24}$$

We estimate the variance of the regression parameters using the nonparametric 0.632 bootstrap (Efron and Tibshirani 1993) in which one samples $\tilde{n} \approx 0.632 \times n$ from the n observations without replacement. We use $\tilde{n} \approx 0.632 \times n$ as the expected number of distinct bootstrap observations is about $0.632 \times n$. We repeat the 0.632 bootstrap for a reasonable number of times (say, B). We use $B = 500$ for each simulation example, and then the sample variance of the bootstrap estimates provides an estimate of the variance of $\hat{\beta}$.

4 Simulation studies

The purpose of this section is to evaluate and compare the performance of the proposed approaches using simulation studies and a real data example. We use six existing model selection approaches for comparison purpose. However, the aim is not to address which approach is superior, rather we identify the similarities among the methods and accordingly provide some suggestions about the situations where one approach may outperform the others. The six approaches are used for the last simulation example and also for the real data example. For AEnet and WEnet methods, we use {0, 0.6, 1.1, 1.7, 2.2, 2.8, 3.3, 3.9, 4.4, 5.0} as the grid for λ_2. For the two censoring constraint-based methods AEnetCC and WEnetCC, the set {0, 1.0, 1.4, 1.8, 2.2, 2.6, 3.0} is used for the penalties of violations of constraints λ_0.

Several alternative penalized regression and Bayesian approaches for variable selection for high-dimensional censored data have been developed. We use the simple elastic net (Enet) approach as defined in Eq. (9) on the weighted data. Another approach is the adaptive elastic net for AFT (ENet-AFT) in Engler and Li (2009). There is also an MCMC selection-based Bayesian method (Bayesian-AFT) for log-normal AFT model introduced in Sha et al. (2006); we use this approach only for the log-normal AFT model. We also use three correlation-based greedy variable selection approaches: sure independence screening (SIS) (Fan and Lv 2008), tilted correlation screening (TCS) (Cho and Fryzlewicz 2012), and PC-simple (Bühlmann et al. 2010) all implemented with the weighted data under the SWLS as defined by the Eq. (6).

4.1 Simulation studies

The logarithm of the survival time is generated from the true AFT model

$$Y_i = \alpha + X_i^T \beta + \sigma \varepsilon_i, \;\; i = 1, \cdots, n \tag{25}$$

with $\varepsilon_i \sim f(\cdot)$, any suitable probability density function and σ, the signal-to-noise ratio. We use correlated datasets. We first create the correlation matrix R where the pairwise correlation (say, r_{ij}) between the i-th and j-th components of X is set to be $0.5^{|i-j|}$. Then multiplying the random covariate matrix X with the Cholesky decomposition of the correlation matrix R gives covariates with the desired correlation structure. Censoring time is generated from particular distributions maintaining a desired censoring level, $P_{\%}$. We consider three $P_{\%}$, 30, 50, and 70, that are indicated as low, medium, and high, respectively. We maintain random censoring except for the case that if the largest observation is found censored ($i.e.$, $Y_{(n)}^{+}$), then we reclassify it as uncensored according to Efron's tail correction (Efron 1967). This is necessary since the WLS method involves the K−M weights that are based on the K−M distribution function (Khan and Shaw 2013b).

4.1.1 Low-dimensional example: $n = 100$, $p = 40$

We consider 40 covariates with three blocks: β coefficients for $j \in \{1, \cdots, 5\}$ are set to be 5 and for $j \in \{6, \cdots, 10\}$ are set to be 2. We treat these two blocks as informative blocks i.e., contain potential covariates (say, $p_\gamma = 10$). The remaining β coefficients (i.e., $j \in \{11, \cdots, 40\}$) are set to be zero and we set $\mathbf{X} \sim U(0, 1)$. We consider log-normal and exponential AFT models. The survival time is generated using (25) with $\varepsilon_i \sim N(0, 1)$ for the log-normal AFT model and using (25) with $\varepsilon_i \sim \log(\text{exponential})$ for the exponential AFT model. More precisely, in Eq. (25), the error is $\sigma\varepsilon$ where $\sigma = 1$ and

$$\varepsilon = \frac{\log[Ex] - E\log[Ex]}{\sqrt{\text{Var}\{\log[Ex]\}}}, \quad \text{where } Ex \sim \text{exponential}\,(5).$$

For both AFT models, the censoring time is generated using the log-normal distribution $\exp[N(c_0\sqrt{1+\sigma}, \;(1+\sigma^2))]$. Here, c_0 is calculated analytically to produce the chosen $P_{\%}$. We fit all four methods, and for both AFT models 100 runs are simulated. For each covariate, we record the frequency of being selected among 100 simulation runs, the minimum, mean, and maximum. The summary statistics for each block are shown in Table 1.

As expected, the covariates of the informatics blocks (blocks 1 and 2) should be selected. For both AFT models, Table 1 shows that when the data are uncorrelated all the methods tend to select most of the informative covariates from block 1 and with a very high percentage of informative covariates from block 2. The inclusion rate of the noninformative covariates into the final models of the methods is shown to be very low particularly with low censoring. When the covariates are correlated, the two methods WEnet and WEnetCC outperform the other two methods. The methods AEnet and AEnetCC tend to select informative covariates with low selection frequencies and tend to include

Table 1 Variable selection frequency percentages for the methods for both log-normal and exponential AFT models

