Variable Selection with The Modified Buckley–James Method and The Dantzig Selector for High–dimensional Survival Data

Md Hasinur Rahaman Khan*
ISRT, University of Dhaka, Bangladesh hasinur@isrt.ac.bd

J. Ewart H. Shaw
Department of Statistics, University of Warwick, UK

Abstract

We develop a group of algorithms for variable selection using the accelerated failure time (AFT) models that are based on the synthesis of the Buckley–James estimating method and the Dantzig selector. In particular, first two algorithms are based on two modified Buckley–James estimating methods that are developed for high–dimensional data. The last three algorithms are based on a weighted Dantzig selector that uses weights obtained from the two synthesis based algorithms and another obtained from a proposed form. The adaptive Dantzig selector is chosen because it satisfies the oracle properties. The methods are easy to understand and scalable and they do estimation and variable selection simultaneously. They also can deal with collinearity among the covariates and the groups of covariates. We conducted several simulation studies and one empirical analysis with a microarray dataset.

Keywords: Accelerated failure time, Buckley–James estimating equation, Dantzig selector, Variable selection

1 Introduction

Suppose $Y_i = \log(T_i)$, where T_i represents the lifetime. Then the accelerated failure time (AFT) with traditional notation, is defined by

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

The standard estimation techniques, such as ordinary least squares (OLS) cannot be employed directly to the model in (1) even if the number of predictors p is small relative to the sample size n. For dataset where $p > n$ the standard estimation techniques are more difficult to apply when variable selection is needed along with estimation.

In survival analysis the popular models are usually well suited for data with few covariates and many observations. In contrast for a typical setting of gene expression data from DNA microarray in genomic, it is necessary to consider the case where the number of covariates p exceeds the number of samples n. Variable selection techniques using the AFT models have been developed by many researchers. Such as the threshold gradient descent (Huang et al., 2006), elastic net (Wang et al., 2008; Engler & Li, 2009) and the Bayesian variable selection (Sha et al., 2006). Many estimation methods have been developed for AFT models. For example, weighted least squares (Stute 1993, 1996), and doubly penalized Buckley–James method (e.g., Wang et al., 2008).

2 Methods

2.1 First Modified Buckley–James Approach (MRBJ)

We focus on an approximating approach for the consistent root of the estimating equation as discussed in Jin et al. (2006). We use a consistent regularized estimator as the initial value in the Buckley–James iteration. This initial estimator allows the Buckley–James method to perform with high-dimensional dataset. Along with the estimation we develop the resampling procedure for estimating the limiting covariance matrix. We refer to this proposed approach as the modified resampling based Buckley–James (MRBJ).

The Buckley–James estimator $\hat{\beta}_{bj}$ is defined by $\beta = \mathrm{L}(b)_{b=\beta}$, where

$$\mathrm{L}(b) = \left\{\sum_{i=1}^{n}(X_i - \bar{X})^{\otimes 2}\right\}^{-1}\left[\sum_{i=1}^{n}(X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b)\}\right], \tag{2}$$

where $a^{\otimes 2}$ means aa^T for a vector. The following iterative algorithm is then obtained from (2).

$$\hat{\beta}_{(m)} = \mathrm{L}(\hat{\beta}_{(m-1)}), \qquad m \geq 1. \tag{3}$$

In Equation (3) we use the Dantzig selector estimator $\hat{\beta}_{ds}$, implemented for the weighted data (weighted response and predictors by the K-M weights) as the initial estimator $\hat{\beta}_{(0)}$. The estimator $\hat{\beta}_{ds}$ can be defined as below.

$$\min_{\beta} \ \|\beta\|_1 \qquad \text{subject to} \qquad \|X^T\, Y - X\,\beta)\|_\infty \leq \lambda. \tag{4}$$

We develop now a resampling procedure under high–dimensional data to approximate the distribution of $\hat{\beta}_{(m)}$ can now be developed by following Jin et al. (2006). We select an active subset $\mathcal{A}$ of important variables using an auxiliary method based on the initial estimator $\hat{\beta}_{ds}$. One potential choice could be selecting $\mathcal{A}$ as $\{j : |\hat{\beta^j}_{ds}| > \mathtt{C}\}$ for a suitable value of $\mathtt{C}$. Typically $\mathtt{C}$ will be very small for a sparse model. We now define X by $X^{\mathcal{A}}$ for those $j \in \mathcal{A}$.

