Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 137-161
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Concave Group Selection of Nonparameter Additive Accelerated Failure Time Model

Ling Zhu

Jinan University, Guangzhou, China
Email: zhuling132@163.com

How to cite this paper: Zhu, L. (2021) Concave Group Selection of Nonparameter Additive Accelerated Failure Time Model. *Open Journal of Statistics*, **11**, 137-161.
https://doi.org/10.4236/ojs.2021.111008

Received: December 28, 2020
Accepted: February 2, 2021
Published: February 5, 2021

Abstract

In this paper, we have studied the nonparameter accelerated failure time (AFT) additive regression model, whose covariates have a nonparametric effect on high-dimensional censored data. We give the asymptotic property of the penalty estimator based on GMCP in the nonparameter AFT model.

Keywords

Accelerated Failure Time Model, Nonparameter Model, Group Minimax Concave Penalty, Weighted Least Squares Estimation

1. Introduction

With the development of the Internet, high-dimensional data has been widely collected in life, especially in the field of medical research and finance, the results or responses of data are censored, so the study of high-dimensional censored data is meaningful. However, due to the impact of "disaster of dimension", the study of high-dimensional data becomes extremely difficult, and some special methods must be adopted to deal with it. As the number of data dimensions increases, the performance of high-dimensional data structures declines rapidly. In low-dimensional spaces, we often use Euclidean distance to measure the similarity between data; but in high-dimensional spaces, this kind of similarity no longer exists, which makes the data mining of high-dimensional data very severely challenging. On the one hand, the performance of the data mining algorithm based on the index structure is reduced; on the other hand, many mining methods based on the entire spatial distance function will fail. By reducing the number of dimensions, the data can be reduced from high to low dimensions, and then using low-dimensional data processing methods. Therefore, the study of effective dimensionality reduction methods becomes significant in statistics.

In many studies, the main results or responses of survival data are censored. Survival analysis is another important theme of statistics, and it has been widely used in medical research and finance. Therefore, the study of survival data has attracted a lot of attention. The Cox model [1] is the most commonly used regression model for survival data. The alternative method of the PH model is the accelerated failure time model, which directly correlates the logarithm of the failure time with the covariate, and is similar to the traditional linear model, which is easier to explain than the PH model. [2] takes into account both Lasso and threshold gradient oriented regularization for high-dimensional AFT model estimation and variable selection. [3] uses partial least squares (PLS) and Lasso methods to select variables in AFT models with high-dimensional covariates. [4] proposed a robust weighted minimum absolute deviation method to estimate the high-dimensional AFT model. [5] uses COSSO penalty in the nonparameter AFT model for variable selection [6] in the high-dimensional nonparameter AFT model, using the reproduction kernel Hilbert norm penalty for estimation, a new enhanced algorithm is proposed for censoring time data. The algorithm is suitable for fitting parameter accelerated failure time models. [7] studied the elastic net method for variable selection under the Cox proportional hazard model and the AFT model with high-dimensional covariates. [8] developed a robust prediction model for event time results through LASSO regularization. This model is aimed at the Gehan estimation of high-dimensional prediction variables accelerated failure time AFT model. [9] extends rank-based Lasso estimation to the estimation and variable selection in the high-dimensional partial linear acceleration failure time model. [10] uses the bridge penalty for regular estimation and parameter selection of high-dimensional AFT models. Based on the high-dimensional semi-parameter accelerated failure time model, [11] proposed the Buckley-James method of double penalty, which can perform variable selection and parameter estimation at the same time. [12] has developed a method for quickly predicting variable selection and contraction estimation of high-dimensional predictive variable AFT models. The model is related to the correlation vector machine (RVM), which relies on maximum posterior estimation to get sparse estimates quickly. [13] proposes a semiparametric regression model whose covariate effect contains parametric and nonparametric parts. The selection of parametric covariates is achieved by iterative LASSO method, and the nonparametric components are estimated using the sieve method [14], and based on kullback-leibler geometry [15], an empirical model selection tool for nonparameter components was obtained. However, they leave behind some theoretical issues that have not yet been resolved. [16] takes into account the estimation and variable selection of LASSO and MCP in AFT models with high covariates. [17] implements regularization in the high-dimensional AFT model $L_{1/2}$ for variable selection. [18] proposed a covariate adjustment screening and variable selection procedure under the accelerated failure time model. It also appropriately adjusted the low-dimensional confounding factors to achieve a more

accurate estimation of regression coefficients. [19] proposed an adaptive elastic net and weighted elastic net with censored data and high-dimensional variable selection in the AFT model. [20] proposed to apply a tensor recursive neural network architecture to extract latent representations from the entire patient medical record of the high-dimensional AFT model. [21] considers a novel Sparse L_2 Boosting algorithm, which is based on a semiparameter variable coefficient accelerated failure time model of right-censored survival data with high-dimensional covariates model prediction and variable selection. [22] developed a variable selection method in an AFT model with high-dimensional predictive variables, which consists of a set of algorithms based on two widely used techniques in the field of variable selection in survival analysis synthesis: Buckley-James method and Dantzig selector.

In this article, based on potential predictors, we applied the GMCP (Group Minimax Concave Penalty) penalty method for the first time to the study of a high-dimensional nonparametric accelerated failure time additive regression model (2.1) (MCP, [23]). The weighted least squares solution of the model based on GMCP penalty is given. We also derived the group coordinate descent algorithm used to calculate the GMCP estimate in this model. Our simulation results show that the weighted least squares estimation based on GMCP penalty works well in the high-dimensional nonparameter accelerated failure time additive regression model, and is superior to the GLasso (Group Least Absolute Shrinkage and Selection Operator) penalty method.

The rest of the paper is organized as follows. In Section 2, we describe the nonparameter accelerated failure time additive regression (NP-AFT-AR) model and our research methods. In Section 3, we give the asymptotic oracle property of GMCP estimation. The simulation results are given in Section 4. Verification of actual data is given in Section 5. The conclusion is given in Section 6.

2. Models and Methods

2.1. Model

In this paper, we study the following nonparametric accelerated failure time additive regression (NP-AFT-AR) model to describe the relationship between the independent predictors or covariates X_j's and the failure time T:

$$T = \exp\left(\boldsymbol{\eta}_0 + \sum_{j=1}^{p} f_j\left(X_j\right) + \varepsilon\right) \tag{2.1}$$

where $\boldsymbol{\eta}_0$ is the intercept, $\boldsymbol{X} = \left(X_1, \cdots, X_p\right)$ is a $p \times 1$ vector of covariates, f_j's are unknown smooth functions with zero means, *i.e.*, $Ef_j\left(X_j\right) = 0$ and ε is the random error term with mean zero and a finite variance σ^2. We consider sample size is small $n < p$, assuming that some additive components f_j are zero, the main purpose of our research is to find the non-zero components and zero components; the second goal is to find the specific functional form of the non-zero components in order to propose a more parsimonious model. In this

study, we apply the GMCP penalty in the proposed NP-AFT-AR model for component selection and estimation. We use B-splines to parameterize the non-parameter components, then invoke the inverse probability-of-censoring weighted least squares method to achieve the goals. We treat the spline approximation for each component as a group of variables subject to selection. By the GMCP penalty approach, we show that the proposed method can select significant component functions by choosing the nonzero spline basis functions.

2.2. Weighted Least Squares Estimation

We define T_i as the i^{th} subject's survival time, and let C_i denote the censoring time and δ_i denote the event indicator, *i.e.*, $\delta_i = I(T_i \le C_i)$; which takes value 1 if the event time is observed, or 0 if the event time is censored. Define Y_i as the minimum of the survival time and the censoring time, *i.e.*, $Y_i = \log(\min(T_i, C_i))$: Then, the observed data are in the form (Y_i, δ_i, X_i), $i = 1, \cdots, n$. which are assumed to be an independent and identically distributed (i.i.d.) sample from $(Y, \delta, \boldsymbol{X})$.