$P_{\%}$	Methods	Parameters	$r_{ij} = 0$		$r_{ij} = 0.5$	
			Log-normal (Min, Mean, Max)	Exponential (Min, Mean, Max)	Log-normal (Min, Mean, Max)	Exponential (Min, Mean, Max)
30	AEnet	Block1	(99, 99, 99)	(95, 95, 95)	(62, 69.6, 79)	(62, 66.6, 74)
		Block2	(66, 74.6, 83)	(73, 94.8, 97)	(39, 43.8, 51)	(35, 41.8, 48)
		Block3[a]	(00, 1.6, 05)	(00, 2.8, 07)	(11, 18.5, 25)	(12, 18.9, 24)
	AEnetCC	Block1	(100, 100, 100)	(100, 100, 100)	(76, 80, 82)	(80, 87, 95)
		Block2	(48, 53.6, 57)	(48, 51.8, 54)	(44, 46.8, 50)	(51, 54.4, 59)
		Block3[a]	(00, 3.1, 06)	(02, 6.3, 12)	(17, 25.8, 33)	(22, 30, 38)
	WEnet	Block1	(99, 99, 99)	(99, 99, 99)	(100, 100, 100)	(100, 100, 100)
		Block2	(63, 73.4, 82)	(74, 77.2, 79)	(51, 62.8, 75)	(55, 62.2, 68)
		Block3[a]	(00, 1.3, 3)	(00, 1.9, 06)	(00, 2.3, 06)	(00, 03, 07)
	WEnetCC	Block1	(100, 100, 100)	(99, 99.8, 100)	(93, 95.4, 97)	(96, 97, 98)
		Block2	(30, 36, 39)	(24, 32.4, 36)	(44, 47.6, 52)	(43, 49, 57)
		Block3[a]	(00, 2.2, 5)	(00, 2.5, 06)	(04, 9.6, 15)	(03, 9.2, 14)
50	AEnet	Block1	(99, 99.6, 100)	(98, 98.8, 99)	(77, 81.8, 85)	(77, 81.6, 84)
		Block2	(63, 69.8, 75)	(61, 66.6, 76)	(48, 52.2, 55)	(48, 53.6, 59)
		Block3[a]	(09, 18.2, 25)	(10, 19, 27)	(22, 31.5, 38)	(26, 31.7, 41)
	AEnetCC	Block1	(97, 98.6, 100)	(98, 99.2, 100)	(53, 56.4, 59)	(52, 57.8, 65)
		Block2	(27, 33.6, 39)	(34, 38.2, 41)	(19, 22.6, 28)	(16, 24, 29)
		Block3[a]	(00, 3.5, 08)	(01, 4.4, 10)	(05, 10.5, 18)	(04, 8.7, 13)
	WEnet	Block1	(99, 99.8, 100)	(100, 100, 100)	(98, 99, 100)	(99, 99.6, 100)
		Block2	(63, 68.6, 74)	(62, 65.8, 71)	(55, 59.4, 62)	(46, 58.2, 68)
		Block3[a]	(05, 9.9, 18)	(05, 9.1, 14)	(01, 5.3, 10)	(02, 4.9, 09)
	WEnetCC	Block1	(91, 95.8, 100)	(91, 95.4, 98)	(71, 76.2, 81)	(71, 74.2, 81)
		Block2	(15, 21.6, 26)	(22, 23.4, 24)	(20, 25.4, 29)	(20, 26, 29)
		Block3[a]	(00, 2.8, 7)	(00, 2.8, 07)	(02, 6.6, 12)	(02, 6.8, 12)
70	AEnet	Block1	(96, 97.4, 99)	(93, 94.8, 97)	(71, 77, 82)	(70, 73.8, 77)
		Block2	(52, 59.2, 62)	(56, 58.2, 62)	(52, 55, 58)	(42, 48.6, 55)
		Block3[a]	(15, 23.5, 31)	(17, 24.1, 34)	(27, 34.9, 43)	(21, 28.7, 36)
	AEnetCC	Block1	(85, 89.8, 94)	(85, 88.6, 92)	(35, 37.2, 41)	(36, 41.6, 48)
		Block2	(19, 27, 33)	(24, 29.4, 36)	(16, 17, 18)	(13, 19.8, 26)
		Block3[a]	(03, 10, 15)	(05, 9.5, 14)	(05, 8.4, 12)	(05, 10, 16)
	WEnet	Block1	(96, 97.2, 99)	(92, 96, 99)	(90, 93.8, 96)	(93, 94.6, 97)
		Block2	(41, 49.4, 57)	(42, 50.2, 55)	(36, 44.8, 52)	(43, 49.2, 57)
		Block3[a]	(07, 16.9, 22)	(13, 17.8, 24)	(08, 12.3, 18)	(05, 10.6, 18)
	WEnetCC	Block1	(82, 85.8, 89)	(83, 86, 89)	(50, 57.8, 63)	(47, 57.6, 65)
		Block2	(14, 23.4, 31)	(21, 25.2, 31)	(15, 19.6, 24)	(19, 22.2, 28)
		Block3[a]	(03, 8.8, 15)	(03, 6.3, 12)	(03, 7.3, 11)	(02, 7.4, 15)

[a] Stands for the noninformative block with zero coefficients

more irrelevant predictors in their models. However, as censoring increases, the overall performance in terms of both selecting informative covariates and excluding noninformative covariates slightly decreases. But the rate of decrease in performance is shown to be higher in selecting informative covariates than in excluding noninformative covariates.

The prediction performance of the methods on the datasets under the log-normal AFT model is given in Fig. 1. We similarly obtained graphs for the exponential AFT model and not reported here since we got almost similar results as for the log-normal model. According to those figures, all the methods fit the test data reasonably well for all levels of censoring. For the AEnet on the correlated dataset with 30 % censoring, the predicted times are considerably biased compared the corresponding observed times (Fig. 1, row 1, column 3). However, this does not happen with 50 % censoring (row 2,

