Communications for Statistical Applications and Methods
2018, Vol. 25, No. 6, 591–604

https://doi.org/10.29220/CSAM.2018.25.6.591
Print ISSN 2287-7843 / Online ISSN 2383-4757

Penalized variable selection for accelerated failure time models

Eunyoung Park[a], Il Do Ha[1,a]

[a]Department of Statistics, Pukyong National University, Korea

Abstract

The accelerated failure time (AFT) model is a linear model under the log-transformation of survival time that has been introduced as a useful alternative to the proportional hazards (PH) model. In this paper we propose variable-selection procedures of fixed effects in a parametric AFT model using penalized likelihood approaches. We use three popular penalty functions, least absolute shrinkage and selection operator (LASSO), adaptive LASSO and smoothly clipped absolute deviation (SCAD). With these procedures we can select important variables and estimate the fixed effects at the same time. The performance of the proposed method is evaluated using simulation studies, including the investigation of impact of misspecifying the assumed distribution. The proposed method is illustrated with a primary biliary cirrhosis (PBC) data set.

Keywords: AFT model, LASSO, penalized likelihood, SCAD, variable selection

1. Introduction

In survival analysis, accelerated failure time (AFT) model has been introduced as a useful alternative to proportional hazards (PH) model (Lawless, 1982). The PH model is modelled by fixed effects (e.g., regression coefficients) acting multiplicatively on the hazard rate of individual survival time. However, in the AFT model the fixed effects act linearly on the individual survival time, thus making the interpretation of the fixed effects easier than in the PH model. AFT model is robust against the misspecification of the assumed model due to its log-linear transformation (Hutton and Monaghan, 2002; Ha *et al.*, 2002). In this paper, we are interested in the development of a variable-selection procedure in the AFT model. Recently, variable-selection methods using a penalized likelihood with penalty functions have been widely studied in various statistical models, such as linear models, generalized linear models (GLMs), and Cox's (1972) PH models (Tibshirani, 1996; Fan and Li, 2001). The advantages of these methods are the ability to select important variables and estimates the regression coefficients of the covariates, simultaneously. Selecting relevant variables from a regression model with a number of covariates is important in data analysis including survival analysis.

Various penalized variable-selection methods in the semiparametric AFT model with an unspecified distribution have been studied (Huang *et al.*, 2006; Cai *et al.*, 2009; Huang and Ma, 2010; Xu *et al.*, 2010; Wang and Song, 2011; Zhang *et al.*, 2018). Parametric survival models and their functional forms (e.g., survival function) are simple and they would be useful in survival analysis if the model assumption is correct or less sensitive against the inference. The fixed effects (i.e., regression

This paper is condensed form of the first author's Master Thesis from the Pukyong National University, Busan, Korea.

[1] Corresponding author: Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-gu, Busan 48513, Korea. E-mail: idha1353@pknu.ac.kr

Published 30 November 2018 / journal homepage: http://csam.or.kr

coefficients) in parametric AFT model with a specified distribution (e.g., lognormal or Weibull) are relatively robust against the misspecification of the assumed distribution as compared to nuisance parameters in random error terms (Hutton and Monaghan, 2002; Ha *et al.*, 2002). Thus, we are interested in studying the behaviors of variable selection of fixed effects under parametric AFT model.

In this paper, we develop variable-selection procedures of fixed effects in parametric AFT model using a penalized likelihood approach. Here we consider two useful parametric distributions, lognormal and Weibull distributions, for survival analysis. For the variable selection, we use three popular penalty functions, least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), adaptive LASSO (ALASSO) (Zou, 2006), and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). We also show how to derive the penalized likelihood procedure. The performance of the proposed method is evaluated using simulation studies. In particular, the simulation shows that the proposed variable-selection method is somewhat robust against the misspecification of the assumed model. The proposed method is illustrated with a primary biliary cirrhosis (PBC) (Tibshirani, 1997) data set which is well known in the literature.

This paper is organized as follows. In Section 2, we briefly review the AFT model, and propose a penalized variable-selection method using AFT model, including the derivations of the estimation procedures. In Section 3, the results of simulation studies are presented to evaluate the validity of the proposed method. The proposed method is illustrated with the PBC data in Section 4. Discussion is given in Section 5. Finally, technical details are given in the Appendix.

2. Variable selection for accelerated failure time models

2.1. Accelerated failure time model

Let T_i be the survival time (failure time) for each subject ($i = 1, \ldots, n$) and let C_i be the corresponding random censoring time. AFT model is to describe a linear relationship between the logarithm of survival time and covariates as:

$$\log T_i = x_i^T \beta + \epsilon_i, \tag{2.1}$$

where $x_i = (1, x_{i1}, \ldots, x_{i,p-1})^T$ is a covariates vector of the i^{th} subject, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ is a $p \times 1$ vector of regression coefficients corresponding to x_i, and ϵ_i is a random error.