Let $Z_i\,(i = 1, \cdots, n)$ be a n iid positive random variables such that $E\,(Z_i) = \text{var}\,(Z_i) = 1$. Then define

$$\mathrm{L}^*(b) = \left\{\sum_{i=1}^{n} Z_i (X_i^{\mathcal{A}} - \bar{X}^{\mathcal{A}})^{\otimes 2}\right\}^{-1} \left[\sum_{i=1}^{n} Z_i (X_i^{\mathcal{A}} - \bar{X}^{\mathcal{A}})\{\hat{Y}_i^*(b) - \bar{Y}^*(b)\}\right], \tag{5}$$

where $\bar{Y}^*(b) = n^{-1}\sum_{i=1}^{n} \bar{Y}_i^*(b)$. Equation (5) then leads to an iterative process $\hat{\beta}^*_{(m)} = \mathrm{L}^*(\hat{\beta}^*_{(m-1)})$ $m \geq 1$. The initial value $\hat{\beta}^*_{(0)}$ of the iteration process becomes $\hat{\beta}^*_{ds}$ which is obtained by optimizing criterion

$$\min_{\beta} \ \|\beta\|_1 \qquad \text{subject to} \qquad \|Z{X^{\mathcal{A}}}^T (Y - X^{\mathcal{A}}\beta)\|_\infty \leq \lambda.$$

For a given sample $(Z_1, \cdots, Z_n)$, the iteration procedure $\hat{\beta}^*_{(k)} = \mathrm{L}^*(\hat{\beta}^*_{(k-1)})$ yields a $\hat{\beta}^*_{(k)}$ $(1 \leq k \leq m)$. The empirical distribution of $\hat{\beta}^*_{(m)}$ is based on a large number of realizations that are computed by repeatedly generating the random sample $(Z_1, \cdots, Z_n)$. Then the empirical distribution is used to approximate the distribution of $\hat{\beta}_{(m)}$. It also holds from their results that the conditional distribution of $n^{\frac{1}{n}}\,(\hat{\beta}^*_{(m)} - \hat{\beta}_{(m)})$ given the data $(T_i^*, \delta_i, X_i;\, i = 1, \cdots, n)$ converges almost surely to the asymptotic distribution of $n^{\frac{1}{n}}\,(\hat{\beta}_{(m)} - \hat{\beta}_{(0)})$.

2.2 Second Modified Buckley−James Approach (BJDS)

We propose a modified Buckley−James iterative procedure by using the Dantzig selector that minimizes the ℓ_1 norm of β subject to an ℓ_∞ constraint on the error terms that are based on the weighted least squares (WLS) (Stute, 1993, 1996). In each iteration we first implement the Dantzig selector and then it follows the imputation procedure. We refer to this approach as the Buckley–James Dantzig selector (BJDS) approach. To our knowledge there are few studies where the Buckley−James method is modified by utilizing regularized estimation technique in its iterative procedure (e.g., Wang et al., 2008). The WLS approach for AFT model implementation entails modifying the least squares approach by imposing Kaplan–Meier weights.

2.2.1 BJDS Algorithm:

1. Initialization: Set the initial value for β and set $k = 0$.
2. Iteration: At the k-th iteration, compute
   (a) $\hat{Y}_i(\hat{\beta}^{k-1}) = \delta_i\, Y_i + (1 - \delta_i)\,\left[\int_{\xi_i}^{\infty} \varepsilon_i \,\frac{\mathrm{d}\hat{F}(\varepsilon_i)}{1-\hat{F}(\xi_i)} + X_i^T\hat{\beta}^{k-1}\right]$, where $\xi_i = Y_i - X_i^T\hat{\beta}^{k-1}$.
   (b) β^k by fitting the problem $\min_{\beta} \ |\beta\|_1 \qquad \text{subject to } \|X^T(\hat{Y}_i(\hat{\beta}^{k-1}) - X\beta)\|_\infty \leq \lambda$ for a chosen m iterations.
3. Stop: Go back to the iteration step until the stopping criterion is satisfied. The stopping criteria is chosen to be $\max|\beta^k| - |\beta^{k-1}| < \gamma$, where γ is pre-specified number.

The BJDS algorithm might generate oscillating estimates among the iterations because of the discontinuous nature of the discrete estimating function for β in relation to the Kaplan–Meier estimator.