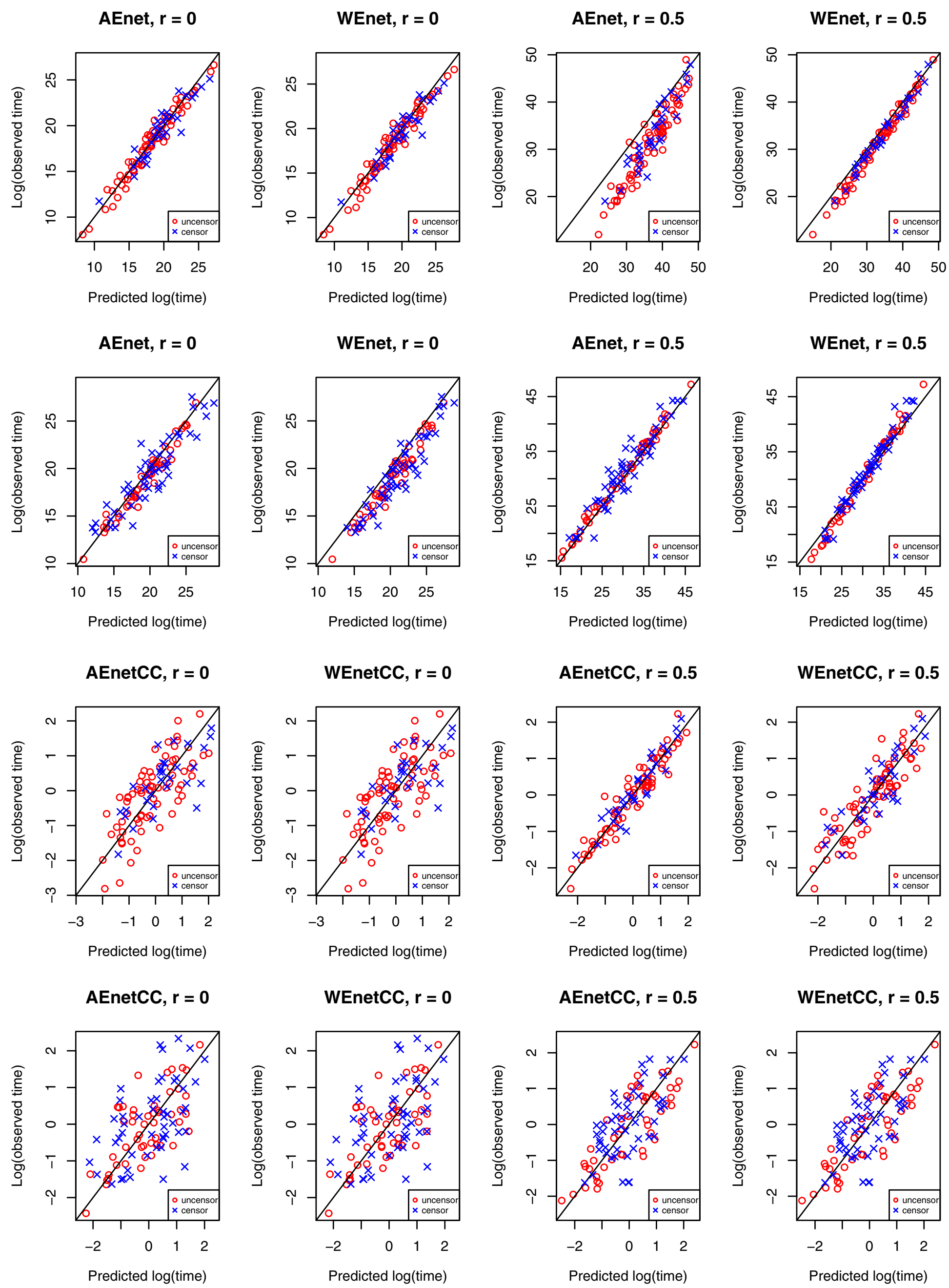

Fig. 1 Predicted versus observed log survival time under log-normal AFT model for the methods AEnet and WEnet for datasets with $P_{\%} = 30$ (*first row panel*) and $P_{\%} = 50$ (*second row panel*) and for the methods AEnetCC and WEnetCC for datasets with $P_{\%} = 30$ (*third row panel*) and $P_{\%} = 50$ (*fourth row panel*)

column 3). So it is hard to draw general conclusion. Given this performance of AEnet, all other methods select the correct set of non-zero coefficients, although the coefficients may be poorly estimated.

4.1.2 High-dimensional example: $n = 100$, $p = 120$

We fix $\alpha = 1$ and set the first 20 coefficients for β's to 4 (i.e., p_γ is 20) and the remaining coefficients of β to zero. We keep everything else similar to low-dimensional example. We fit all four proposed methods together with six other existing methods and compare them in terms of variable selection.

The Bayesian-AFT MCMC sampler was run for 200,000 iterations at which the first 100,000 iterations are used as burn-in. A starting model with 40 randomly selected variables is considered. For the Bayesian-AFT, we choose 0.01 to be the cut-off for the marginal posterior probabilities.

Table 2 shows the results from 100 simulation runs. We evaluate the frequency of being selected among 100 simulation runs and then compute the minimum, mean, and maximum of those frequencies. The results are presented in the table for two censoring levels, 30 and 50 %. In terms of the mean selection frequencies of informative variables, all proposed methods outperform the Enet-AFT, SIS, TCS, and

Table 2 Variable selection frequency percentages for 20 p_γ variables and 100 $p - p_\gamma$ (non-relevant) variables with the methods for both log-normal and exponential AFT models