For the distribution of ϵ_i, we consider two popular parametric distributions, i.e., normal and extreme value (EV) distributions. If $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ having the density

$$f(\epsilon_i) = \left(2\pi\sigma_\epsilon^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\epsilon_i^2}{2\sigma_\epsilon^2}\right), \tag{2.2}$$

T_i has the lognormal (LN) distribution with location parameter $x_i^T\beta$ and scale parameter $\psi = \sigma_\epsilon^2$. If ϵ_i follows an EV distribution with scale parameter σ having the density

$$f(\epsilon_i) = \sigma^{-1} \exp\left\{\left(\frac{\epsilon_i}{\sigma}\right) - \exp\left(\frac{\epsilon_i}{\sigma}\right)\right\}, \tag{2.3}$$

T_i follows Weibull distribution with scale parameter $\lambda_0 = \exp\{-(x_i^T\beta)\psi\}$ and shape parameter $\psi = 1/\sigma$. In particular, the Weibull distribution is a flexible model because of an unique distribution satisfying both AFT and PH models (Lawless, 1982).

In this paper, if T_i in AFT model (2.1) follows LN distribution, we call the model LN AFT model. If T_i follows Weibull distribution, we call it Weibull AFT model. We follow two usual assumptions under non-informative censoring (Ha *et al.*, 2002; Zhou, 2005; Zhang *et al.*, 2018):

Assumption 1. *Given covariates x_i, T_i's and C_i's are conditionally independent and the pairs (T_i, C_i)'s are also conditionally independent for $i = 1, \ldots, n$.*

Assumption 2. *Given covariates x_i, C_i's are conditionally non-informative about T_i's.*

Based on these two assumptions, we make inferences as shown below.

2.2. Variable selection procedure

Now, we present how to derive a variable selection procedure using a penalized likelihood. In survival analysis with random censoring, observable random variables are given by

$$Y_i = \min(\log T_i, \log C_i) \quad \text{and} \quad \delta_i = I(T_i \leq C_i).$$

Let $\lambda(t)$ be the hazard function of T_i, and let $\Lambda(t) = \int_0^t \lambda(k)dk$ be the corresponding cumulative hazard function. Under Assumptions 1 and 2, the log-likelihood for AFT model (2.1) is defined by

$$\ell = \ell(\theta) = \sum_{i=1}^{n} \{\delta_i \log \lambda_\theta(y_i) - \Lambda_\theta(y_i)\}, \tag{2.4}$$

where $\theta = (\beta, \psi)^T$ and ψ is a parameter in random error term of ϵ_i. Note here that $\psi = 1/\sigma$ in EV of ϵ_i and $\psi = \sigma_\epsilon^2$ in normal.

For variable selection of fixed effects β in model (2.1), we use the following penalized log-likelihood (Fan and Li, 2001), denoted by ℓ_p, given by

$$\ell_p = \ell_p(\theta) = \ell(\theta) - n \sum_{k=0}^{p-1} J_\gamma(|\beta_k|), \tag{2.5}$$

where $J_\gamma(\cdot)$ is a penalty function with a tuning parameter γ. A larger value of γ tends to choose a simple model, whereas a smaller value of γ inclines to a complex model. Here, we use the three penalty functions, LASSO, ALASSO, and SCAD. The forms of three penalty functions are:

(1) LASSO (Tibshirani, 1996):

$$J_\gamma(|\beta|) = \gamma|\beta|. \tag{2.6}$$

(2) ALASSO (Zou, 2006):

$$J_\gamma(|\beta|) = \gamma|\beta|w, \tag{2.7}$$

where w is a known weights vector.

(3) SCAD (Fan and Li, 2001):

$$J'_\gamma(|\beta|) = \gamma I(|\beta| \leq \gamma) + \frac{(a\gamma - |\beta|)_+}{a-1} I(|\beta| > \gamma), \tag{2.8}$$

where $a = 3.7$ and x_+ denotes the positive part of x.

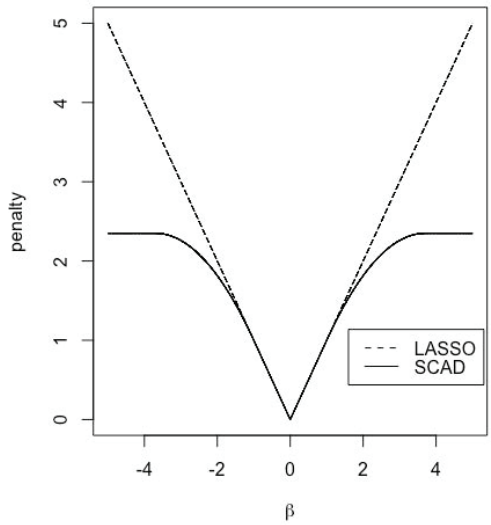

Figure 1: *Penalty functions of LASSO and SCAD. LASSO = least absolute shrinkage and selection operator; SCAD = smoothly clipped absolute deviation.*

Figure 1 displays the shapes of LASSO and SCAD functions under $\gamma = 1$. A good penalty function should produce estimates that satisfy unbiasedness, sparsity, and continuity (Fan and Li, 2001, 2002). The LASSO is a well-known penalty, but it does not satisfy these three properties. Thus, Fan and Li (2001, 2002) and Zou (2006) have shown that SCAD and ALASSO satisfy the three properties and that they can perform well as the oracle procedure in terms of selecting the correct subset models and estimating the true non-zero coefficients, simultaneously.