Let $Y_{(1)} \le \cdots \le Y_{(n)}$ be the order statistics of Y_i's, $\delta_{(1)}, \cdots, \delta_{(n)}$ and $X_{(1)}, \cdots, X_{(n)}$ are the associated censoring indicators and covariates. Let F be the distribution of T and $\widehat{F}_n$ be its Kaplan-Meier estimator $\widehat{F}_n(y) = \sum_{i=1}^{n} \omega_{ni} 1(Y_{(i)} \le y)$, where the ω_{ni}'s are Kaplan-Meier weights ([24]) calculated by

$$\omega_{n1} = \frac{\delta_{(1)}}{n}, \quad \omega_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \quad i = 2, \cdots, n$$

[4] showed that the weights, ω_{ni}'s, are the jumps in the Kaplan-Meier estimator. These are equivalent to the inverse probability-of-censoring weights ([25] [26]), $\omega_{ni} = \delta_{(i)} \big/ \widehat{G}_n(Y_{(i)} -)$; where $\widehat{G}_n$ is the Kaplan-Meier estimator of G, the distribution function of C. The Stute's weighted least squares loss function for the NP-AFT-AR model (2.1) is defined as

$$Q_n = \frac{1}{2} \sum_{i=1}^{n} n\omega_{ni} \left\{ Y_{(i)} - \eta_0 - \sum_{j=1}^{p} f_j\left(X_{(i)j}\right) \right\}^2 \tag{2.2}$$

Here, we use B-spline basis functions to approximated unknown functions f_j's. For every function component, assuming that X_j is bounded; and $E\{f_j(X_j)\} = 0, j = 1, \cdots, p$; The basis functions are determined by the order $(p+1)$ and the number of interior knots κ. The total number of B-spline basis functions for each function component would be $p + \kappa + 1$: For identifiability, satisfy $Ef_j(X_j) = 0$; we take the total number of basis functions to be $M_n = p + \kappa$ only and center all the basis functions at their means. Then the B-splines approximation for each function component, $f_j(X_j), j = 1, \cdots, p$; is given by

$$f_j(X_j) \approx \sum_{k=1}^{M_n} \beta_{jk} B_{jk}(X_j)$$

where $B_{jk}(X_j)$ are the B-spline basis functions and $\boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \beta_{jM_n})^{\mathrm{T}}$ is

the corresponding coefficient parameter vector. Let $\boldsymbol{B}_j$ denote the $n\times M_n$ design matrix of B-spline basis of the j^{th} predictor and $\boldsymbol{B}_{j(i)}$ be its i^{th} row vector corresponding to the sorted data. Denote the $n\times pM_n$ design matrix as $\boldsymbol{B}=\left(\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_p\right)$; the i^{th} row of $\boldsymbol{B}$ as $\boldsymbol{B}_{(i)}$; and the corresponding parameter vector as $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_p^{\mathrm{T}}\right)^{\mathrm{T}}$. Then we have

$$\begin{aligned}\sum_{j=1}^{p} f_j\left(X_{(i)j}\right) &= f_1\left(X_{(i)1}\right)+\cdots+f_p\left(X_{(i)p}\right)\\ &\approx \sum_{k=1}^{M_n}\beta_{1k}B_{1k}\left(X_{(i)1}\right)+\cdots+\sum_{k=1}^{M_n}\beta_{pk}B_{pk}\left(X_{(i)p}\right)\\ &= \sum_{j=1}^{p}\boldsymbol{B}_{(i)j}\boldsymbol{\beta}_j\end{aligned} \tag{2.3}$$

By plugging Equation (2.3) into Equation (2.2), we will get the new loss function as following:

$$Q_n\left(\eta_0,\boldsymbol{\beta}\right)=\frac{1}{2}\sum_{i=1}^{n}n\omega_{ni}\left\{Y_{(i)}-\eta_0-\sum_{j=1}^{p}\boldsymbol{B}_{(i)j}\boldsymbol{\beta}_j\right\}^2 \tag{2.4}$$

By centering $\boldsymbol{B}_{(i)j}$ and $Y_{(i)}$ with their ω_{ni}-weighted means, the intercept becomes 0. Denote $\tilde{\boldsymbol{B}}_{(i)j}=\left(n\omega_{ni}\right)^{1/2}\left(\hat{\boldsymbol{B}}_{(i)j}-\bar{\boldsymbol{B}}_{j\omega}\right)$ and $\tilde{Y}=\left(n\omega_{ni}\right)^{1/2}\left(Y_{(i)}-\bar{Y}_{\omega}\right)$; where $\bar{\boldsymbol{B}}_{j\omega}=\sum_{i=1}^{n}\omega_{ni}\boldsymbol{B}_{(i)j}\big/\sum_{i=1}^{n}\omega_{ni}$ and $\bar{Y}_{\omega}=\sum_{i=1}^{n}\omega_{ni}Y_{(i)}\big/\sum_{i=1}^{n}\omega_{ni}$ Let $\|\boldsymbol{a}\|_2=\left(\sum_{j=1}^{m}\left|a_j\right|^2\right)^{1/2}$ denote the L_2 norm of any vector $\boldsymbol{a}\in R^m$. For simplicity, we use $\tilde{\boldsymbol{B}}_j=\left(\tilde{B}_{(1)j},\cdots,\tilde{B}_{(n)j}\right)^{\mathrm{T}}$ and $\tilde{Y}=\left(\tilde{Y}_{(1)},\cdots,\tilde{Y}_{(n)}\right)^{\mathrm{T}}$. Then we can rewrite the Stute's weighted least squares loss function Equation (2.4) as

$$Q_n\left(\boldsymbol{\beta}\right)=\frac{1}{2}\sum_{i=1}^{n}\left\{\tilde{Y}_{(i)}-\sum_{j=1}^{p}\tilde{B}_{(i)j}\boldsymbol{\beta}_j\right\}^2=\frac{1}{2}\left\|\tilde{Y}-\sum_{j=1}^{p}\tilde{\boldsymbol{B}}_j\boldsymbol{\beta}_j\right\|_2^2 \tag{2.5}$$

2.3. Weighted Least Square Estimation of GMCP Penalty

B-splines approximation is used on the unknown functions, which transforms the nonparameter regression into a parameter regression that makes variable selection and parameter estimation easier to solve. Meanwhile, the grouped variables in $\tilde{\boldsymbol{B}}_j$; *i.e.*, $\tilde{B}_{jk};k=1,\cdots,M_n$; for each $j=1,\cdots,p$, are all related to the variable X_j; so we can consider B-spline basis functions for each nonparameter function f_j to be a group. Instead of selecting the significant nonparameter functions, our task converts to choosing the significant B-spline basis functions from $\tilde{\boldsymbol{B}}_j$ or nonzero coefficients from $\boldsymbol{\beta}_j$.

In order to carry out variable selection at the group and individual variable levels simultaneously. In our case, the GMCP penalty function is

$$\rho_{\gamma}\left(\left\|\boldsymbol{\beta}_j\right\|_{A_j},\lambda\right)=\lambda\int_0^{\|\boldsymbol{\beta}_j\|_{A_j}}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x \tag{2.6}$$

where γ is a parameter that controls the concavity of ρ and λ is the penalty parameter. Here $x_+=x1_{\{x\geq 0\}}$. We require $\lambda\geq 0$ and $\gamma>1$. The term

MCP comes from the fact that it minimizes the maximum concavity measure defined at (2.2) of [23], subject to conditions on unbiasedness and selection feature. The MCP can be easily understood by considering its derivative

$$\dot{\rho}_\gamma\left(\left\|\boldsymbol{\beta}_j\right\|_{A_j},\lambda\right)=\lambda\left(1-\frac{\left\|\boldsymbol{\beta}_j\right\|_{A_j}}{\gamma\lambda}\right)_+ \tag{2.7}$$

where for any $m\times 1$ vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_1$ is the L_1 norm: $\|\boldsymbol{a}\|_1=|a_1+\cdots+a_m|$, $\lambda>0$ is the penalty tuning parameter and $A_j=\{k:\beta_{jk}\in\boldsymbol{\beta}_j\}$. In our case, each A_j represents the j^{th} group of basis functions, *i.e.*, $\tilde{B}_{jk},k=1,\cdots,M_n$; the values of the basis functions for each nonparameter function f_j may be different from those for another function $f_{j'}$; and when $j\neq j'$; we assume there is no overlap between groups. Now combining the objective function in Equation (2.5) and the penalty function in Equation (2.6), we have the penalized weighted least squares objective function for the proposed NP-AFT-AR model as follows:

$$Q_{n\lambda}(\boldsymbol{\beta})=\frac{1}{2}\left\|\tilde{Y}-\sum_{j=1}^{p}\tilde{\boldsymbol{B}}_j\boldsymbol{\beta}_j\right\|_2^2+\sum_{j=1}^{p}\lambda\int_0^{\|\boldsymbol{\beta}_j\|_{A_j}}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x \tag{2.8}$$

We can conduct group or component selection and estimation by minimizing $Q_{n\lambda}(\boldsymbol{\beta})$: If $\|\boldsymbol{\beta}_j\|_{A_j}=0$; it implies that the function component f_j is deleted, otherwise, it is selected, further, the individual basis functions within a group can be selected.

2.4. Computation

We derive a group coordinate descent algorithm for computing $\boldsymbol{\beta}$. This algorithm is a natural extension of the standard coordinate descent algorithm ([27]). It has also been used in calculating the penalized estimates based on concave penalty functions ([28]).

The group coordinate descent algorithm optimizes a target function with respect to a single group at a time, iteratively cycling through all groups until convergence is reached. It is particularly suitable for computing $\boldsymbol{\beta}$, since it has a simple closed form expression for a single-group model, see (2.11) below.