$P_{\%}$	Methods	Parameters	$r_{ij} = 0$		$r_{ij} = 0.5$	
			Log-normal (Min, Mean, Max)	Exponential (Min, Mean, Max)	Log-normal (Min, Mean, Max)	Exponential (Min, Mean, Max)
30	AEnet	p_γ	(77, 84.8, 94)	(82, 87.9, 94)	(42, 50.7, 58)	(44, 54.2, 63)
		$p - p_\gamma$	(02, 9.4, 19)	(02, 9.8, 17)	(11, 21.1, 29)	(17, 23.8, 31)
	AEnetCC	p_γ	(77, 89.8, 98)	(83, 89.5, 94)	(49, 62, 70)	(60, 67, 77)
		$p - p_\gamma$	(08, 14.3, 21)	(04, 13.3, 21)	(30, 40.3, 52)	(35, 44.67, 53)
	WEnet	p_γ	(72, 80, 90)	(73, 80, 89)	(36, 43.3, 50)	(36, 43.2, 55)
		$p - p_\gamma$	(05, 12.3, 21)	(05, 13.1, 22)	(00, 1.4, 05)	(00, 1.5, 05)
	WEnetCC	p_γ	(82, 87.3, 94)	(82, 86.5, 91)	(52, 61.8, 73)	(55, 65, 75)
		$p - p_\gamma$	(06, 12.4, 19)	(03, 11.6, 20)	(03, 11, 22)	(03, 10.5, 21)
	Enet	p_γ	(96, 97.2, 98)	(97, 97.3, 98)	(96, 98.2, 100)	(96, 98.8, 100)
		$p - p_\gamma$	(05, 6.5, 8)	(4, 6.5, 9)	(01, 4.3, 10)	(00, 4.1, 10)
	Enet-AFT	p_γ	(50, 53.2, 56)	(50, 53.3, 55)	(36, 40.3, 48)	(32, 39.5, 50)
		$p - p_\gamma$	(03, 4.3, 06)	(03, 4.4, 06)	(05, 16.7, 27)	(08, 16.7, 27)
	Bayesian-AFT	p_γ	(45, 64, 75)	–	(45, 72, 85)	–
		$p - p_\gamma$	(14, 29, 41)	–	(36, 48.3, 62)	–
	SIS	p_γ	(11, 20.5, 29)	(13, 21.3, 29)	(06, 11.9, 18)	(06, 11.5, 18)
		$p - p_\gamma$	(00, 0.9, 03)	(00, 0.8, 04)	(00, 2.6, 06)	(00, 2.7, 07)
	TCS	p_γ	(38, 45.9, 53)	(44, 47.4, 52)	(42, 52.8, 61)	(44, 55.1, 64)
		$p - p_\gamma$	(00, 3.4, 08)	(00, 3.7, 09)	(05, 15.1, 26)	(08, 15.0, 27)
	PC-simple	p_γ	(11, 23.5, 30)	(16, 25.2, 33)	(11, 18.1, 26)	(13, 19.7, 29)
		$p - p_\gamma$	(00, 1.1, 04)	(00, 0.9, 05)	(00, 5.6, 13)	(01, 5.4, 15)
50	AEnet	p_γ	(59, 68.1, 75)	(60, 66.2, 73)	(41, 50.7, 56)	(43, 50.6, 60)
		$p - p_\gamma$	(03, 8.8, 16)	(03, 9.7, 18)	(16, 26, 35)	(15, 25.6, 38)
	AEnetCC	p_γ	(69, 76.4, 85)	(68, 72.7, 80)	(47, 57.2, 67)	(41, 52.5, 64)
		$p - p_\gamma$	(16, 23.6, 31)	(12, 22.2, 33)	(26, 37.9, 50)	(25, 35.2, 50)
	WEnet	p_γ	(45, 56.8, 65)	(46, 52.2, 60)	(21, 28.4, 36)	(21, 28.6, 37)
		$p - p_\gamma$	(04, 10.6, 18)	(03, 11.5, 20)	(00, 1.8, 06)	(00, 02, 06)
	WEnetCC	p_γ	(69, 75.4, 87)	(68, 72.8, 80)	(49, 55.5, 66)	(46, 55.8, 64)
		$p - p_\gamma$	(12, 21.1, 29)	(11, 21.2, 30)	(06, 12.4, 19)	(03, 12.3, 21)
	Enet	p_γ	(64, 67.4, 70)	(66, 68.1, 70)	(70, 76.2, 85)	(71, 80.4, 85)
		$p - p_\gamma$	(11, 12.7, 15)	(10, 12.7, 15)	(03, 10.3, 18)	(03, 8.8, 16)
	Enet-AFT	p_γ	(33, 35.3, 38)	(33, 35.8, 38)	(25, 17.8, 41)	(26, 33.6, 41)
		$p - p_\gamma$	(03, 04, 05)	(03, 04, 06)	(09, 17.8, 27)	(08, 17.9, 26)
	Bayesian-AFT	p_γ	(40, 53.5, 70)	–	(30, 51.5, 60)	–
		$p - p_\gamma$	(27, 50, 67)	–	(32, 45.1, 56)	–
	SIS	p_γ	(11, 17.4, 25)	(09, 16.8, 23)	(05, 10.5, 15)	(06, 09.9, 17)
		$p - p_\gamma$	(00, 1.5, 06)	(00, 1.6, 05)	(00, 2.9, 08)	(00, 3.0, 09)
	TCS	p_γ	(48, 56.9, 64)	(46, 55.6, 64)	(40, 47.5, 55)	(38, 45.5, 59)
		$p - p_\gamma$	(25, 37.6, 48)	(27, 37.9, 47)	(28, 39.5, 53)	(28, 39.9, 51)
	PC-simple	p_γ	(13, 19.4, 27)	(12, 18.8, 26)	(07, 14.8, 21)	(10, 15.9, 23)
		$p - p_\gamma$	(00, 1.9, 07)	(00, 1.9, 05)	(00, 5.3, 16)	(00, 5.1, 13)

PC-simple methods. Their performances are very close to the performance of the Enet method at the higher censoring level, although they show slightly poorer performances with lower censoring. However, for the uncorrelated dataset and both AFT models, the two methods AEnet and AEnetCC tend to exclude fewer noninformative covariates. The three

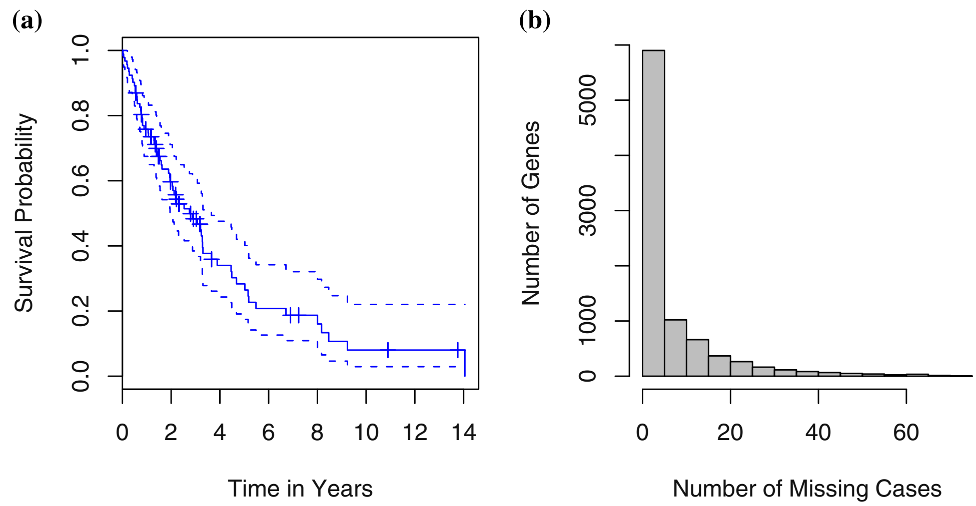

Fig. 2 **a** K–M plots of overall survival of patients for MCL data (*left*). **b** Histogram of number of genes with missingness (*right*)

greedy approaches tend to select far fewer variables (both informative and spurious) in the final model. As expected, for all approaches, the number of variables selected from p_γ decreases as the censoring increases.