2.3 The Two Stage Adaptive Dantzig Selector (DSSD)

Here we modify the typical Dantzig selector (4) in the same way as that done for the adaptive lasso (Zou, 2006) approach. In the first stage the weights are estimated from an initial estimator β^{ini}. In the second stage a weighted Dantzig selector is developed using the weights. The adaptive Dantzig selector estimator is the solution to

$$\min_{\beta} \|w\beta\|_1 \qquad \text{subject to} \qquad \|X^T(\hat{Y}(\hat{\beta}^{\text{ini}}) - X\beta)\|_\infty \leq \lambda w, \tag{6}$$

where w is the data-dependent weights vector and $\hat{Y}(\hat{\beta}^{\text{ini}})$ are the imputed failure times that are obtained by using the Buckley–James estimating equation.

By analogy with the adaptive lasso (Zou, 2006) and adaptive Dantzig selector (Li et al., 2012) we conjecture that the adaptive Dantzig selector for censored data (6) satisfies oracle properties if w is chosen cleverly, for example, $\hat{w}_j = 1/|\hat{\beta}_j^{\text{ini}}|^\nu$ for $j = 1, \cdots, p$ and for some $\nu > 0$. The reason for using such weights is to allow a relatively higher penalty for zero coefficients and lower penalty for nonzero coefficients to reduce the estimation bias and improve variable selection accuracy. The most typical choice is $\hat{w}_j = 1/|\hat{\beta}_j^{ls}|$ where $\hat{\beta}^{ls}$ is the OLS estimator of β. We introduce here an alternative choice that depends on the Z statistic value of the initial estimates. We choose $\hat{w}_j = 1/|Z(\hat{\beta}_j^{\text{ini}})|^\nu$, where $Z(\hat{\beta}_j^{\text{ini}}) = \hat{\beta}_j^{\text{ini}}/\text{SE}(\hat{\beta}_j^{\text{ini}})$ for $j = 1, \cdots, p$. Note that this alternative choice is based on the significance test of the initial estimates using Z statistic under the asymptotic normality assumption. We recommend $\hat{w}_j = 1/|Z(\hat{\beta}_j^{ls})|$ by analogy with the above $\hat{w}_j = 1/|\hat{\beta}_j^{ls}|$.

2.3.1 Computational Algorithm for DSSD:

1. Compute weights using $\hat{w} = 1/|\hat{\beta}^{\text{ini}}|$ or $\hat{w} = 1/|Z(\hat{\beta}^{\text{ini}})|$.
2. Define $X^* = X/\hat{w}$.
3. Compute the DS estimates $\hat{\beta}^*$ by solving the following problem using DASSO algorithm for each λ
   $\min_{\beta} |\beta\|_1 \qquad \text{subject to } \|X^{*T}(\hat{Y}(\hat{\beta}^{\text{ini}}) - X^*\beta)\|_\infty \leq \lambda.$
4. Now compute the adaptive DS estimates $\hat{\beta}$ as $\hat{\beta} = \hat{\beta}^*/\hat{w}$.

So both approaches the Dantzig selector and the adaptive Dantzig selector appear to have the same computation cost. Note that if we use $\hat{w} = 1/|\hat{\beta}^{\text{ini}}|^\nu$ for a specific $\hat{\beta}^{\text{ini}}$ then a two–dimensional cross–validation can be used to obtain the optimal pair of (λ, ν) for the DSSD method.

3 Numerical Examples

We have two synthesis based approaches, MRBJ and BJDS and a two stage adaptive Dantzig approach DSSD with its three implementations—DSSD-BJ1 that is the DSSD where initial estimator is used as the MRBJ estimator, DSSD-BJ2 that is the DSSD where initial estimator is used as the BJDS estimator, and DSSD-BJ1* that is similar to the DSSD-BJ1 method except that the weights are estimated based on the pivotal quantity (e.g., Z value for normal test) of the MRBJ estimators.

3.1 Simulation Studies

We estimate ρ^2 that measures the ratio of the squared error between estimated and true β to the theoretical optimum squared error assuming that the identity of the non–zero β coefficients was known. The ρ^2 is defined by

$$\rho^2 = \frac{\sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2}{\sum_{j=1}^p \min(\beta_j^2, \sigma^2)}.$$

For estimating predictive performance we use MSE_{pred} We also use M that calculates the average number of variables selected in the final model, the false positive rate denoted by F^+ (the proportion of irrelevant variables that are estimated as non-zero coefficients), and the false negative rate denoted by F^- (the proportion of relevant variables that are estimated as zero coefficients). The pairwise correlation (r_{ij}) between the i-th and the j-th components of X is set to be $0.5^{|i-j|}$. Random censorship is maintained throughout the study. We use three level of censoring $P_{\%}$– 30, 50, and 70. For underlying simulation examples the intercepts were not counted in computing the number of non-zero coefficients.