We write $\boldsymbol{A}_j=\boldsymbol{R}_j$ for an $M_n\times M_n$ upper triangular matrix $\boldsymbol{R}_j$ via the Cholesky decomposition. Let $\boldsymbol{\theta}_j=\boldsymbol{R}_j\boldsymbol{\beta}_j$ and $\hat{\boldsymbol{B}}_j=\tilde{\boldsymbol{B}}_j\boldsymbol{R}_j^{-1}$. Simple algebra shows that

$$Q(\boldsymbol{\theta},\lambda,\gamma)=\frac{1}{2}\left\|\tilde{\boldsymbol{Y}}-\sum_{j=1}^{p}\hat{\boldsymbol{B}}_j\boldsymbol{\theta}_j\right\|_2^2+\sum_{j=1}^{p}\lambda\int_0^{\|\boldsymbol{\theta}_j\|}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x \tag{2.9}$$

Note that $n^{-1}\hat{\boldsymbol{B}}_j'\hat{\boldsymbol{B}}_j=\boldsymbol{R}_j'\left(n^{-1}\hat{\boldsymbol{B}}_j'\hat{\boldsymbol{B}}_j\right)\boldsymbol{R}_j^{-1}=I_{m_n}$. $\hat{Y}_j=\hat{Y}-\sum_{k\neq j}^{p}\hat{\boldsymbol{B}}_k\boldsymbol{\theta}_k$ and

$$Q_j(\boldsymbol{\theta}_j,\lambda,\gamma)=\frac{1}{2}\left\|\tilde{Y}_j-\hat{\boldsymbol{B}}_j\boldsymbol{\theta}_j\right\|_2^2+\lambda\int_0^{\|\boldsymbol{\theta}_j\|}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x \tag{2.10}$$

Let $\boldsymbol{\eta}_j=\hat{\boldsymbol{B}}_j\left(\hat{\boldsymbol{B}}_j'\hat{\boldsymbol{B}}_j\right)^{-1}\hat{Y}_j$. For $\gamma>1$, it can be verified that the value that mi-

nimizes $Q_j(\boldsymbol{\theta},\lambda,\gamma)$ is

$$\tilde{\theta}_{j,GM}(\lambda,\gamma)=M(\boldsymbol{\eta}_j;\lambda,\gamma)\equiv\begin{cases}0 & \text{if } \|\boldsymbol{\eta}_j\|\le\lambda\\ \dfrac{\gamma}{\gamma-1}\left(1-\dfrac{\lambda}{\|\boldsymbol{\eta}_j\|}\right)\boldsymbol{\eta}_j & \text{if } \lambda<\|\boldsymbol{\eta}_j\|\le\gamma\lambda\\ \boldsymbol{\eta}_j & \text{if } \|\boldsymbol{\eta}_j\|>\gamma\lambda\end{cases}\quad(2.11)$$

In particular, when $\gamma=\infty$, we have

$$\tilde{\theta}_{j,GL}=\left(1-\frac{\lambda}{\|\boldsymbol{\eta}_j\|}\right)_+\boldsymbol{\eta}_j,$$

which is the GLasso estimate for a single-group model ([29]).

The group coordinate descent algorithm can now be implemented as follows. Suppose the current values for the group parameter $\tilde{\boldsymbol{\theta}}_k^{(s)}, k\neq j$ are given. We want to minimize $Q(\boldsymbol{\theta},\lambda,\gamma)$ with respect to $\boldsymbol{\theta}_j$. Let

$$Q_j(\boldsymbol{\theta}_j,\lambda,\gamma)=\frac{1}{2}\left\|\tilde{Y}-\sum_{k\neq j}\hat{\boldsymbol{B}}_k\tilde{\boldsymbol{\theta}}_k^{(s)}-\hat{\boldsymbol{B}}_j\boldsymbol{\theta}_j\right\|_2^2+\lambda\int_0^{\|\boldsymbol{\theta}_j\|}\left(1-\frac{x}{\gamma\lambda}\right)_+\mathrm{d}x\quad(2.12)$$

and write $\tilde{Y}_j=\sum_{k\neq j}\hat{\boldsymbol{B}}_k\tilde{\boldsymbol{\theta}}_k^{(s)}$ and $\eta_j=n^{-1}\hat{\boldsymbol{B}}_j'(\tilde{Y}-\tilde{Y}_j)$. Let $\tilde{\boldsymbol{\theta}}_j$ be the minimizer of $Q_j(\boldsymbol{\theta}_j,\lambda,\gamma)$. When $\gamma>1$, we have $\tilde{\boldsymbol{\theta}}_j=M(\eta_j,\lambda,\gamma)$, where M is defined in (2.11).

For any given (λ,γ), we use (2.11) to cycle through one component at a time. Let $\tilde{\boldsymbol{\theta}}^{(0)}=\left(\tilde{\boldsymbol{\theta}}_1^{(0)\prime},\cdots,\tilde{\boldsymbol{\theta}}_p^{(0)\prime}\right)'$ be the initial value. The proposed coordinate descent algorithm is as follows.

Initial vector of residuals $r=Y-\tilde{Y}$, where $\tilde{Y}=\sum_{j=1}^{p}\hat{\boldsymbol{B}}_j\boldsymbol{\theta}_j^{(0)}$, For $s=0,1,\cdots$, carry out the following calculation until convergence. For $j=1,\cdots,p$, repeat the following steps.

Step 1: Calculate $\tilde{\boldsymbol{\eta}}_j=n^{-1}\hat{\boldsymbol{B}}_j'r+\tilde{\boldsymbol{\theta}}_j^{(s)}$.

Step 2: Update $\boldsymbol{\theta}_j^{(s+1)}=M(\tilde{\boldsymbol{\eta}}_j;\lambda,\gamma)$.

Step 3: Update $r=r-\hat{\boldsymbol{B}}_j\left(\boldsymbol{\theta}_j^{(s+1)}-\boldsymbol{\theta}_j^{(s)}\right)$ and $j=j+1$.

The last step ensures that r holds the current values of the residuals. Although the objective function is not necessarily convex, it is convex with respect to a single group when the coefficients of all the other groups are fixed.

3. Asymptotic Oracle Properties of GMCP

Let $|A|$ denote the cardinality of any set $A\in\{1,\cdots,p\}$ and $d_A=|A|M_n$. Define

$$B_A=\left(\tilde{B}_{jk},k=1,\cdots,M_n;j\in A\right)\text{ and }\Sigma_A=\frac{B_A'B_A}{n}$$

Here B_A is $n\times nd_A$ dimensional sub-design matrix corresponding to the variables in A, Denote $\|f_j\|_2=\left[Ef_j^2(X_j)\right]$ We make the following assumptions.

Similar to [5], we assume:

(C1) T and C are independent.

(C2) $Pr(T \le C \mid T, X) = Pr(T \le C \mid T)$.

(C3) $E(T^2) < \infty$ and $E(\varepsilon \mid \boldsymbol{X}) = 0$.

(C4) Denote τ_T and τ_C as the least upper bounds of T and C, respectively. Then $\tau_T < \tau_C$ or $\tau_T = \tau_C = \infty$.

(C5) f^2 has finite envelope function.

(C6) $E\{f(\boldsymbol{X}) - f^0(\boldsymbol{X})\}^2$ for $f \ne f^0$.

These assumptions correspond to the conditions in [30]. In the random censorship model, (C1) is a basic assumption. (C2) given the failure time T, the censoring indicator is independent of the $\boldsymbol{X}$. (C3) in least-squares estimation, we need the second moment. (C4) assumes the probability of an event being observed is greater than zero, which guarantees the consistency of the estimator. (C5) is a fundamental condition for the consistency and convergence rate in the proofs, and is used in the entropy calculation. (C6) guarantees that f_0 is identifiable.

(C7) There exist constant $q^* > 0, c_1 > 0$ and $c_2 > 0$ where $0 < c_1 \le c_2 < \infty$ such that

$$c_1 \le \frac{\|B_A v\|_2^2}{n} \le c_2, \forall |A| = q^*, \|v\|_2 = 1 \text{ and } v \in R^{d_A}$$

(C8) There is a small constant $\eta_1 \ge 0$ such that $\Sigma_{k \in A_0} \|f_j\|_2 \le \eta_1$.

(C9) The random errors $\varepsilon_i, i = 1, \cdots, n$ are independent and identically distributed as ε, where $E(\varepsilon) = 0$ and $E(\varepsilon^2) = \sigma^2 < \infty$; moreover, the tail probabilities satisfy $P(|\varepsilon| > x) \le K \exp(-Cx^2)$ for $x > 0$ and some constants C and K.

(C10) There exists a positive constant M such that
$|x_{ik}| \le M, i = 1, \cdots, n; k = 1, \cdots, p$.

(C7) is the sparse Riesz condition (SRC) formulated for the nonparameter AFT model (2.1), which controls the range of eigenvalues of the matrix Z. This condition was introduced to study the properties of Lasso for the linear regression model by [31]. (C8) assumes that the unimportant predictors are small in the L_2 sense, but do not need to be exactly zero. If $\eta_1 = 0$, (C8) becomes $f_j = 0$ for all $k \in A_0$. The problem of variable selection is equivalent to distinguishing nonzero functions from zero functions. (C9) assumes that the distribution of the error terms has sub-Gaussian tails. This condition holds when the error distribution is normal. (C10) assumes that all the predictors are uniformly bounded, which is satisfied in many practical situations.