For the variable selection, we want to find the estimators $\hat{\beta}$ which maximize the penalized log-likelihood ℓ_p in (2.5), given by

$$\hat{\beta} = \arg\max_{\beta} \ell_p.$$

We call the resulting estimators penalized maximum likelihood estimators (PMLEs). The PMLEs are obtained by solving the following estimating equations:

$$\frac{\partial \ell_p}{\partial \beta_k} = \frac{\partial \ell}{\partial \beta_k} - n \sum_{k=0}^{p-1} [J_\gamma(|\beta_k|)]' = 0, \quad (k = 0, 1, \ldots, p-1). \tag{2.9}$$

Here we use $[J_\gamma(|\beta_k|)]' = J'_\gamma(|\beta_k|)\mathrm{sgn}(|\beta_k|) \approx \{J'_\gamma(|\beta_k^{(0)}|)/|\beta_k^{(0)}|\}\beta_k$ for $\beta_k \approx \beta_k^{(0)}$ by local quadratic approximation (LQA) (Fan and Li, 2001), and $\mathrm{sgn}(\cdot)$ is the sign function. It can be shown that the negative Hessian matrix of β based on ℓ_p can be explicitly written as a simple matrix form:

$$H_p = H_p(\ell_p; \beta) = -\frac{\partial^2 \ell_p}{\partial \beta \partial \beta^T} = X^T W X + n \sum_\gamma, \tag{2.10}$$

where X is a $n \times p$ model matrix of covariates x_i's and W is a weight matrix with a diagonal element w_i, i.e.,

$$W = -\frac{\partial^2 \ell}{\partial \eta \partial \eta^T} = \mathrm{diag}(w_i), \tag{2.11}$$

with a linear predictor $\eta = X\beta$ in AFT model (2.1) and $\sum_\gamma = \text{diag}\{J'_\gamma(|\beta_j|)/|\beta_j|\}$. Let $U(\cdot) = \varphi(\cdot)/\{1 - \Phi(\cdot)\}$ be the hazard function of $N(0,1)$, where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of $N(0,1)$, respectively. In LN AFT model, $w_i = \{\delta_i + (1-\delta_i)\xi(m_i)\}/\sigma_\epsilon^2$, where $\xi(m_i) = U(m_i)\{U(m_i) - m_i\}$, $U(m_i) = \varphi(m_i)/\{1 - \Phi(m_i)\}$, and $m_i = (y_i - x_i^T\beta)/\sigma_\epsilon$. In Weibull AFT model, $w_i = \Lambda_i/\sigma^2$, where $\Lambda_i = \exp(m_i)$ and $m_i = (y_i - x_i^T\beta)/\sigma$. We can obtain the PMLEs of β from the Newton-Raphson method; its one-step formula is given by

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + \left[-\ell''_p\left(\hat{\beta}^{(0)}\right)\right]^{-1} \ell'_p\left(\hat{\beta}^{(0)}\right), \tag{2.12}$$

where $\hat{\beta}^{(0)}$ is the initial values of β, $\ell'_p(\beta) = \partial\ell_p(\beta)/\partial\beta$, and $-\ell''_p(\beta) = H_p(\beta)$. The nuisance parameter ψ in the error term of model (2.1) is obtained from the following estimating equation:

$$\frac{\partial \ell_p}{\partial \psi} = \frac{\partial \ell}{\partial \psi} = 0 \tag{2.13}$$

since ψ does not depend on the penalty function. More details for the estimating equations are given in Appendix. Then we compute the sandwich standard error (Fan and Li, 2001; Ha *et al.*, 2014) for $\hat{\beta}$, from variance-covariance matrix

$$\text{cov}\left(\hat{\beta}\right) = \left(H_{\beta\beta} + n\sum_\gamma\right)^{-1} H_{\beta\beta}\left(H_{\beta\beta} + n\sum_\gamma\right)^{-1},$$

where $H_{\beta\beta} = -\partial^2\ell/\partial\beta\partial\beta^T = X^T W X$.

Wang *et al.* (2007) showed that the generalized cross validation (GCV) approach cannot select the tuning parameters satisfactorily, with a nonignorable overfitting effect in the resulting model. For the selection of tuning parameter γ, we use a Bayesian information criterion type (BIC-type) criterion (Ha *et al.*, 2014), given by

$$\text{BIC}^*(\gamma) = -2\ell\left(\hat{\beta}, \hat{\psi}\right) + \log(n)\text{df}, \tag{2.14}$$

where $\text{df} = \text{tr}[(H_{\beta\beta} + n\sum_\gamma)^{-1} H_{\beta\beta}]$ is an effective degree of freedom.

In summary, an outline of the proposed variable-selection algorithm is described as follows.

Step 1. Find initial values of β and ψ.

Step 2. In the inner loop, we maximize ℓ_p in (2.5) for β and ψ.

Step 3. In the outer loop, we find γ that minimizes $\text{BIC}^*(\gamma)$ in (2.14).

After convergence, we compute the estimated standard errors for $\hat{\beta}$. For the initial values of β in LASSO and ALASSO, we use the estimates from the AFT model without penalty. For the weights w of the ALASSO in (2.7), following Zhang and Lu (2007) and Wang and Song (2011), we use

$$w = \frac{1}{|\tilde{\beta}|},$$

where $\tilde{\beta}$ are non-penalized coefficient estimates. Following Ha *et al.* (2014, 2017), for the initial values of β in SCAD, we use the LASSO solutions. Our procedures were implemented by using R programs.