Table 1: Simulation results for case 1 ($n = 40, 100$; $p = 40$) of example I. Here block with (*) stands for the noninformative block. The theoretical value for M is 15.

		$P_{\%}=30$					$P_{\%}=50$				
Parameter	Blocks	MRBJ	BJDS	DSSD-BJ1	DSSD-BJ1*	DSSD-BJ2	MRBJ	BJDS	DSSD-BJ1	DSSD-BJ1*	DSSD-BJ2
					n = 40, p = 40						
ρ^2	-	16.17	2.39	1.35	1.35	2.39	23.56	18.49	11.32	11.32	18.49
MSE_{pred}	-	1237.2	1431.5	1587.2	1445.95	1655.12	1932.8	2093.4	1772.1	1964.0	1704.6
M	-	11.64	14.98	14.9	14.9	14.98	10.46	13.16	12.48	12.48	13.16
F^+	-	0.06	0.01	0.00	0.00	0.01	0.07	0.07	0.04	0.04	0.07
F^-	-	0.33	0.02	0.01	0.01	0.02	0.42	0.23	0.23	0.23	0.23
					n = 100, p = 40						
ρ^2	-	1.49	0.32	0.11	0.11	0.32	7.09	5.40	0.37	0.37	5.43
MSE_{pred}	-	1324.8	1331.9	1193.9	1194.1	1234.8	2195.0	2640.3	1947.7	1949.3	2174.5
M	-	14.7	15	15	15	15	13.62	16.44	15	15	15.08
F^+	-	0.01	0.00	0.00	0.00	0.00	0.04	0.06	0.00	0.00	0.02
F^-	-	0.04	0.00	0.00	0.00	0.00	0.16	0.01	0.00	0.00	0.02
					n = 40, p = 40						
	Block1	0.68	0.98	0.99	0.99	0.98	0.56	0.78	0.78	0.78	0.78
	Block2*	0.06	0.00	0.00	0.00	0.00	0.04	0.05	0.03	0.03	0.05
	Block3	0.62	0.97	0.98	0.98	0.97	0.58	0.77	0.75	0.75	0.77
PM	Block4*	0.08	0.02	0.01	0.01	0.02	0.09	0.07	0.05	0.05	0.07
	Block5	0.70	0.99	1.00	1.00	0.99	0.60	0.75	0.77	0.77	0.75
	Block6*	0.06	0.01	0.00	0.00	0.01	0.07	0.07	0.04	0.04	0.07
					n = 100, p = 40						
	Block1	0.96	1.00	1.00	1.00	1.00	0.85	0.99	1.00	1.00	0.97
	Block2*	0.04	0.00	0.00	0.00	0.00	0.02	0.05	0.00	0.00	0.02
	Block3	0.94	1.00	1.00	1.00	1.00	0.85	0.99	1.00	1.00	0.96
PM	Block4*	0.00	0.00	0.00	0.00	0.00	0.03	0.09	0.00	0.00	0.05
	Block5	0.98	1.00	1.00	1.00	1.00	0.83	0.99	1.00	1.00	0.96
	Block6*	0.01	0.00	0.00	0.00	0.00	0.05	0.06	0.00	0.00	0.02

Table 2: Simulation results for case 2 of example I ($n = 50$, $p = 60$). Here block with (*) stands for the noninformative block. The theoretical value for M is 30.

		$P_{\%}=30$					$P_{\%}=50$				
Parameter	Blocks	MRBJ	BJDS	DSSD-BJ1	DSSD-BJ1*	DSSD-BJ2	MRBJ	BJDS	DSSD-BJ1	DSSD-BJ1*	DSSD-BJ2
ρ^2	-	32.82	8.07	4.56	4.56	8.07	44.79	30.73	20.81	20.81	30.73
MSE_{pred}	-	11148.3	6694.4	6364.2	9781.1	8106.2	13025.6	12136.5	10682.6	9757.2	10442.5
M	-	15.56	34.28	34.26	34.26	34.34	13.28	24.3	24.34	24.28	24.36
F^+	-	0.01	0.19	0.16	0.16	0.19	0.01	0.07	0.06	0.06	0.07
F^-	-	0.49	0.03	0.02	0.02	0.03	0.57	0.25	0.25	0.25	0.25
PM	Block1	0.52	0.95	0.98	0.98	0.95	0.42	0.74	0.74	0.74	0.74
	Block2	0.50	0.96	0.98	0.98	0.96	0.44	0.74	0.76	0.75	0.75
	Block3*	0.01	0.19	0.19	0.18	0.20	0.01	0.07	0.06	0.06	0.07
	Block4*	0.00	0.18	0.14	0.14	0.18	0.01	0.07	0.06	0.06	0.07