5 Real data example

5.1 Mantle cell lymphoma data

Rosenwald et al. (2003) reported a study using microarray expression analysis of mantle cell lymphoma (MCL). The primary goal of the study was to discover gene expression signatures that correlate with survival in MCL patients. MCL accounts for 6 % of all non-Hodgkins lymphomas and a higher fraction of deaths from lymphoma, given that it is an incurable malignancy (Swerdlow and Williams 2002). Among 101 untreated patients with no history of previous lymphoma included in the study, 92 were classified as having MCL, based on established morphologic and immunophenotypic criteria. Survival times of 64 patients were available, and the remaining 28 patients were censored (i.e., censoring rate $P_{\%}$ is 30). The median survival time was 2.8 years (range 0.02–14.05 years). The length of survival of MCL patients following diagnosis is quite heterogeneous (Fig. 2a). Many patients died within the first 2 years following diagnosis, yet 15 % (14/92) of the patients survived more than 5 years and three patients survived more than 10 years. Lymphochip DNA microarrays were used to quantify mRNA expression in the lymphoma samples from the 92 patients. The gene expression dataset that contains expression values of 8810 cDNA elements is available at http://llmpp.nih.gov/MCL/. The data do not provide any further relevant covariates for MCL patients.

We apply the AFT model with all methods to this dataset. Although these methods have in principle no limit to the number of genes that can be used, we pre-process the data in a simple way. Pre-processing is important to gain further stability by reducing potential noise from the dataset. The pre-processing steps can be summarized as follows: (1) First, missing values of the original dataset are imputed by their sample means (for example, for a particular gene the missing gene expression value for a patient is replaced by the mean of the gene expression values for the observed patients). (2) Secondly, we compute correlation coefficients of the uncensored survival times with gene expressions. (3) Finally, a reasonable number of genes are selected based on their correlation with the response. After pre-processing, 574 genes with the largest absolute correlation (>0.3) have been identified and selected for analysis. We then standardize these 574 gene expressions to have zero mean and unit variance and take logarithms of the observed times. We use mean imputation to impute missing values for the MCL dataset.

We employ all the approaches and select the optimal tuning parameter with 5-fold cross-validation. The results are reported in Table 3. The results suggest that most of the methods select a considerable number of genes and there are many common genes that are found between the methods. The three greedy methods tend to select fewer genes. The TCS selects the lowest number of genes (2), while the Enet selects the largest number of genes (68). Among the four proposed methods, the AEnetCC selects the lowest number of genes (18). The Enet, Enet-AFT, Bayesian-AFT, SIS, TCS, and PC-simple select 3, 5, 4, 5, 1, and 1 genes, respectively.

In the final model, out of 574 genes, the Bayesian-AFT method finds only 25 genes that have the largest marginal posterior probabilities (we choose 0.38 to be the cut-off) (see also Fig. 3). The left panel of Fig. 3 shows that MCMC chains mostly visited models with 20–40 genes. The right panel of Fig. 3 shows that not many genes have high marginal probabilities (only 25 genes with marginal probabilities greater than 0.38).

Table 3 Number of genes selected by the methods (diagonal elements) and number of common genes found between the methods (off diagonal elements)

Methods	AEnet	AEnetCC	WEnet	WEnetCC	Enet	Enet-AFT	Bayesian-AFT	SIS	TCS	PC-simple
AEnet	45	12	10	03	10	07	05	05	01	02
AEnetCC	12	18	02	01	02	05	04	05	01	01
WEnet	10	02	39	09	03	01	01	00	00	00
WEnetCC	03	01	09	40	02	01	02	01	00	00
Enet	10	02	03	02	68	08	03	01	00	01
Enet-AFT	07	05	01	01	08	25	05	03	01	00
Bayesian-AFT	05	04	01	02	03	05	25	02	01	00
SIS	05	05	00	01	01	03	02	05	01	01
TCS	01	01	00	00	00	01	01	01	02	00
PC-simple	02	01	00	00	01	00	00	01	00	03

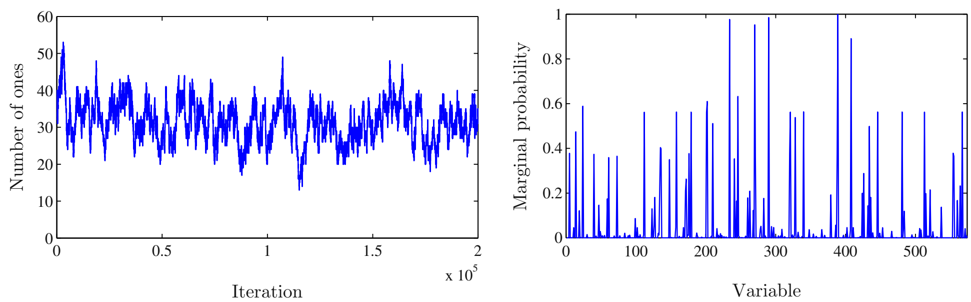

Fig. 3 Number of included genes (*left*) in each iteration and marginal posterior probabilities of inclusion (*right*)