Table 1: Simulation results under LN AFT model ($\sigma_\epsilon^2 = 1$)

n	Method	C	IC	PT	MSE
100	LASSO	2.62	0.00	0.02	0.132
	ALASSO	4.18	0.00	0.41	0.079
	SCAD	4.37	0.01	0.59	0.077
300	LASSO	2.42	0.00	0.00	0.052
	ALASSO	4.39	0.00	0.45	0.021
	SCAD	4.46	0.00	0.61	0.017
500	LASSO	2.68	0.00	0.03	0.032
	ALASSO	4.50	0.00	0.59	0.015
	SCAD	4.71	0.00	0.78	0.014

LN = lognormal; AFT = accelerated failure time; MSE = mean squared error; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

3. Simulation study

Simulation studies, based upon 100 replications of simulated data, are presented to evaluate the performance of the proposed variable-selection procedure for AFT models. Here, we compare the performances of the variable-selection methods using LASSO, ALASSO, and SCAD. Below we consider the two distributions (LN, Weibull) for this purpose. Following the simulation scheme of Fan and Li (2001), we generate the data from the AFT model (2.1) with the true regression parameters

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)^T = (1, 0.8, 0, 0, 1, 0, 0, 0.6, 0)^T.$$

Here, the corresponding covariates $x = (1, x^*)$ and covariates $x^* = (x_1, \ldots, x_8)^T$ are generated with and AR(1) structure with a correlation coefficient $\rho = 0.5$. Note that x_1, x_4, and x_7 are important covariates. We also consider three sample sizes $n = 100, 300$, and 500. The corresponding censoring times C_i's are generated from an uniform distribution with a parameter empirically determined to achieve approximately the right censoring rate about 45%. As the measures of variable selection, we consider the average number of zero coefficients (C and IC), the probability of choosing the true model (PT) and mean squared error (MSE). Following Zhang and Lu (2007) and Wang and Song (2011), we summarize the median of MSEs over 100 replications to measure prediction accuracy; it is defined by $\text{MSE}(\hat{\beta}) = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)$, where Σ is the population covariance matrix of the covariates. Here, the "C"(5 is the best) indicates the average number of regression coefficients, of the five true zeros, correctly found to zero, and "IC" (0 is the best) indicates the average number of the four true non-zeros incorrectly set to zero.

For the LN case we consider $\sigma_\epsilon^2 = 1$, and for the Weibull $\sigma = 0.5$ (i.e., $\psi = 1/\sigma = 2$; increasing hazard), $\sigma = 1$ (i.e., $\psi = 1/\sigma = 1$; exponential distribution with constant hazard) and $\sigma = 2$ (i.e., $\psi = 1/\sigma = 0.5$; decreasing hazard). Table 1 (LN case) and Tables 2–4 (Weibull case) summarize the simulation results. Tables 1–4 indicate that the ALASSO and SCAD overall perform well as compared to the LASSO. The ALASSO and SCAD methods are further improved with n, while the LASSO method is not. In particular, the SCAD method outperforms the LASSO and ALASSO in terms of "C", "PT", and "MSE".

(1) LN case: Table 1.

(2) Weibull case: Table 2–4.

We also investigated the robustness of the proposed method when the true distribution of ϵ in the AFT model (2.1) is misspecified. Following Xu *et al.* (2010), we considered two misspecified

Table 2: Simulation results under Weibull AFT model ($\sigma = 0.5$)

n	Method	C	IC	PT	MSE
100	LASSO	2.18	0	0.00	0.569
	ALASSO	4.29	0	0.51	0.027
	SCAD	4.01	0	0.36	0.027
300	LASSO	2.43	0	0.02	0.017
	ALASSO	4.53	0	0.63	0.008
	SCAD	4.49	0	0.69	0.008
500	LASSO	2.57	0	0.02	0.011
	ALASSO	4.66	0	0.72	0.005
	SCAD	4.68	0	0.75	0.004

AFT = accelerated failure time; MSE = mean squared error; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

Table 3: Simulation results under Weibull AFT model ($\sigma = 1$)

n	Method	C	IC	PT	MSE
100	LASSO	2.41	0	0.02	0.203
	ALASSO	4.02	0	0.35	0.112
	SCAD	4.44	0	0.61	0.100
300	LASSO	2.59	0	0.03	0.074
	ALASSO	4.47	0	0.52	0.034
	SCAD	4.64	0	0.72	0.029
500	LASSO	2.51	0	0.04	0.052
	ALASSO	4.52	0	0.58	0.018
	SCAD	4.65	0	0.75	0.015

AFT = accelerated failure time; MSE = mean squared error; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

Table 4: Simulation results under Weibull AFT model ($\sigma = 2$)

n	Method	C	IC	PT	MSE
100	LASSO	2.95	0.12	0.05	0.746
	ALASSO	4.05	0.32	0.30	0.596
	SCAD	4.63	0.63	0.44	0.625
300	LASSO	2.84	0.00	0.06	0.268
	ALASSO	4.42	0.01	0.50	0.166
	SCAD	4.92	0.04	0.91	0.102
500	LASSO	2.92	0.00	0.02	0.212
	ALASSO	4.48	0.00	0.55	0.091
	SCAD	4.94	0.00	0.94	0.058