3.1.1 Example I: Group Selection of Correlated Variables

Case 1: $p < n$

We consider $n = 40$ and 100, $p = 40$, $\varepsilon \sim N(0, 1)$ and $\sigma = 1$. There are six blocks of covariates $\mathbf{X}$ and each of the first five blocks contains five correlated covariates. Coefficients for β's for $j \in \{1, \cdots, 5\} \cup \{11, \cdots, 15\} \cup \{21, \cdots, 25\}$ are set to be 5 (i.e. $p_\gamma = 15$) and the remaining 25 β's are set to be 0 and $\mathbf{X} \sim MVN(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Sigma_0 & & & & & \\ & \Sigma_0 & & & & \\ & & \Sigma_0 & 0.2J & & \\ & & 0.2J & \Sigma_0 & & \\ & & & & \Sigma_0 & 0.2J \\ & & & & 0.2J & \Sigma_0 \\ & & & & & & \Sigma_1 \end{pmatrix},$$

where Σ_0 is a 5×5 matrix with diagonal elements chosen to be 1 and off–diagonal elements 0.7, Σ_1 is a 10×10 identity matrix, and J is a 5×5 matrix with all elements 1. Here $C \sim U(-a, a)$ with suitable value for a calculated analytically for two $P_{\%}$, 30 and 50. Results from 1000 simulation runs are reported in Table 1. For this example we calculate one further summary statistic called PM that represents the proportion of the total number of selected relevant variables to the total number of true relevant variables. It is worth mentioning that the practical implication of grouped variables typically is seen in gene expression data where p is large and n is small. The analysis of such grouped variables has attracted attention in many studies.

Case 2: $p > n$

We use $n = 50$, $p = 60$ and consider first 30 coefficients of β to be non-zero (i.e. $p_\gamma = 30$) and they are assumed to be 5. The remaining coefficients are fixed to be zero. We maintain everything else the same as for the first case example except that $\mathbf{X} \sim MVN(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Sigma_0 & & & \\ & \Sigma_0 & 0.2J & \\ & 0.2J & \Sigma_0 & \\ & & & \Sigma_0 \end{pmatrix},$$

where Σ_0 is a 15×15 matrix with diagonal elements 1 and off–diagonal elements 0.7, and J is a 15×15 matrix with all elements 1. The results are summarized in Table 2.

Table 3: Simulation results for example II ($n = 30$, $p = 60$). The theoretical value for M is 20.