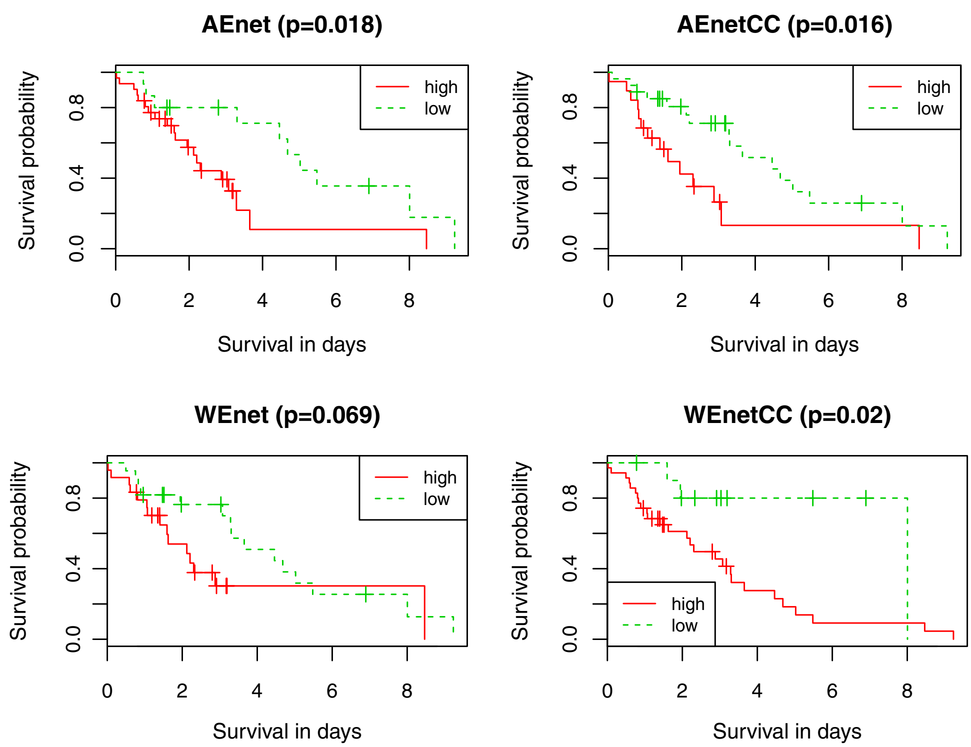

Fig. 4 Survival comparison between the high-risk group and low-risk group using different methods

There are four genes with uniqid 24761, 27434, 27844, and 29888 that are identified by all four proposed methods, and there are another five genes with uniqid 22136, 24383, 29876, 30034, and 33704 that are identified by three of the proposed methods. The overall analysis of the MCL data suggests that all four proposed methods are capable of identifying sets of genes that are potentially related to the response variable. In the analysis, the AEnetCC, as in the simulations with $r_{ij} = 0.5$, selects a smaller number of genes than do the other methods. However, with gene expression data, a smaller number of identified genes mean a more focused hypothesis for future confirmations studies, and is thus usually preferred.

We evaluate the predictive performance using the four proposed methods. We use the obtained models to predict the risk of death in the MCL test dataset. We first partition the data randomly into two equal parts called training and test datasets. We then implement the methods to the training dataset and compute the risk scores ($X^T\hat{\beta}$) based on the model estimates and the test dataset. The subjects are classified to be in the high-risk group or low-risk group based on whether the risk score exceeds the median survival time in the training dataset. We compare the K–M curves between the two groups, and then a log-rank test is used to identify the difference between the two K–M curves (see Fig. 4). The corresponding predictive MSE for the methods AEnet, AEnetCC, WEnet, and WEnetCC are 15.7, 19.6, 29.4, and 15.0, respectively. The log-rank test suggests that the high- and low-risk groups are significantly different from each other under almost all the methods. So it seems that the methods can group very well the subjects' survival time into two risk sets.

5.1.1 Mantle cell lymphoma data under adaptive pre-processing

A challenge with the MCL dataset (with ultra-high dimensionality, $p \gg n$) is that the important genes might be highly correlated with some unimportant ones, which usually increases with dimensionality. The maximum spurious correlation between a gene and the survival time also grows with dimensionality. Here, we focus on a smart pre-processing technique for MCL data that addresses this issue and also reduces circularity bias (Kriegeskorte et al. 2009) by reducing false discovery rate. Fan and Lv (2008) introduced the SIS idea that reduces the ultra-high dimensionality to a relatively large-scale d_n, where $d_n < n$. In Fan and Lv (2008), asymptotic theory is proved to show that, with high probability, SIS keeps all important variables with vanishing false discovery rate. Then, the lower dimension methods such as SCAD (Fan and Li 2001) can be used to estimate the sparse model. This procedure is referred to as SIS+SCAD. For MCL data, we first apply SIS to reduce the dimensionality from 8810 to $d_n = [3\,n^{2/3}] = 61$, and then fit the data using all ten *methods* (our four proposed methods and six competitors, including the three greedy methods). We call this procedure $SIS + methods$. The results are reported in Table 4.

From Table 4, we see that the proposed methods select a considerable number of genes in common with the other methods. The three greedy methods tend to return final mod-

Table 4 Number of genes selected by the SIS followed by the methods (diagonal elements) and number of common genes found between the methods (off diagonal elements)

Methods	SIS+AEnet	SIS+AEnetCC	SIS+WEnet	SIS+WEnetCC	SIS+Enet	SIS+Enet-AFT	SIS+Bys-AFT	SIS+SCAD	SIS+TCS	SIS+PC
SIS+AEnet	05	02	02	01	02	04	02	04	00	02
SIS+AEnetCC	02	12	05	03	07	07	02	03	00	02
SIS+WEnet	02	05	10	02	05	07	02	02	00	02
SIS+WEnetCC	01	03	02	08	05	02	01	01	00	01
SIS+Enet	02	07	05	05	32	13	02	03	01	02
SIS+Enet-AFT	04	07	07	02	13	30	02	04	00	02
SIS+Bys-AFT	02	02	02	01	02	02	02	02	00	02
SIS+SCAD	04	03	02	01	03	04	02	05	00	02
SIS+TCS	00	00	00	00	01	00	00	00	01	00
SIS+PC	02	02	02	01	02	02	02	02	00	02

Table 5 Number of genes selected by the methods with and without implementing the SIS

Methods	AEnet	AEnetCC	WEnet	WEnetCC	Enet	Enet-AFT	Bayesian-AFT	SIS	TCS	PC-simple
SIS+AEnet	1	1	0	0	0	0	0	1	0	0
SIS+AEnetCC	1	1	0	0	1	0	0	1	0	1
SIS+WEnet	1	1	0	0	0	1	0	1	0	0
SIS+WEnetCC	1	1	0	0	0	0	0	0	0	0
SIS+Enet	6	6	0	1	1	2	0	3	1	1
SIS+Enet-AFT	4	4	0	1	1	2	0	4	0	1
SIS+Bys-AFT	0	0	0	0	0	0	0	0	0	0
SIS+SCAD	1	1	0	0	1	0	0	1	0	1
SIS+TCS	0	0	0	0	0	0	0	0	0	0
SIS+PC	0	0	0	0	0	0	0	0	0	0