AFT = accelerated failure time; MSE = mean squared error; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

distributions, a t-distribution with degree of freedom 3 (denoted by t_3) and a mixture distribution with $0.5N(0,1) + 0.5N(0,9)$ (denoted by Mix). Here, t_3 and Mix are non-normal distributions with a common mean 0, but their variances are 3 and 5, respectively. For this purpose, the LN AFT model is fitted when the distribution of ϵ is N(0,1), t_3 or Mix. Investigating the behavior of fitting LN AFT model is interesting because it becomes a classical normal regression model under log-transformation of survival time and its covariate effect is estimated unbiasedly even if the baseline distribution is misspecified under no censoring (Hutton and Monaghan, 2002). The simulation scheme is the same as before, except for considering an additional high censoring with 70%. Table 5 summarizes the results using a moderate sample size as in $n = 300$; Table 5 also shows that fitting the proposed LN

Table 5: Simulation results of fitting LN AFT model (n = 300) when the normal assumption is satisfied or violated

Error	Censoring	Method	C	IC	PT	MSE
$N(0,1)$	45%	LASSO	2.42	0.00	0.00	0.052
		ALASSO	4.39	0.00	0.45	0.021
		SCAD	4.46	0.00	0.61	0.017
	70%	LASSO	2.40	0.00	0.02	0.128
		ALASSO	4.38	0.00	0.45	0.045
		SCAD	4.57	0.00	0.63	0.032
t_3	45%	LASSO	2.80	0.00	0.00	0.080
		ALASSO	4.51	0.00	0.56	0.046
		SCAD	4.69	0.00	0.71	0.052
	70%	LASSO	2.61	0.00	0.00	0.176
		ALASSO	4.32	0.01	0.43	0.109
		SCAD	4.90	0.03	0.89	0.172
Mix	45%	LASSO	2.90	0.00	0.04	0.139
		ALASSO	4.58	0.01	0.65	0.097
		SCAD	4.89	0.02	0.88	0.087
	70%	LASSO	2.88	0.00	0.02	0.252
		ALASSO	4.42	0.02	0.52	0.192
		SCAD	4.91	0.11	0.84	0.227

Note: t_3, t-distribution with degree of freedom 3; Mix, mixture distribution with $0.5N(0,1) + 0.5N(0,9)$.
LN = lognormal; AFT = accelerated failure time; MSE = mean squared error; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

AFT model is overall robust against misspecified distributions, t_3 and Mix. As expected, the MSEs are increased with censoring rate from 45% to 70%. We find that the proposed method is still robust, except for a higher IC under SCAD with Mix distribution, when censoring rate is high as in 70%.

In addition, we investigated the robustness of Weibull AFT model against mis-specifying distribution. Here, Weibull AFT model is fitted when the distribution of ϵ is non-Weibull (i.e., t_3 or Mix) under the same simulation scheme above. We again find that the simulation results (not shown) are similar to those evident in Table 5.

4. Illustration

For the illustration of the proposed method in Section 2, we consider the PBC data of the liver (Tibshirani, 1997). A total of 424 PBC patients met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. Here we consider 312 patients who participated in the randomized trial. Censoring rate due to survival was 59.8%. Table 6 summarizes the variables used in the analysis. For the analyses, all covariates (i.e., all variables except for Id, Futime and Status in Table 6) are standardized.

As presented above, we consider the two AFT models (i.e., LN and Weibull cases) with covariates in Table 7. First, we use two standard criteria of model selection: Akaike information criterion (AIC) and BIC, given by AIC= $-2\ell + 2p$ and BIC = $-2\ell + \log(n) * p$. We conduct model selection under no penalty and choose a model with lower AIC and BIC values. Table 7 indicates the results.

From Table 7, we select the LN AFT model because the values of AIC and BIC in the LN are all smaller than those of the Weibull. Here we checked the adequacy of the lognormal assumption of survival time. This can be checked by a normal hazard plot (Klein and Moeschberger, 2003, p.410), i.e., we plot $\Phi^{-1}(1 - \hat{S}_0(t))$ versus $\log t$ as shown in Figure 2. Here, $\Phi^{-1}(\cdot)$ is the inverse (i.e., probit function) of standard normal cumulative distribution function and $\hat{S}_0(t)$ is the Kaplan-Meier estimate

Table 6: Explanation of variables for primary biliary cirrhosis data

Variable	Explanation
Id	Case number
Futime	Number of days from registration to death
Status	Status at endpoint (0: survival (59.8 %), 1: death)
Drug	Types of drugs (1: D-penicillmain, 2: placebo)
Age	In years
Sex	Sex (0: male, 1: female)
Ascites	Presence of ascites (0: no, 1: yes)
Hepato	Presence of hepatomegaly or enlarged liver (0: no, 1: yes)
Spiders	Blood vessel malformations in the skin (0: no, 1: yes)
Edema	Presence of edema (0: no edema, 0.5: untreated or successfully treated, 1: edema despite diuretic therapy)
Bili	Serum bilirunbin (mg/dl)
Chol	Serum cholesterol (mg/dl)
Albumin	Serum albumin (g/dl)
Copper	Urine copper (ug/day)
Alk_phos	Alkaline phosphotase (U/liter)
Sgot	SGOT (U/ml)
Trig	Triglycerides (mg/dl)
Platelet	Platelets per cubic (ml/1000)
Protime	Prothrombin time
Stage	Histologic stage of disease

Table 7: Model selection for AFT model with primary biliary cirrhosis data

	ℓ	AIC	BIC
LN	−195.41	426.82	492.00
Weibull	−197.91	431.82	496.99

AFT = accelerated failure time; AIC = Akaike information criterion; BIC = Bayesian information criterion; LN = lognormal.

of the baseline survival function $S_0(t)$. This is expected to show an approximate straight line if the assumption of lognormal distribution is appropriate. Figure 2 shows approximatively a linear trend for the probit survival against the log of time. Therefore, the assumption of lognormal as the baseline distribution seems appropriate: see also Royston (2001) for the usefulness of lognormal AFT model. Accordingly, we use the LN AFT model for the variable selection.