		$r_{ij} = 0$					$r_{ij} = 0.5$				
$P_{\%}$	Methods	ρ^2	MSE_{pred}	M	F^+	F^-	ρ^2	MSE_{pred}	M	F^+	F^-
					Log-normal AFT						
	MRBJ	23.21	971.1	14.2	0.26	0.55	23.01	4877.8	14.3	0.20	0.49
	BJDS	21.13	799.4	19.5	0.34	0.36	20.85	835.4	19.3	0.36	0.40
30	DSSD-BJ1	21.04	1136.6	19.4	0.34	0.36	20.70	1000.4	19.7	0.36	0.40
	DSSD-BJ1*	21.04	913.5	19.4	0.34	0.36	20.70	941.0	19.7	0.38	0.39
	DSSD-BJ2	21.13	1827.2	19.5	0.34	0.36	20.85	1874.4	19.3	0.36	0.40
	MRBJ	31.78	763.3	11.8	0.24	0.65	27.98	5687.0	11	0.21	0.66
	BJDS	26.16	934.6	13.8	0.30	0.61	25.47	673.1	13.8	0.28	0.59
50	DSSD-BJ1	25.96	1003.4	13.8	0.30	0.61	24.85	879.6	13.9	0.29	0.60
	DSSD-BJ1*	25.96	859.8	13.8	0.29	0.60	24.85	998.2	13.9	0.29	0.60
	DSSD-BJ2	26.16	1552.8	13.8	0.20	0.61	25.47	1420.1	13.8	0.28	0.59
	MRBJ	30.19	665.7	7.3	0.12	0.76	31.42	6390.3	7.2	0.13	0.77
	BJDS	28.39	659.3	8	0.15	0.75	33.56	662.6	7.9	0.16	0.76
70	DSSD-BJ1	27.53	764.2	8	0.14	0.74	30.63	588.9	8	0.17	0.77
	DSSD-BJ1*	27.53	672.3	8	0.14	0.74	30.63	625.8	8	0.17	0.77
	DSSD-BJ2	28.39	994.5	8	0.15	0.75	33.56	920.0	7.9	0.16	0.76
					Weibull AFT						
	MRBJ	22.47	963.0	14.1	0.21	0.50	27.07	6819.8	13.7	0.24	0.55
	BJDS	20.89	711.7	19.3	0.34	0.38	22.55	712.9	13.7	0.402	0.43
30	DSSD-BJ1	20.90	1060.1	19.3	0.34	0.38	22.46	1209.8	19.3	0.40	0.43
	DSSD-BJ1*	20.90	2598.1	19.3	0.34	0.38	22.46	1111.8	19.3	0.40	0.43
	DSSD-BJ2	20.89	1851.7	19.3	0.34	0.38	22.55	1902.1	19.5	0.40	0.43
	MRBJ	32.11	1085.4	10.4	0.20	0.68	24.68	3866.5	11.1	0.17	0.62
	BJDS	27.08	846.4	13.4	0.26	0.59	22.42	489.1	13.8	0.24	0.55
50	DSSD-BJ1	27.10	999.0	13.1	0.26	0.60	22.35	579.7	13.9	0.24	0.55
	DSSD-BJ1*	27.10	1020.5	13.1	0.26	0.60	22.35	570.0	13.9	0.24	0.55
	DSSD-BJ2	27.08	1403.6	13.4	0.26	0.59	22.42	1412.9	13.8	0.24	0.55
	MRBJ	29.25	740.0	7.1	0.16	0.80	31.58	4605.2	7.5	0.16	0.79
	BJDS	29.48	720.4	7.7	0.17	0.78	30.89	669.0	7.9	0.16	0.76
70	DSSD-BJ1	28.90	707.9	7.6	0.16	0.78	31.28	729.6	7.9	0.16	0.76
	DSSD-BJ1*	28.90	727.5	7.6	0.16	0.78	31.28	736.4	7.9	0.16	0.76
	DSSD-BJ2	29.48	900.5	7.7	0.17	0.78	30.89	1000.4	7.9	0.16	0.76

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

Methods	MRBJ	BJDS	DSSD-BJ1	DSSD-BJ1*	DSSD-BJ2
MRBJ	20	01	13	13	04
BJDS	01	08	02	02	07
DSSD-BJ1	13	02	64	64	10
DSSD-BJ1*	13	02	64	64	10
DSSD-BJ2	04	07	10	10	41

3.1.2 Example II

We set $n = 30$, $p = 60$ and the first 20 coefficients for β to be 5 (i.e. p_γ is 20) and the remaining coefficients of β to be zero, $\mathbf{X} \sim U(0, 1)$. We choose two AFT models, the log-normal (i.e., (1) with $\varepsilon_i \sim N(0, 1)$) and the Weibull (i.e., (1) $\varepsilon_i \sim \log(\text{Weibull})$ leading to Weibull distributed lifetimes). Now for both AFT models the censoring time is generated from a log-normal distribution that has the form $\exp[N(c_0\sqrt{1+\sigma},\ (1+\sigma^2))]$ where we choose $\sigma = 1$. The c_0 is determined analytically based on the pre-specified three $P_{\%}$.

Table 3 summarizes the results from 1000 simulation runs. The summary results from all above simulations reveal the following findings: If there is collinearity among the covariates then the error and the prediction error generally increase for the methods. The rate of false positive and false negative also slightly increase. If censoring increases the methods underestimate the coefficient estimation severely and the error, predicted error, and false negative rate deteriorate. Only the false positive rate gets better (reduces) as censoring increases. This happens to both type of datasets, low and high–dimensional. Among the approaches MRBJ is the worst overall. All the remaining methods perform fairly similarly to each other in terms of coefficient estimation, identifying the correct coefficients and the error although BJDS tends to generate less predicted error.