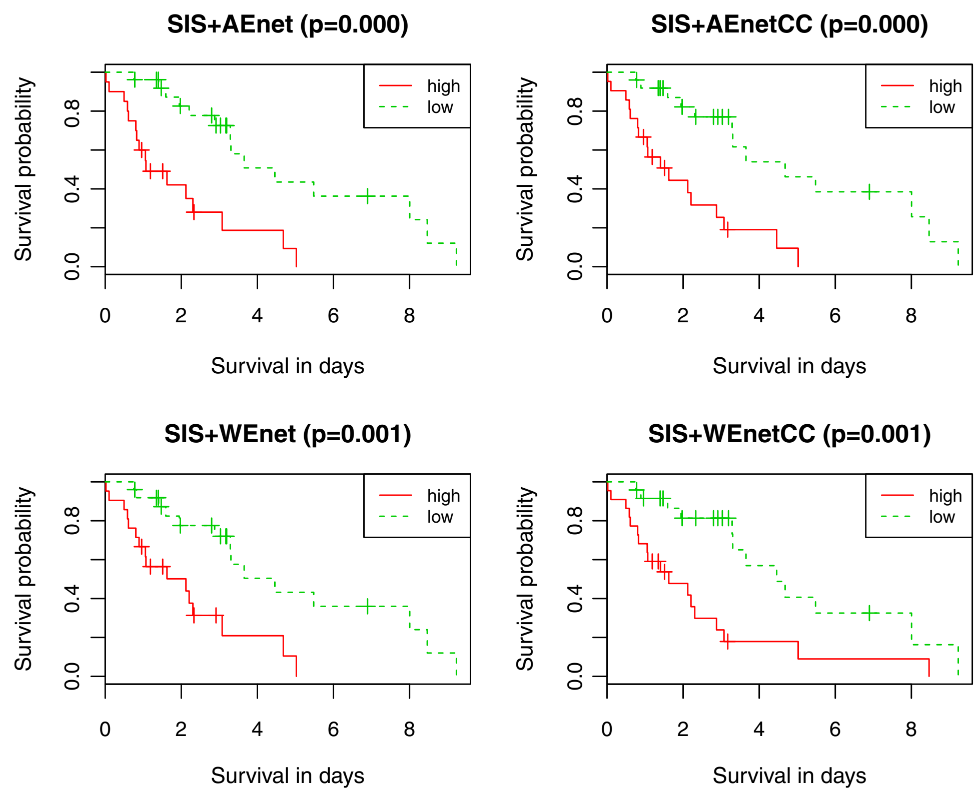

Fig. 5 Survival comparison between the high-risk group and low-risk group using different methods with SIS implementation

els with fewer genes. TCS selects the lowest number of genes (1), while Enet selects the largest number of genes (32). Among the four proposed methods, AEnet selects the lowest number of genes (5). The results also suggest that the implementation of SIS followed by all the methods (proposed and competitors) pick smaller sets of genes, most of which are not in the set of genes found by the methods without SIS. Table 5 shows the number of common genes between the methods with and without SIS implementation. The predictive performance for the four proposed methods with SIS implementation has been evaluated (see Fig. 5) similar to what was done before for methods without SIS (Fig. 4). The predictive MSE for methods SIS+AEnet, SIS+AEnetCC, SIS+WEnet, and SIS+WEnetCC are 1.2, 1.1, 2.8, and 1.4, respectively. It is clear from the predictive performance graph Fig. 5 (also Fig. 4 for methods without SIS) and the predictive MSE's that the predictive performance improves considerably after implementation of SIS.

6 Discussion

In this study, we propose adaptive elastic net and weighted elastic net regularized variable selection approaches for the AFT model. We conjecture that the proposed approaches enjoy oracle properties under some regularity assumptions in analogous with the adaptive elastic net (Ghosh 2007) and weighted elastic net (Hong and Zhang 2010). They produce sparse solutions. We propose another two variable selection algorithms, where censored observations are used as extra constraints in the optimization function of the methods. The censoring constraints in the optimization equations limit the model space using the right censored data. It is shown how all the methods apart from the AEnetCC can be optimized after transforming them into an adaptive lasso problem in some augmented space.

The analysis of simulated datasets with both dimensional—low and high—and MCL gene expression data shows that the regularized SWLS approach for variable selection with its four implementations (AEnet, AEnetCC, WEnet, and WEnetCC) can be used for selecting important variables. They also can be used for future prediction for survival time under AFT models. The MCL gene expression data analysis also suggests that the sure independence screening improves the performance of all the proposed methods in the AFT model. It is observed that the methods AEnetCC and WEnetCC seem only to perform well under moderately high-dimensional censored datasets such as with variables at most four or five times higher than the sample size. However, the two methods with SIS implementation have no such limits. A variable selection strategy such as ours, which allows the adaptive elastic net (Ghosh 2007) and weighted elastic net (Hong and Zhang 2010) to be used for censored data, is new. The extensions of these methods, which use right censored data to limit the model space and improve the parameter estimation, are also new. Both the adaptive and weighted elastic net together with two extensions enjoy the computational advantages of the lasso. The methods show that various regularized techniques will continue to be an important tool in variable selection with survival data. Our new variable selection techniques for both low and high-dimensional censored data are promising alternatives to the existing methods. For implementing all proposed methods, we have provided a publicly available package `AdapEnetClass` (Khan and Shaw 2014) implemented in the R programming system.

Acknowledgments The first author is grateful to the centre for research in Statistical Methodology (CRiSM), Department of Statistics, University of Warwick, UK for offering research funding for his PhD study.