Table 8 shows the estimated coefficients and SEs for the PBC in the LN case. As the result of the penalized variable selection, the values of the tuning parameters γ that minimize the BIC* in (2.14) are 0.073 for LASSO, 0.013 for ALASSO and 0.110 for SCAD, respectively. The estimates of σ_ϵ^2 are 0.850, 0.629, 0.697 and 0.727 under no penalty ($\gamma = 0$), LASSO, ALASSO, and SCAD, respectively. The LASSO chooses eleven covariates (Age, Sex, Ascites, Spiders, Edema, Bili, Albumin, Copper, Sgot, Protime, and Stage) out of the 17 covariates except for the intercept. We also confirm that these variable-selection results are similar with the LASSO results in Cox's PH model by Tibshirani (1997) even if signs of both estimates are opposite. The SCAD choose eight covariates (Age, Edema, Bili, Albumin, Copper, Sgot, Protime, and Stage). The ALASSO selects one more variable (i.e., Ascites) than in SCAD, which is not significant under no penalty. In particular, the non-zero estimates by the SCAD are generally similar to the corresponding estimates under no penalty. LASSO selects many covariates, which are not significant under no penalty. This may be because the LASSO selects unimportant variables more than ALASSO and SCAD, as evident in the lower "C" values of the LASSO in Table 1. The findings indicate that the LASSO might not properly identify important variables in the AFT models; for the frailty models, see Ha *et al.* (2014).

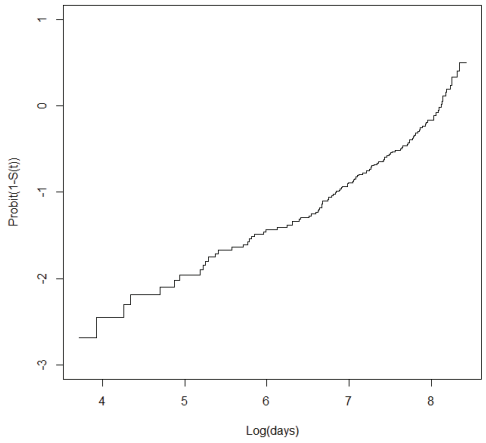

Figure 2: *Plot of probit(1 − survival) against log of days.*

Table 8: Variable selection using LN AFT model for primary biliary cirrhosis data

Variable	No penalty	LASSO	LASSO†	ALASSO	SCAD
Intercept	8.073(0.086)	7.885(0.060)	-	7.994(0.065)	7.989(0.066)
Drug	−0.002(0.069)	0.000(0.000)	0.00(0.00)	0.000(0.000)	0.000(0.000)
Age	−0.221(0.080)	−0.139(0.039)	0.17(0.09)	−0.179(0.047)	−0.099(0.028)
Sex	0.091(0.068)	0.016(0.011)	−0.01(0.03)	0.000(0.000)	0.000(0.000)
Ascites	−0.112(0.076)	−0.092(0.032)	0.04(0.07)	−0.023(0.009)	0.000(0.000)
Hepato	−0.005(0.080)	0.000(0.000)	0.00(0.00)	0.000(0.000)	0.000(0.000)
Spiders	−0.116(0.072)	−0.051(0.024)	0.02(0.05)	0.000(0.000)	0.000(0.000)
Edema	−0.185(0.081)	−0.191(0.042)	0.18(0.11)	−0.246(0.046)	−0.304(0.053)
Bili	−0.202(0.086)	−0.204(0.043)	0.35(0.12)	−0.244(0.047)	−0.306(0.053)
Chol	−0.048(0.074)	0.000(0.000)	0.00(0.00)	0.000(0.000)	0.000(0.000)
Albumin	0.106(0.077)	0.100(0.034)	−0.22(0.10)	0.029(0.011)	0.051(0.018)
Copper	−0.148(0.073)	−0.152(0.040)	0.21(0.11)	−0.143(0.037)	−0.116(0.031)
Alk_phos	−0.040(0.061)	0.000(0.000)	0.00(0.00))	0.000(0.000)	0.000(0.000)
Sgot	−0.187(0.075)	−0.103(0.035)	0.09(0.08)	−0.118(0.038)	−0.030(0.012)
Trig	0.022(0.072)	0.000(0.000)	0.00(0.00)	0.000(0.000)	0.000(0.000)
Platelet	0.004(0.072)	0.000(0.000)	0.00(0.00)	0.000(0.000)	0.000(0.000)
Protime	−0.167(0.073)	−0.123(0.038)	0.09(0.09)	−0.133(0.038)	−0.080(0.024)
Stage	−0.244(0.091)	−0.181(0.044)	0.21(0.09)	−0.259(0.055)	−0.275(0.057)

† indicates the results of variable selection from Cox's PH model by Tibshirani (1997).
LN = lognormal; AFT = accelerated failure time; LASSO = least absolute shrinkage and selection operator; ALASSO = adaptive LASSO; SCAD = smoothly clipped absolute deviation.