In this subsection, we simply write $\hat{f}_j(X_j) = \sum_{k=1}^{M_n} \hat{\beta}_{jk} B_{jk}$ is GMCP estimator. Let $\boldsymbol{\beta}_*^o = \min\{\|\boldsymbol{\beta}_j^o\|_2, j \in A_0^c\}$ and set $\boldsymbol{\beta}_*^o = \infty$ if A_0^c is empty. Define

$$\hat{\boldsymbol{\beta}}^o = \arg\min_b \left\{ \frac{1}{2} \left\| \tilde{\boldsymbol{Y}} - \sum_{j=1}^{p} \tilde{\boldsymbol{B}}_j \boldsymbol{\beta}_j^o \right\|_2^2 ; \|\boldsymbol{\beta}_j^o\|_2 = 0, \forall j \notin A_0^c \right\} \tag{3.1}$$

and

$$f_j^o(X) = \sum_{k=1}^{M_n} \hat{\boldsymbol{\beta}}_{jk}^o \boldsymbol{B}_{jk}$$

This is the oracle least squares estimator. Of course, it is not a real estimator, since the oracle set is unknown.

We first consider the case where the 2-norm GMCP objective function is convex. This necessarily requires $c_{\min} > 0$ where $c_{\min}$ be the smallest eigenvalue of Σ, and recall $\Sigma = n^{-1}\boldsymbol{B}'\boldsymbol{B}$. As in [32], define the function

$$h(t,k) = \exp\left(-k\left(\sqrt{2t-1}-1\right)^2 \Big/ 4\right), \ t > 1, \ k = 1,2,\cdots \tag{3.2}$$

This function arises from an upper bound for the tail probabilities of the chi-square distributions given in Lemma A.2 in **Appendix**. This is derived from an exponential inequality for chi-square random variables of [33].

Theorem 3.1. *Suppose* $\varepsilon_1,\cdots,\varepsilon_n$ *are independent and identically distributed as* $N(0,\sigma^2)$ *and* $(C1)$-$(C10)$. *Then for any* (λ,γ) *satisfying* $\gamma > 1/c_{\min}$, $\boldsymbol{\beta}_*^o > \lambda\gamma$ *and* $n\lambda^2 > \sigma^2$, *we have*

$$P\left(\hat{\boldsymbol{\beta}}(\lambda,\gamma) \neq \hat{\boldsymbol{\beta}}^o\right) \leq \eta_{1n}(\lambda) + \eta_{2n}(\lambda)$$

and

$$P\left(\hat{f} \neq \hat{f}^o\right) \leq \eta_{1n}(\lambda) + \eta_{2n}(\lambda)$$

where $\eta_{1n}(\lambda) = (p-q)h\left(n\lambda^2/\sigma^2, M_n\right)$ and $\eta_{2n}(\lambda) = qh\left(c_1 n\left(\boldsymbol{\beta}_*^o - \gamma\lambda\right)^2/\sigma^2, M_n\right)$.

We give the proof of Theorem 3.1 in **Appendix**. It provides an upper bound on the probability that $\hat{f}$ is not equal to the oracle estimator in terms of the tail probability function h in (3.2). The key condition $\gamma > 1/c_{\min}$ ensures that the 2-norm GMCP criterion is strictly convex. Nonetheless, this result is a starting for a similar result in $p > n$ case. The following corollary is an immediate consequence of Theorem 3.1.

Corollary 1. *suppose that the condition of Therorm* 3.1 *are satisfied. Also suppose that* $\boldsymbol{\beta}_*^o \geq \gamma\lambda + a_n\tau_n$ *for* $a_n \to \infty$ *as* $n \to \infty$. *If* $\lambda \geq a_n\lambda_n$, *then*

$$P\left(\hat{\boldsymbol{\beta}}(\lambda,\gamma) \neq \hat{\boldsymbol{\beta}}^o\right) \to 0 \text{ as } n \to \infty$$

and

$$P\left(\hat{f} \neq \hat{f}^o\right) \to 0 \text{ as } n \to \infty$$

where

$$\lambda_n = \sigma\sqrt{2\log\left(\max\{p-q,1\}\right)/(nM_n)} \text{ and } \tau_n = \sigma\sqrt{2\log\left(\max\left(\{q,1\}\right)/\left(nc_1M_n\right)\right)}$$

By Corollary 1, the 2-norm GMCP estimator equals the oracle least squares estimator with probability converging to one. This implies it is group selection consistent. We now consider the high-dimensional case where $p > n$. Under

condition (C7), let $K_* = \bar{c} - 1/2$, $m_* = K_* q$ and $\xi = 1/\left(4c^* M_n\right)$. Define

$$\eta_{3n}(\lambda) = (p-q)^{m_*} \frac{e^{m_*}}{m_*^{m_*}} h\left(\xi n \lambda^2 \sigma^{-2} / M_n, m_* M_n\right)$$

Theorem 3.2. *suppose* $\varepsilon_1, \cdots, \varepsilon_n$ *are independent and identically distributed as* $N(0, \sigma^2)$ *and B satisfies the* $SRC\left(q^*, c_1, c_2\right)$ *in (C7) with* $q^* \geq (\bar{c} - 1/2)$, $m_* = K_* q$ *and* $\xi = 1/\left(4c^* M_n\right)$, *we have*

$$P\left(\hat{\beta}(\lambda, \gamma) \neq \hat{\boldsymbol{\beta}}^o\right) \leq \eta_{1n}(\lambda) + \eta_{2n}(\lambda) + \eta_{3n}(\lambda)$$

and

$$P\left(\hat{f} \neq \hat{f}^o\right) \leq \eta_{1n}(\lambda) + \eta_{2n}(\lambda)$$

where $\eta_{1n}(\lambda) = (p-q) h\left(n\lambda^2 / \sigma^2, M_n\right)$ and

$\eta_{2n}(\lambda) = q h\left(c_1 n \left(\boldsymbol{\beta}_*^o - \gamma\lambda\right)^2 / \sigma^2, M_n\right)$.

Corollary 2. *suppose that the condition of Therorm* 3.2 *are satisfied. Also suppose that* $\boldsymbol{\beta}_*^o \geq \gamma\lambda + a_n \tau_n$ *for* $a_n \to \infty$ *as* $n \to \infty$. *If* $\lambda \geq a_n \lambda_n^*$, *then*

$$P\left(\hat{\boldsymbol{\beta}}(\lambda, \gamma) \neq \hat{\boldsymbol{\beta}}^o\right) \to 0 \text{ as } n \to \infty$$

and

$$P\left(\hat{f} \neq \hat{f}^o\right) \to 0 \text{ as } n \to \infty$$

where $\lambda_n^* = 2\sigma\sqrt{2c_2 M_n \log(p-q)/n}$.

Theorem 3.2 and Corollary 2 provide sufficient conditions for the asymptotic oracle property of the global 2-norm GMCP estimator in the $p > n$ situations. Here we allow $p - \left|A_0^c\right| = \exp\left\{O\left(n/\left(c_2 M_n\right)\right)\right\}$. So p can be greater than n. The condition $n\lambda^2\xi > \sigma^2 M_n$ is stronger than the corresponding condition $n\lambda^2 > \sigma^2$ in Theorem 3.5 ([34]). The condition $\gamma \geq c_1^{-1}\sqrt{4+\bar{c}}$ ensures that the GMCP criterion is convex in any q-dimensional subspace. It is stronger than the minimal sufficient condition $\gamma > c_1^{-1}$ for convexity in q-dimensional subspaces. This is the price we need to pay in search for a lower-dimensional space that contains the true model.

4. Numerical Simulation

In this section, we conduct simulation studies to evaluate the performance of the GMCP and GLasso penalties in a high-dimensional NP-AFT-AR model with limited samples. We therefore focus on the comparisons of the group selection methods with only the BIC ([35]) selected tuning parameter (λ, M_n), is given:

$$\mathrm{BIC}(\lambda, M_n) = \log\left(\mathrm{RSS}_{\lambda, M_n}\right) + \log(n) \frac{df_{\lambda, M_n}}{n}$$

Where RSS is the sum of squared residuals, *df* is the number of selected variables given (λ, M_n). We choose M_n from the increasing sequence in Section 5, for any given value of M_n, We choose from a sequence of 100 values λ, from $0.01\lambda_{\max}$ to $\lambda_{\max}$, Where $\lambda_{\max} = \max_{1 \leq j \leq p} \left\|\tilde{\boldsymbol{B}}_j' \tilde{\boldsymbol{Y}}\right\|_2 / \sqrt{M_n \tilde{\boldsymbol{B}}_j'}$ is corresponding to

the covariate $X_j, j=1,\cdots,p$ with $n\times M_n$ "design" matrix. $\lambda_{\max}$ is the maximum penalty value, which compresses all estimated coefficients to zero.