References

Akaike, H.: Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)

Antoniadis, A., Fryzlewicz, P., Letue, F.: The Dantzig selector in Cox's proportional hazards model. Scand. J. Stat. **37**(4), 531–552 (2010)

Buckley, J., James, I.: Linear regression with censored data. Biometrika **66**, 429–436 (1979)

Bühlmann, P., Kalisch, M., Maathuis, M.H.: Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. Biometrika **97**(2), 261–278 (2010)

Cai, T., Huang, J., Tian, L.: Regularized estimation for the accelerated failure time model. Biometrics **65**, 394–404 (2009)

Candes, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)

Cho, H., Fryzlewicz, P.: High dimensional variable selection via tilting. J. R. Stat. Soc. Ser. B **74**(3), 593–622 (2012)

Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. Ser. B **34**, 187–220 (1972)

Datta, S., Le-Rademacher, J., Datta, S.: Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics **63**, 259–271 (2007)

Efron, B.: The two sample problem with censored data. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, pp. 831–853. Prentice Hall, New York (1967)

Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)

Engler, D., Li, Y.: Survival analysis with high-dimensional covariates: an application in microarray studies. Stat. Appl. Genet. Mol. Biol. **8**(1), 1–22 (2009). (Article 14)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

Fan, J., Li, R.: Variable selection for Cox's proportional hazards model and frailty model. Ann. Stat. **30**, 74–99 (2002)

Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B **70**(5), 849–911 (2008)

Faraggi, D., Simon, R.: Bayesian variable selection method for censored survival data. Biometrics **54**, 1475–1485 (1998)

Frank, I.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**(2), 109–135 (1993)

Gehan, E.A.: A generalized Wilcoxon test for comparing arbitrarily singlecensored samples. Biometrika **52**, 203–223 (1965)

Ghosh, S.: On the grouped selection and model complexity of the adaptive elastic net. Stat. Comput. **21**(3), 451–462 (2011)

Ghosh, S.: Adaptive elastic net: an improvement of elastic net to achieve oracle properties. Technical Reports, Indiana University-Purdue University, Indianapolis, (PR no. 07–01) (2007)

Gui, J., Li, H.: Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics **21**, 3001–3008 (2005)

Hong, D., Zhang, F.: Weighted elastic net model for mass spectrometry imaging processing. Math. Model. Nat. Phenom. **5**(3), 115–133 (2010)

Hu, S., Rao, J.S.: Sparse penalization with censoring constraints for estimating high dimensional AFT models with applications to microarray data analysis. Technical Reports, University of Miami (2010)

Huang, J., Harrington, D.: Iterative partial least squares with rightcensored data analysis: a comparison to other dimension reduction techniques. Biometrics **61**(1), 17–24 (2005)

Huang, J., Ma, S.: Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal. **16**, 176–195 (2010)

Huang, J., Ma, S., Xie, H.: Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics **62**, 813–820 (2006)

Hunter, D.R., Li, R.: Variable selection using MM algorithms. Ann. Stat. **33**(4), 1617–1642 (2005)

Jin, Z., Lin, D., Wei, L.J., Ying, Z.L.: Rank-based inference for the accelerated failure time model. Biometrika **90**, 341–353 (2003)

Jin, Z., Lin, D.Y., Ying, Z.: On least-squares regression with censored data. Biometrika **93**(1), 147–161 (2006)

Khan, M.H.R., Shaw, J.E.H.: AdapEnetClass: a class of adaptive elastic net methods for censored data. R package version 1.1 (2014)

Khan, M.H.R.: Variable selection and estimation procedures for high-dimensional survival data. Ph.D. Thesis, Department of Statistics, University of Warwick (2013)

Khan, M.H.R., Shaw, J.E.H.: On dealing with censored largest observations under weighted least squares. CRiSM Working Paper, No 13–07 Department of Statistics, University of Warwick (2013b)

Khan, M.H.R., Shaw, J.E.H.: Variable selection with the modified Buckley- James method and the dantzig selector for high-dimensional survival data. In: 59th ISI World Statistics Congress Proceedings, Hong Kong, pp. 4239–4244, 25–30 Aug 2013c

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I.: Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. **12**(5), 535–540 (2009)

Li, H., Luan, Y.: Kernel Cox regression models for linking gene expression profiles to censored survival data. Pac. Symp. Biocomput. **8**, 65–76 (2003)

Meinshausen, N., Bühlmann, P.: Stability selection. J. R. Stat. Soc. Ser. B **72**(4), 417–473 (2010)

Peduzzi, P.N., Hardy, R.J., Holford, T.R.: A stepwise variable selection procedure for nonlinear regression models. Biometrics **36**, 511–516 (1980)

Radchenko, P., James, G.M.: Improved variable selection with Forward-Lasso adaptive shrinkage. Ann. Appl. Stat. **5**(1), 427–448 (2011)

Rosenwald, A., Wright, G., Wiestner, A., Chan, W., Connors, J., Campo, E., Gascoyne, R., Grogan, T., Muller Hermelink, H., Smeland, E., Chiorazzi, M., Giltnane, J., Hurt, E., Zhao, H., Averett, L., Henrickson, S., Yang, L., Powell, J., Wilson, W., Jaffe, E., Simon, R., Klausner, R., Montserrat, E., Bosch, F., Greiner, T., Weisenburger, D., Sanger, W., Dave, B., Lynch, J., Vose, J., Armitage, J., Fisher, R., Miller, T., LeBlanc, M., Ott, G., Kvaloy, S., Holte, H., Delabie, J., Staudt, L.: The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell **3**, 185–197 (2003)

Sha, N., Tadesse, M.G., Vannucci, M.: Bayesian variable selection for the analysis of microarray data with censored outcome. Bioinformatics **22**(18), 2262–2268 (2006)

Stute, W.: Consistent estimation under random censorship when covariables are available. J. Multivar. Anal. **45**, 89–103 (1993)

Stute, W.: Distributional convergence under random censorship when covariables are present. Scand. J. Stat. **23**, 461–471 (1996)

Swerdlow, S., Williams, M.: From centrocytic to mantle cell lymphoma: a clinicopathologic and molecular review of 3 decades. Hum. Pathol. **33**, 7–20 (2002)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)

Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**, 385–395 (1997)

Wang, S., Nan, B., Zhu, J., Beer, D.G.: Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics **64**, 132–140 (2008)

Wu, Y.: Elastic net for Cox's proportional hazards model with a solution path algorithm. Stat. Sin. **22**, 271–294 (2012)

Ying, Z.: A large sample study of rank estimation for censored regression data. Ann. Stat. **21**(1), 76–99 (1993)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B **68**, 49–67 (2006)

Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)

Zou, H., Zhang, H.H.: On the adaptive elastic-net with a diverging number of parameters. Ann. Stat. **37**(4), 1733–1751 (2009)