3.2 Real Data Example: DLBCL Data of Rosenwald et al.

We use the Diffuse large–B–cell lymphoma (DLBCL) dataset of Rosenwald et al. (2002) to illustrate the application of our proposed variable selection algorithms. The data consist of measurements of 7399 genes from 240 patients of which, 127 were deceased (uncensored) and 95 were alive (censored) at the end of the study. The authors fitted univariate Cox PH models on each probe after dividing the data into a training set with 160 patients and a test set with 80 patients. Genes that were associated with a good or a bad outcome (overall survival after chemotherapy) at a significance level of $p < 0.01$ were assigned to gene–expression signatures. This leads to a set of 473 genes. We fitted univariate log-normal AFT models to each individual gene using the training data of 160 patients. We found 463 genes that are significant at level $p < 0.01$. For further analysis with our proposed approaches we finally select 392 genes that are found common to the both pre-filtering processes the univariate Cox PH model and the AFT model.

We employ all five approaches to the training data to select the predictive genes among these 392 candidate genes. The MRBJ selects 20 genes, the BJDS returns a final model of 8 genes, the DSSD-BJ1 identifies 64, the DSSD-BJ1* finds 64, the DSSD-BJ2 selects 41. Table 4 represents the number of genes identified by the proposed methods along with the number of common genes between the methods. There are 22 genes that are found common to at least three proposed methods. We calculated the predictive MSE

for the methods. Slightly lower predictive errors are found for two methods: the BJDS and the DSSD-BJ2. The MRBJ seems to be the worst as it tends to give the highest predictive MSE.

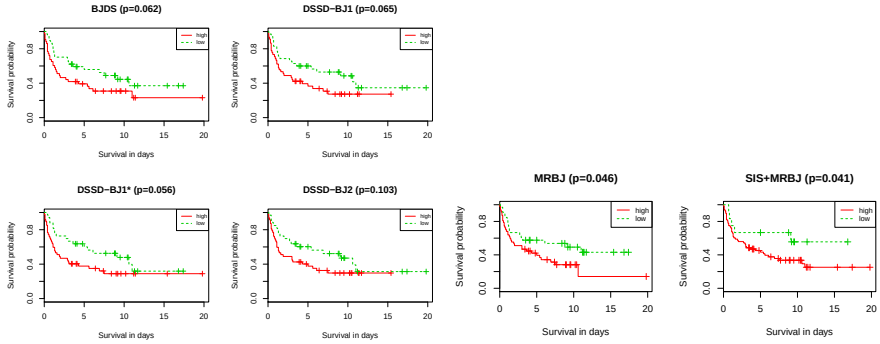

Figure 1: Survival comparison between the high risk group and low risk group using different methods.

To evaluate and validate the predictive performance of the methods, we use the obtained models to predict the risk of death in the DLBCL test dataset. The Kaplan–Meier plots of overall survival clearly show distinct differences in the survival times in the validation dataset (Figure 1). The p-value suggests that the methods BJDS, DSSD-BJ1, DSSD-BJ1*, and MRBJ perform very well to group the patients' survival time into two risk sets with their respective risk score. The other method, DSSD-BJ2 performs slightly worst, particularly with its risk score based on the much higher genes (41). The p-values for comparing high and low risk groups by the methods are 0.062 for the BJDS, 0.065 for the DSSD-BJ1, 0.056 for the DSSD-BJ1*, 0.102 for the DSSD-BJ2, and 0.046 for the MRBJ.

4 Discussion

We have proposed five variable selection methods that are suitable to apply for AFT models for the dataset with high-dimensionality. The MRBJ provides consistent estimators for coefficients using a resampling method and that also produces a limiting covariance matrix for the parameters for high–dimensional sparse linear model. The BJDS shows how the Dantzig selector can be utilized in the Buckley–James iterative process to establish a sparse variable selection process and it is able to perform simultaneous parameter estimation and variable selection. The last three Dantzig selector for survival data (DSSD) approaches uses an adaptive Dantzig selector designed to work for both low and high–dimensional survival data. We also introduce, under the DSSD-BJ1*, a new data–dependent weighting scheme obtained using the normal Z statistic of the initial estimators.

References

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology* **8,** Article 14.

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

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

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, K., Smeland, E., and Staut, L. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine* **346,** 1937–1947.

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

Stute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multivariate Analysis* **45,** 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23,** 461–471.

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

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