We compute the empirical prediction mean square error (MSE) to reveal the estimation accuracy. Let $\hat{f}_j$ be the estimator of $f_j, j=1,\cdots,p$; and we define MSE as

$$\mathrm{MSE}_{f_j}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}_j\left(X_{ij}\right)-f_p\left(X_{ij}\right)\right|^2$$

Three scenarios are considered in the following, where some nonzero components are linear and the response variable is subject to various censoring rates. The sample size $n=400,200$ and a total of 100 simulation runs are used. The logarithm of censoring time C_i is generated from a uniform distribution $U\left(c_1,c_2\right), c_1>0; c_2>0$, where c_1 and c_2 are determined by a Monte-Carlo method to achieve the censoring rates of 35% and 40% respectively. For example, the censoring rate $cr=Pr\left(T>C\right)$ is approximated by $\widehat{cr}=\sum_{i=1}^{M}I\left(T_i>C_i\right)/M$ where T_i is drawn from the proposed model (2.1) and C_i is drawn from $U\left(-c_1,c_2\right), c_1>0; c_2>0$, M is the Monte-Carlo simulation runs used to compute *cr*. When we chose $c_1=0, c_2=4, \widehat{cr}\approx 40\%$, which is considered to be the desired censoring rate. To take account of the computational efficiency and accuracy, we use the cubic B-spline with five evenly distributed interior knots for all the functions $f_j, j=1,\cdots,p$, which gives the number of $3+1+5=9$ basis functions for each nonparametric component. Due to the identifiability constraint, $E\left\{f_j\left(X_j\right)\right\}=0$; the actual number of basis functions used is 8. This choice is made because our simulation studies indicated that using a larger number of knots does not improve the finite sample performance (results are not shown).

4.1. Scenario 1 (Covariates Are Independent)

In this scenario, we consider independent covariates and set the intercept $\eta_0=0$: The logarithm of failure times, $T_i, i=1,\cdots,n$, are generated from

$$T=\exp\Big(f_1\left(X_1\right)+f_2\left(X_2\right)+f_3\left(X_3\right)+f_4\left(X_4\right)+f_5\left(X_5\right)$$
$$+f_6\left(X_6\right)+\sum_{j=7}^{p}f_j\left(X_j\right)+\varepsilon\Big)$$

where

$$f_1\left(X_1\right)=2\left(\sin\left(0.25\pi X_1\right)\right)^3,\ f_2\left(X_2\right)=2\sin\left(2X_2\right),\ f_3\left(X_3\right)=X_3^2-\frac{3}{4},$$

$$f_4\left(X_4\right)=1.2X_4,\ f_5\left(X_5\right)=\exp\left(-X_5\right)-\frac{25}{12},$$

$$f_6\left(X_6\right)=\frac{1}{4}X_6^3,\ f_7\left(X_7\right)=\cdots=f_p\left(X_p\right)\equiv 0.$$

The predictors are sampled from the $N\left(0,1\right)$.

We set $p=500$ and consider the cases where $n=400,200$, respectively to

see the performance of our proposed methods as the sample size increases. The penalty parameters are selected using CV as described above.

The results for the the GMCP, GSCAD and GLasso methods are given in Table 1 and Table 2 based on 100 replications. The columns in Table 1 include the average number of variables (NV) being selected, model error (ER), percentage of occasions on which correct variables are included in the selected model (%IN) and percentage of occasions on which the exactly correct variables are selected (%CS) with standard error in parentheses. Table 2 summarizes the mean square errors for the six important functions $n^{-1}\sum_{i=1}^{n}\left|\hat{f}_j\left(X_{ji}\right)-f_j\left(X_{ji}\right)\right|^2, j=1,\cdots,6$ with standard error in parentheses.

Several observations can be obtained from Tables 1-4. The model that was selected by the GMCP and is better than the one selected by the GLasso in terms of model error, the percentage of occasions on which the true variables being selected and the mean square errors for the important coefficient functions. The GMCP includes the correct variables with high probability. When the sample size increases, the performance of both methods becomes better as expected. To examine the estimated nonparametric functions from Concave group Selection methods, we plot GMCP along with the true function components in Figure 1 and Figure 2. The estimated nonparametric coefficient functions are from GMCP method in one run when 100. From the graph, the estimators of the

Table 1. Simulation results. NV, number of selected variables; ER, model error; IN%, percentage of occasions on which the correct variables are included in the selected model; CS%, percentage of occasions on which exactly correct variables are selected, averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	Results for high dimension, $p=500$			
	NV	ER	IN%	CS%
	$n=400(\mathrm{CR}=35\%)$			
Group Lasso	6.0	0.0004	100.0	100.0
	(0.00)	(0.0003)	(0.00)	(0.00)
Group SCAD	6.0	0.0001	100.0	100.0
	(0.00)	(0.0001)	(0.00)	(0.00)
Group MCP	6.0	0.00009	100.0	100.0
	(0.00)	(0.00009)	(0.00)	(0.00)
	$n=200(\mathrm{CR}=35\%)$			
Group Lasso	8.0	0.0015	97.0	97.0
	(1.83)	(0.0018)	(0.171)	(0.171)
Group SCAD	6.1	0.0007	99.0	99.0
	(0.35)	(0.0014)	(0.100)	(0.100)
Group MCP	6.2	0.0005	99.0	98.0
	(1.62)	(0.0010)	(0.100)	(0.140)

Table 2. Simulation results. Mean Square errors for the important coefficient functions based on 100 replications. Enclosed in parentheses are the corresponding standard errors.

	$f_1(X_1)$	$f_2(X_2)$	$f_3(X_3)$	$f_4(X_4)$	$f_5(X_5)$	$f_6(X_6)$
			$n=400(\mathrm{CR}=35\%)$			
group Lasso	0.109	0.312	0.324	0.150	0.682	0.624
	(0.054)	(0.087)	(0.105)	(0.065)	(0.584)	(0.299)
group SCAD	0.076	0.227	0.262	0.114	0.683	0.519
	(0.049)	(0.086)	(0.089)	(0.068)	(0.626)	(0.319)
group MCP	0.073	0.226	0.258	0.111	0.644	0.516
	(0.048)	(0.085)	(0.094)	(0.064)	(0.540)	(0.319)
			$n=200(\mathrm{CR}=35\%)$			
group Lasso	0.259	0.803	1.399	0.415	0.864	0.711
	(0.101)	(0.268)	(0.558)	(0.168)	(0.610)	(0.313)
group SCAD	0.202	0.378	0.724	0.175	0.603	0.584
	(0.125)	(0.274)	(0.374)	(0.142)	(0.533)	(0.662)
group MCP	0.200	0.365	0.720	0.162	0.639	0.547
	(0.117)	(0.262)	(0.367)	(0.123)	(0.687)	(0.646)

Table 3. Simulation results. NV, number of selected variables; ER, model error; IN%, percentage of occasions on which the correct variables are included in the selected model; CS%, percentage of occasions on which exactly correct variables are selected, averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	Results for high dimension, $p=500$			
	NV	ER	IN%	CS%
		$n=400(\mathrm{CR}=40\%)$		
Group Lasso	6.6	0.0003	100.0	100.0
	(1.16)	(0.0003)	(0.0)	(0.0)
Group SCAD	6.1	0.0001	100.0	100.0
	(0.29)	(0.0001)	(0.0)	(0.0)
Group MCP	6.1	0.00009	100.0	100.0
	(0.37)	(0.00009)	(0.0)	(0.0)
		$n=200(\mathrm{CR}=40\%)$		
Group Lasso	8.4	0.0016	96.0	95.0
	(2.31)	(0.0031)	(0.196)	(0.219)
Group SCAD	6.1	0.0010	97.0	95.0
	(2.04)	(0.0031)	(0.171)	(0.219)
Group MCP	6.2	0.0007	97.0	96.0
	(2.23)	(0.0027)	(0.171)	(0.196)

Table 4. Simulation results. Mean Square errors for the important coefficient functions based on 100 replications. Enclosed in parentheses are the corresponding standard errors.