5. Discussion

Through penalized likelihood approach, we have shown the procedures that select important variables in the AFT model. We have demonstrated via simulation studies and illustration that the proposed variable-selection methods generally work well. Here we have found that the SCAD method performs better than the LASSO and ALASSO methods. The results confirm those in semi-parametric frailty hazard models by Ha *et al.* (2014).

The AFT model has some advantages over Cox's PH model as follows (Ha *et al.*, 2017, pp.31–32): (i)AFT model does not require a PH assumption (i.e., a strong assumption) as in the Cox's model; (ii) The interpretation of regression coefficients is easier in the AFT model than in the Cox's model;

(iii) The estimated regression parameters in AFT model are relatively robust against misspecification of the model assumption, while ones in the Cox's model can be biased. In addition, following Reid (1994), Cox pointed out that "AFT models are in many ways more appealing" than the PH models "because of their quite direct physical interpretation".

We have also demonstrated via a simulation study that the proposed method is somewhat robust against misspecification of the assumed distribution. It would be also interested to investigate the robustness of the LN or Weibull AFT model against a further mis-specifying distribution, for example, when the true distribution of survival time T is not smooth and has change points. However, comparing with an existing variable selection procedure for semiparametric AFT model will be more informative about the setting in which the proposed method is useful; this would be an interesting future work.

We have developed the variable-selection methods in AFT models with low-dimensional covariates ($n > p$). Developing the penalized AFT models with high-dimensional covariates ($n < p$) would be an interesting topic. The proposed methods are based on parametric penalized-likelihood approaches that allow for LN and Weibull distributions. Therefore, an extension to semi-parametric AFT models (Huang *et al.*, 2006; Huang and Ma, 2010) with an unspecified error distribution would be suitable for further work.

Furthermore, the proposed method can be extended to AFT models allowing for random effects that can be useful for analyzing correlated survival data (Ha *et al.*, 2017).

Acknowledgements

This research was supported by an NRF grant funded by Korea government (MEST) (No.2011-0030810).

Appendix: Derivations

(1) LN AFT model

The log-likelihood for LN AFT model is given by

$$\ell\left(\beta, \sigma_\epsilon^2; y, \delta\right) = \sum_i \left[\delta_i \left\{-\frac{1}{2}\log\sigma_\epsilon^2 + \log\varphi(m_i) - \log(1-\Phi(m_i))\right\} + \log(1-\Phi(m_i))\right],$$

where $m_i = (y_i - x_i^T\beta)/\sigma_\epsilon$. The estimating equations for β_k $(k = 0, 1, \ldots, p-1)$ and σ_ϵ^2 are as follows:

$$\frac{\partial\ell}{\partial\beta_k} = \sigma_\epsilon^{-1}\sum_i \{\delta_i m_i + (1-\delta_i)U(m_i)\}\, x_{ik} = 0,$$
$$\frac{\partial\ell}{\partial\sigma_\epsilon^2} = \frac{1}{2\sigma_\epsilon^2}\sum_i \left\{\delta_i\left(m_i^2 - 1\right) + (1-\delta_i)U(m_i)m_i\right\} = 0,$$

where $U(\cdot) = \varphi(\cdot)/(1-\Phi(\cdot))$.

For the variable selection, we have to solve the following estimating equation:

$$\frac{\partial\ell_p}{\partial\beta_k} = \frac{\partial\ell}{\partial\beta_k} - n\sum_k J_\gamma'(|\beta_k|)\mathrm{sgn}(|\beta_k|) = 0 \quad (k = 0, 1, \ldots, p-1),$$

where sgn(·) is the sign function. The negative Hessian matrix is given by

$$H_p = -\frac{\partial^2 \ell_p}{\partial(\beta, \sigma_\epsilon^2)^2} = \begin{pmatrix} -\dfrac{\partial^2 \ell_p}{\partial\beta\partial\beta^T} & -\dfrac{\partial^2 \ell_p}{\partial\beta\partial\sigma_\epsilon^2} \\ -\dfrac{\partial^2 \ell_p}{\partial\sigma_\epsilon^2\partial\beta^T} & -\dfrac{\partial^2 \ell_p}{\partial\sigma_\epsilon^2\partial\sigma_\epsilon^2} \end{pmatrix},$$

where

$$-\frac{\partial^2 \ell_p}{\partial\beta\partial\beta^T} = X^T W X + n\sum_\gamma,$$
$$-\frac{\partial^2 \ell_p}{\partial\beta_k\partial\sigma_\epsilon^2} = -\frac{\partial^2 \ell_p}{\partial\sigma_\epsilon^2\partial\beta_k} = \frac{1}{2\sigma_\epsilon^3}\sum_i \left[2\delta_i m_i + (1-\delta_i)\{U(m_i) + m_i\xi(m_i)\}\right] x_{ik},$$

and

$$-\frac{\partial^2 \ell}{\partial\sigma_\epsilon^2\partial\sigma_\epsilon^2} = \frac{1}{2\sigma_\epsilon^4}\sum_i \left[\delta_i\left(2m_i^2 - 1\right) + (1-\delta_i)m_i\left\{\frac{m_i\xi(m_i) + 3U(m_i)}{2}\right\}\right].$$

Here $\xi(m_i) = U(m_i)\{U(m_i) - m_i\}$, $W = \mathrm{diag}(w_i)$, and $w_i = \{\delta_i + (1-\delta_i)\xi(m_i)\}/\sigma_\epsilon^2$.