	$f_1(X_1)$	$f_2(X_2)$	$f_3(X_3)$	$f_4(X_4)$	$f_5(X_5)$	$f_6(X_6)$
			$n=400(\mathrm{CR}=40\%)$			
group Lasso	0.111	0.247	0.176	0.132	0.750	0.666
	(0.055)	(0.102)	(0.140)	(0.074)	(0.702)	(0.313)
group SCAD	0.077	0.202	0.110	0.109	0.681	0.563
	(0.051)	(0.100)	(0.059)	(0.087)	(0.592)	(0.357)
group MCP	0.074	0.202	0.113	0.107	0.655	0.555
	(0.050)	(0.100)	(0.074)	(0.087)	(0.552)	(0.345)
			$n=200(\mathrm{CR}=40\%)$			
group Lasso	0.392	0.746	0.777	0.543	1.304	0.439
	(0.144)	(0.343)	(0.456)	(0.272)	(0.834)	(0.197)
group SCAD	0.133	0.441	0.271	0.217	0.894	0.297
	(0.159)	(0.357)	(0.369)	(0.283)	(0.769)	(0.244)
group MCP	0.122	0.428	0.286	0.197	0.916	0.289
	(0.148)	(0.346)	(0.462)	(0.240)	(1.125)	(0.243)

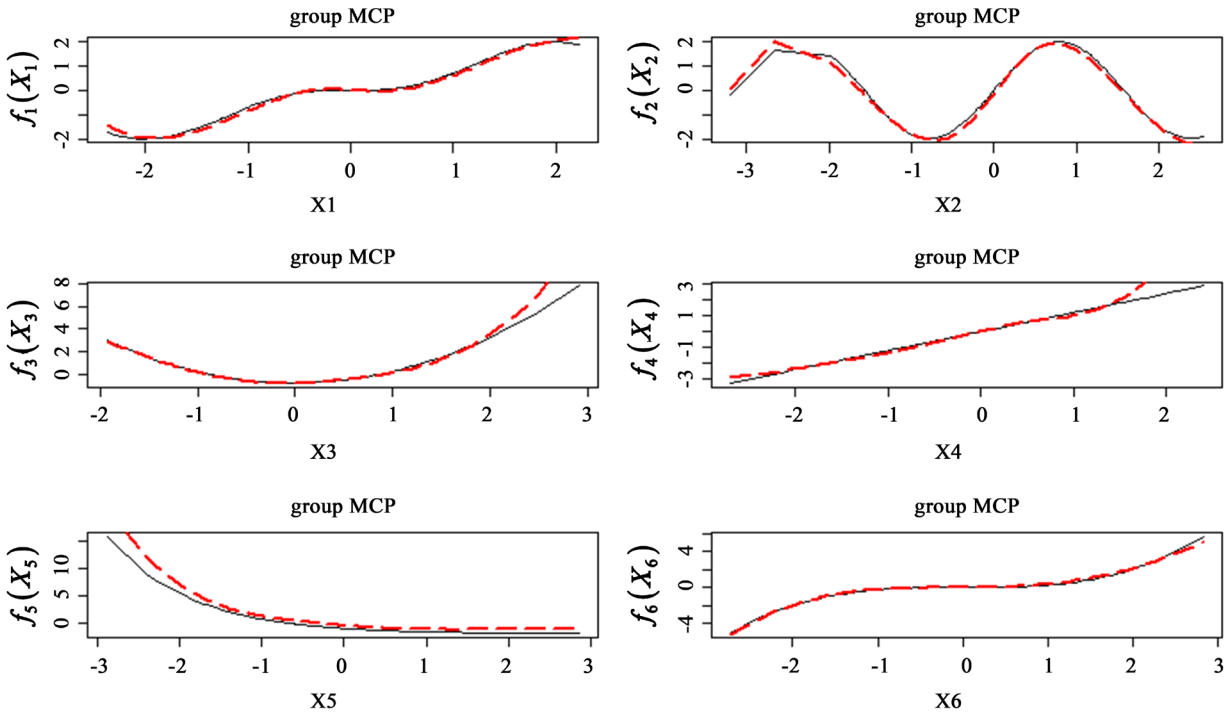

Figure 1. $n=200$, the solid black line is the real function, the dotted red line is the GMCP estimation, CR = 35%.

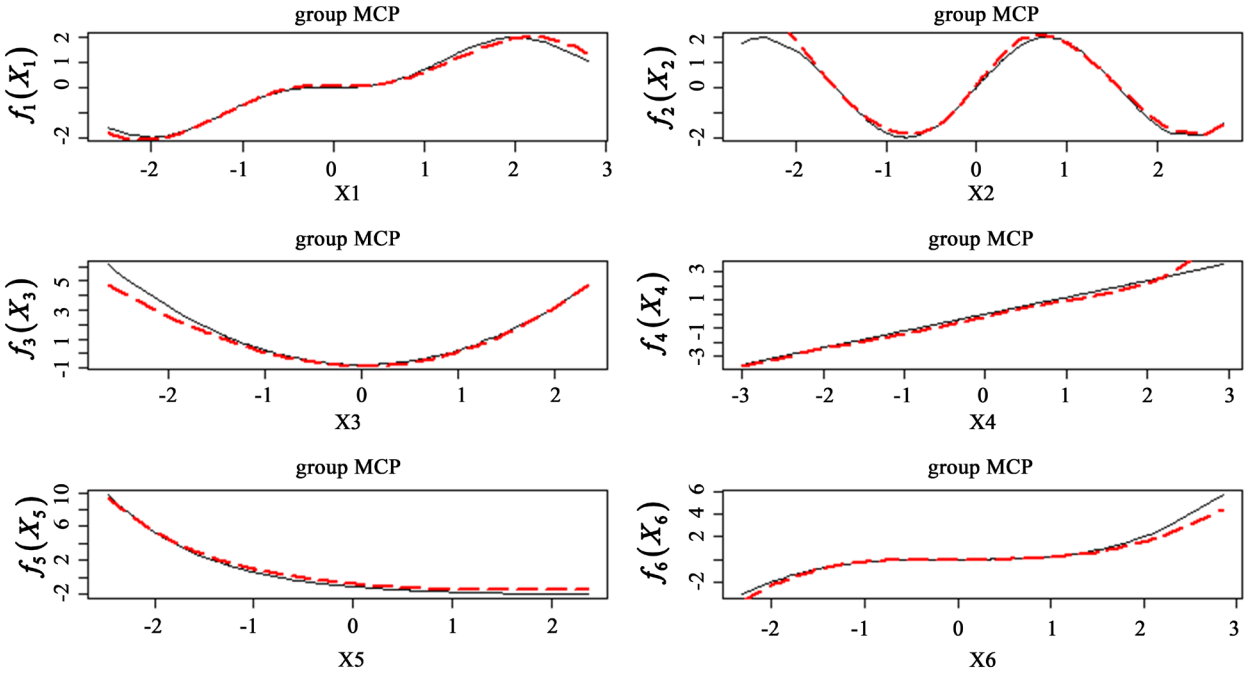

Figure 2. $n = 200$, the solid black line is the real function, the dotted red line is the GMCP estimation, CR = 40%.

nonparameter $f_j(X_j), j = 1, \cdots, 6$, fit the true functions well, which are consistent with the mean square errors for the functions reported in Table 2, Table 4.

4.2. Scenario 2 (Covariates Are Correlated)

In this scenario, we consider correlated covariates and set the intercept $\eta_0 = 0$: The logarithm of failure times, $T_i, i = 1, \cdots, n$, are generated from?

$$T = \exp\left(f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + f_5(X_5) + f_6(X_6) + \sum_{j=7}^{p} f_j(X_j) + \varepsilon \right)$$

$$f_1(X_1) = 1.2X_1,\ f_2(X_2) = 2\sin(2X_2),\ f_3(X_3) = \left(X_3^2 - \frac{3}{4}\right),$$

$$f_4(X_4) = \exp(-X_5) - \frac{25}{12},\ f_5(X_5) = \sin(0.5\pi X_5),$$

$$f_6(X_6) = 2\left(\sin(0.25\pi X_6)\right)^3,\ f_7(X_7) = \cdots = f_p(X_p) \equiv 0.$$

where the covariates $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ are generated from $X_p = (W_p + 0.5U)/1.5$ where $W_1, \cdots, W_p$ and U are i.i.d. $N(0,1)$. This provides a design with a correlation coefficient of 0.5 between all of the covariates.

The simulation study results are reported in Tables 5-8. The conclusions for Scenario 2 are very similar to those for Scenario 1. When the censoring rate increases, the estimation and selection performance decreases for all methods. The results in Table 6, Table 8 show that the GMCP estimator is more accurate than the GLasso estimator for both the individual component functions and the full

Table 5. Simulation results. NV, number of selected variables; ER, model error; IN%, percentage of occasions on which the correct variables are included in the selected model; CS%, percentage of occasions on which exactly correct variables are selected, averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	Results for high dimension, $p = 500$			
	NV	ER	IN%	CS%
	$n = 400(\mathrm{CR} = 35\%)$			
Group Lasso	9.0	0.0024	100.0	99.0
	(2.05)	(0.0011)	(0.00)	(0.10)
Group SCAD	7.8	0.0015	100.0	100.0
	(1.58)	(0.0009)	(0.00)	(0.00)
Group MCP	8.3	0.0013	100.0	100.0
	(2.07)	(0.0007)	(0.00)	(0.00)
	$n = 200(\mathrm{CR} = 35\%)$			
Group Lasso	12.9	0.0043	86.5	86.0
	(3.47)	(0.0033)	(0.343)	(0.347)
Group SCAD	8.3	0.0033	93.5	92.0
	(1.63)	(0.0031)	(0.247)	(0.271)
Group MCP	8.6	0.0024	93.5	93.5
	(1.72)	(0.0024)	(0.247)	(0.247)

Table 6. Simulation results. Mean Square errors for the important coefficient functions based on 100 replications. Enclosed in parentheses are the corresponding standard errors.