(2) Weibull AFT model

The log-likelihood for Weibull AFT model is given by

$$\ell(\beta, \sigma; y, \delta) = \sum_i \left[\delta_i\left\{-\log\sigma + \frac{y_i - x_i^T\beta}{\sigma}\right\} - \Lambda_i\right],$$

where $\Lambda_i = \exp\{(y_i - x_i^T\beta)/\sigma\}$. The estimating equations for β and σ are:

$$\frac{\partial \ell}{\partial \beta_k} = -\frac{1}{\sigma}\sum_i (\delta_i - \Lambda_i)x_{ik} = 0, \quad (k = 0, 1, \ldots, p-1),$$
$$\frac{\partial \ell}{\partial \sigma} = -\frac{1}{\sigma}\sum_i \{\delta_i(1 + m_i) - m_i\Lambda_i\} = 0,$$

where $\Lambda_i = \exp(m_i)$ with $m_i = (y_i - x_i^T\beta)/\sigma$.

For the variable selection, the corresponding estimating equations are given by:

$$\frac{\partial \ell_p}{\partial \beta_k} = \frac{\partial \ell}{\partial \beta_k} - n\sum_k J'_\gamma(|\beta_k|)\mathrm{sgn}(|\beta_k|) = 0,$$

The negative Hessian matrix is again given by

$$H_p = -\frac{\partial^2 \ell_p}{\partial(\beta, \sigma)^2} = \begin{pmatrix} -\dfrac{\partial^2 \ell_p}{\partial\beta\partial\beta^T} & -\dfrac{\partial^2 \ell_p}{\partial\beta\partial\sigma} \\ -\dfrac{\partial^2 \ell_p}{\partial\sigma\partial\beta^T} & -\dfrac{\partial^2 \ell_p}{\partial\sigma\partial\sigma} \end{pmatrix},$$

where

$$-\frac{\partial^2 \ell_p}{\partial\beta\partial\beta^T} = X^T W X + n\sum_{\gamma},$$
$$-\frac{\partial^2 \ell_p}{\partial\beta\partial\sigma} = -\frac{\partial^2 \ell_p}{\partial\sigma\partial\beta^T} = \frac{1}{\sigma^2}\sum_i \{\Lambda_i(1+m_i) - \delta_i\}x_{ik},$$

and

$$-\frac{\partial^2 \ell_p}{\partial\sigma\partial\sigma} = \frac{1}{\sigma^2}\sum_i \{m_i\Lambda_i(2+m_i) - \delta_i(2m_i+1)\}.$$

Here $W = \mathrm{diag}(w_i)$ with $w_i = \Lambda_i/\sigma^2$.

References

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

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

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

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

Ha ID, Jeong JH, and Lee Y (2017). *Statistical Modelling of Survival Data with Random Effects: h-Likelihood Approach*, Springer, Singapore.

Ha ID, Lee Y, and Song JK (2002). Hierarchical likelihood approach for mixed linear models with censored data, *Lifetime Data Analysis*, **8**, 163–176.

Ha ID, Pan J, Oh S, and Lee Y (2014). Variable selection in general frailty models using penalized h-likelihood, *Journal of Computational and Graphical Statistics*, **23**, 1044–1060.

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

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

Hutton JL and Monaghan PF (2002). Choice of parametric accelerated life and proportional hazard models for survival data: asymptotic results, *Lifetime Data Analysis*, **8**, 375–393.

Klein JP and Moeschberger S (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed), Springer, Berlin.

Lawless JF (1982). *Statistical Models and Methods for Lifetime data* (1st ed), Wiley, New York.

Reid N (1994). A conversation with Sir David Cox, *Statistical Science*, **9**, 439–455.

Royston P (2001). The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors, *Statistica Neerlandica*, **55**, 89–104.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B*, **58**, 267–288.

Tibshirani R (1997). The LASSO method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385–395.

Wang H, Li R, and Tsai CL (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553–568.

Wang X and Song L (2011). Adaptive lasso variable selection for the accelerated failure models, *Communications in Statistics - Theory and Methods*, **40**, 4372–4386.

Xu J, Leng C, and Ying Z (2010). Rank-based variable selection with censored data, *Statistics and computing*, **20**, 165–176.

Zhang HH and Lu W (2007). Adaptive LASSO for Cox's proportional hazards model, *Biometrika*, **94**, 691–703.

Zhang Z, Sinha S, Maiti T, and Shipp E (2018). Bayesian variable selection in the accelerated failure time model with an application to the surveillance, epidemiology, and end results breast cancer data, *Statistical Methods Medical Research*, **27**, 971–990.

Zhou M (2005). Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model, *Biometrika*, **92**, 492–498.

Zou H (2006). The adaptive Lasso and its oracle properties, *Journal of American Statistical Association*, **101**, 1418–1429.

Received March 19, 2018; Revised June 29, 2018; Accepted September 15, 2018