	$f_1(X_1)$	$f_2(X_2)$	$f_3(X_3)$	$f_4(X_4)$	$f_5(X_5)$	$f_6(X_6)$
	$n = 400(\mathrm{CR} = 35\%)$					
group Lasso	0.149	0.173	0.224	0.998	0.073	0.124
	(0.059)	(0.082)	(0.218)	(0.166)	(0.026)	(0.069)
group SCAD	0.086	0.117	0.191	0.757	0.032	0.114
	(0.047)	(0.082)	(0.527)	(0.139)	(0.017)	(0.122)
group MCP	0.070	0.133	0.177	0.715	0.028	0.113
	(0.042)	(0.089)	(0.479)	(0.132)	(0.013)	(0.124)
	$n = 200(\mathrm{CR} = 35\%)$					
group Lasso	0.404	0.597	0.586	1.406	0.233	0.256
	(0.149)	(0.264)	(0.143)	(0.304)	(0.109)	(0.065)
group SCAD	0.221	0.365	0.441	0.956	0.119	0.206
	(0.162)	(0.503)	(0.326)	(0.338)	(0.125)	(0.120)
group MCP	0.175	0.374	0.363	0.849	0.082	0.177
	(0.138)	(0.496)	(0.337)	(0.275)	(0.1044)	(0.106)

Table 7. Simulation results. NV, number of selected variables; ER, model error; IN%, percentage of occasions on which the correct variables are included in the selected model; CS%, percentage of occasions on which exactly correct variables are selected, averaged over 100 replications. Enclosed in parentheses are the corresponding standard errors.

	Results for high dimension, $p = 500$			
	NV	ER	IN%	CS%
	$n = 400(\mathrm{CR} = 40\%)$			
Group Lasso	9.0	0.0019	100.0	98.0
	(1.71)	(0.0010)	(0.00)	(0.14)
Group SCAD	7.7	0.0012	100.0	98.0
	(1.18)	(0.0007)	(0.00)	(0.14)
Group MCP	8.4	0.0009	100.0	98.0
	(1.54)	(0.00055)	(0.00)	(0.14)
	$n = 200(\mathrm{CR} = 40\%)$			
Group Lasso	13.0	0.0044	89.0	85.0
	(3.57)	(0.0031)	(0.313)	(0.357)
Group SCAD	8.2	0.0033	95.0	85.0
	(1.66)	(0.0030)	(0.218)	(0.357)
Group MCP	8.4	0.0024	95.0	86.0
	(1.50)	(0.0023)	(0.218)	(0.347)

Table 8. Simulation results. Mean Square errors for the important coefficient functions based on 100 replications. Enclosed in parentheses are the corresponding standard errors.

	$f_1(X_1)$	$f_2(X_2)$	$f_3(X_3)$	$f_4(X_4)$	$f_5(X_5)$	$f_6(X_6)$
	$n = 400(\mathrm{CR} = 40\%)$					
group Lasso	0.104	0.192	0.157	0.781	0.071	0.152
	(0.071)	(0.103)	(0.0844)	(0.191)	(0.036)	(0.058)
group SCAD	0.070	0.103	0.132	0.737	0.037	0.100
	(0.037)	(0.076)	(0.160)	(0.269)	(0.030)	(0.065)
group MCP	0.065	0.099	0.127	0.740	0.034	0.081
	(0.037)	(0.074)	(0.171)	(0.298)	(0.028)	(0.059)
	$n = 200(\mathrm{CR} = 40\%)$					
group Lasso	0.414	0.578	0.495	1.466	0.213	0.231
	(0.134)	(0.232)	(0.132)	(0.332)	(0.087)	(0.069)
group SCAD	0.224	0.262	0.389	1.213	0.115	0.211
	(0.176)	(0.246)	(0.301)	(0.489)	(0.109)	(0.172)
group MCP	0.176	0.204	0.351	1.141	0.084	0.207
	(0.148)	(0.190)	(0.365)	(0.527)	(0.099)	(0.166)

model, since the MSE under the GMCP approach is always smaller than that under the GLasso approach. The results in Table 5, Table 7 show that the GMCP method conducts component selection more precisely than the GLasso method, while the GLasso method chooses many zero component functions as nonzero functions. To examine the estimated nonparametric functions from the GMCP, we plot them along with the true function components in Figure 3, Figure 4. The estimated functions are from the GMCP method in one run when $n = 200$. The estimation and selection accuracy decrease when covariates are correlated, we can still see that the estimated curves under the GMCP method are close to the true curves compared with the estimated curves under the GLasso method.

5. Application in NA-AFT-Model

In this section, we will use Shedden 2008 (for short) to conduct an empirical analysis of part of the collected lung adenocarcinoma data to illustrate the proposed method. For more information, see [36]. Retrospective data of 442 lung adenocarcinoma patients were collected at multiple locations, including their survival time, some other clinical and demographic data, and the expression level of the 22,283 gene from the following genes: tumor samples. However, most samples have small changes. Therefore, in our application, we randomly select 321 samples, the first 500, 1000 genes. Therefore, $n = 321, p = 1000$, and the survival rate is 35.8%.

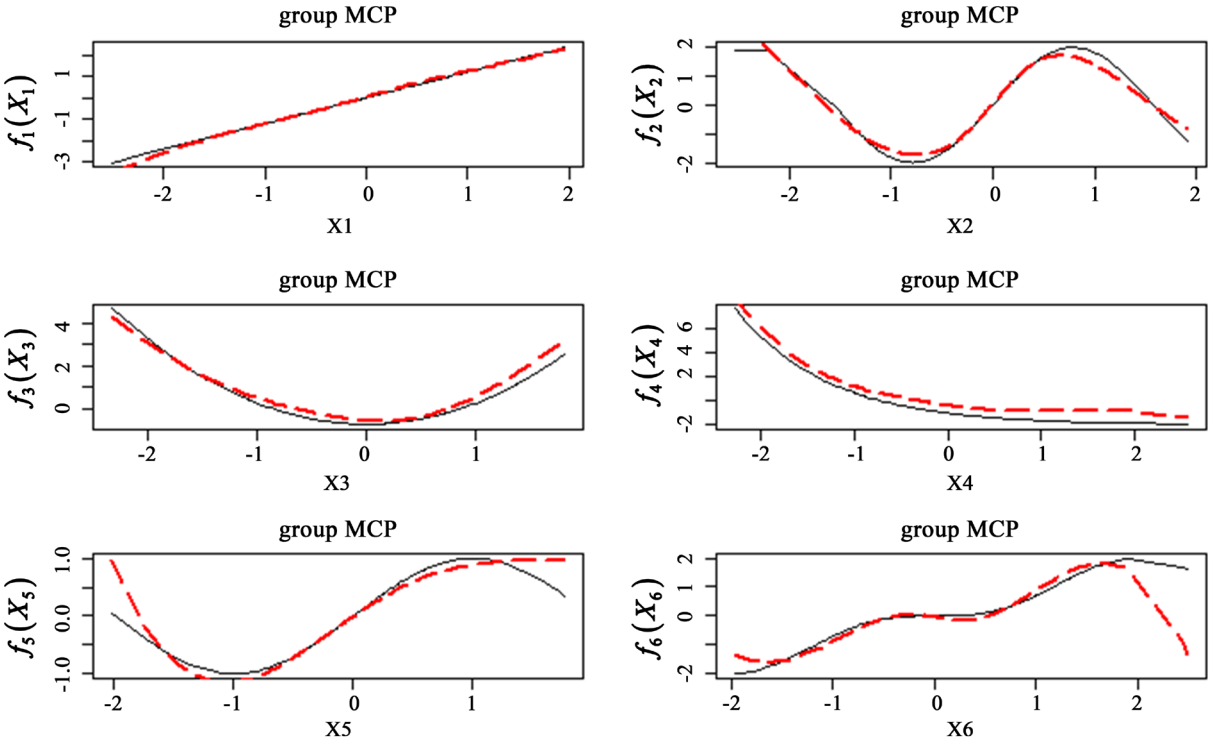

Figure 3. $n = 200$, the solid black line is the real function, the dotted red line is the group MCP estimation, the dotted blue line is the group SCAD estimation, and the black line is the group lasso estimation, CR = 35%.

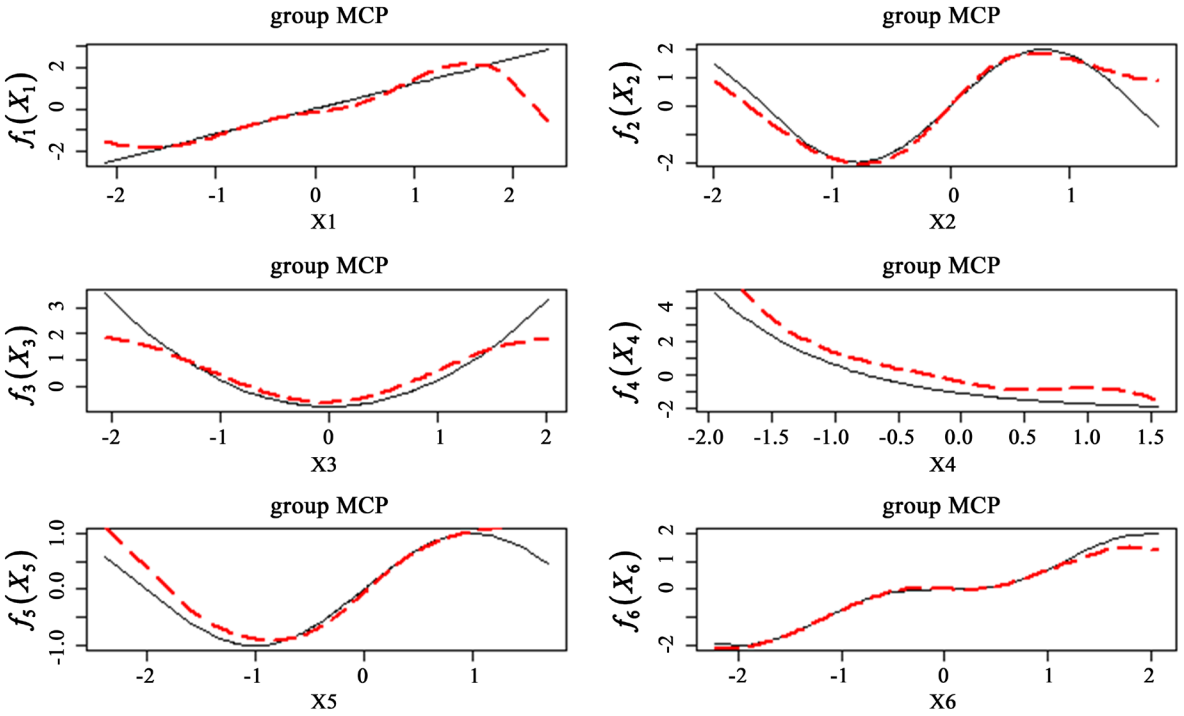

Figure 4. $n = 200$, the solid black line is the real function, the dotted red line is the group MCP estimation, the dotted blue line is the group SCAD estimation, and the black line is the group lasso estimation, CR = 40%.

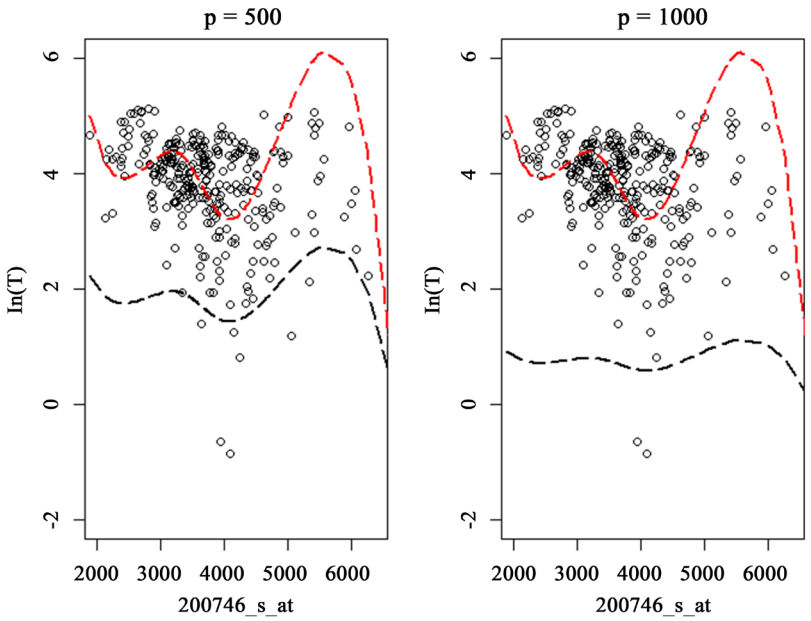

Figure 5. The estimation function graph based on GLasso and GMCP approximates the corresponding 200746_*s_at* by using the same covariate, where the red dotted line is the GMCP estimate and the gray dotted line is the GLasso estimate.

Here, we are interested in the effect of tumor gene expression levels on the survival time of lung adenocarcinoma patients. Since the linear assumption is always latent in high dimensions, the proposed method may be more suitable for analyzing feature selection problems considering nonlinear effects. In our analy-

sis, we set the spline base $M_n = 5$ for each gene. The proposed method selects 1 gene locus under GMCP (ie 200746_*s_at*). However, when $p = 500,1000$, the method under GLasso penalized regression alone selected the 6, 10 gene.

From **Figure 5**, we find that the larger the dimension, the worse the GLasso method estimation, but it has little effect on the GMCP estimation. Therefore, the verification of the actual data shows that the GMCP penalty is better than the GLasso penalty, and the accuracy is higher, and the calculation cost of the two is the same. Under the same conditions, the GMCP method is more suitable than the GLasso.

6. Concluding Remarks

In this paper, we study the weighted least squares estimation and selection attributes of GMCP in the NP-AFT-AR model with high-dimensional data. For the GMCP method, our simulation results show that GLasso tends to select some unimportant variables. In contrast, GMCP has progressive predictability, which shows that it also has selection consistency.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Cox, D.R. (1972) Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society*: *Series B*, **2**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[2] Huang, J., Ma, S. and Xie, H.L. (2006) Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. *Biometrics*, **62**, 813-820. https://doi.org/10.1111/j.1541-0420.2006.00562.x

[3] Datta, S., Jennifer, L.R. and Datta, S. (2007) Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO. *Biometrics*, **63**, 259-271. https://doi.org/10.1111/j.1541-0420.2006.00660.x

[4] Huang, J., Ma, S. and Xie, H.L. (2007) Least Absolute Deviations Estimation for the Accelerated Failure Time Model. *Statistica Sinica*, **17**, 1533-1548.

[5] Leng, C. and Ma, S. (2007) Accelerated Failure Time Models with Nonlinear Covariates Effects. *Australian and New Zealand Journal of Statistics*, **49**, 155-172. https://doi.org/10.1111/j.1467-842X.2007.00470.x

[6] Schmid, M. and Hothorn, T. (2008) Flexible Boosting of Accelerated Failure Time Models. *BMC Bioinformatics*, **9**, Article No. 269. https://doi.org/10.1186/1471-2105-9-269

[7] David, E. and Li, Y. (2009) Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies. *Statistical Applications in Genetics and Molecular Biology*, **8**, 1-122. https://doi.org/10.2202/1544-6115.1423

[8] Cai, T., Huang, J. and Tian, L. (2009) Regularized Estimator for the Accelerated Failure Time Model. *Biometric*, **65**, 394-404. https://doi.org/10.1111/j.1541-0420.2008.01074.x

[9] Johnson, B.A.(2009) Rank-Based Estimation in the ℓ_1-Regularized Partly Linear Model for Censored Outcomes with Application to Integrated Analyses of Clinical Predictors and Gene Expression data. *Biostatistics*, **10**, 3659-466.

[10] Huang, J. and Ma, S. (2010) Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, **16**, 176-195. https://doi.org/10.1007/s10985-009-9144-2

[11] Wang, Z. and Wang, C.Y. (2010) Buckley-James Boosting for Survival Analysis with High-Dimensional Biomarker Data. *Statistical Applications in Genetics and Molecular Biology*, **9**, Article No. 24. https://doi.org/10.2202/1544-6115.1550

[12] Liu, F., Dunson, D. and Zou, F. (2011) High-Dimensional Variable Selection in Meta-Analysis for Censored Data. *Biostatistics*, **67**, 504-512. https://doi.org/10.1111/j.1541-0420.2010.01466.x

[13] Ma, S. and Du, P. (2012) Variable Selection in Partly Linear Regression Model with Diverging Dimensions for Right Censored Data. *Statistica Sinica*, **22**, 1003-1020. https://doi.org/10.5705/ss.2010.267

[14] Schumaker, L. (1981) Spline Functions: Basic Theory. Cambridge University Press, Cambridge.

[15] Gu, C. (2004) Model Diagnostics for Smoothing Spline ANOVA Models. *Canadian Journal of Statistic*, **32**, 347-358. https://doi.org/10.2307/3316020

[16] Hu, J.W. and Chai, H. (2013) Adjusted Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates. *Journal of Multivariate Analysis*, **122**, 96-114. https://doi.org/10.1016/j.jmva.2013.07.011

[17] Chai, H., Liang, Y. and Liu, X.Y. (2015) The $L_{1/2}$ Regularization Approach for Survival Analysis in the Accelerated Failure Time Model. *Computers in Biology and Medicine*, **64**, 283-290. https://doi.org/10.1016/j.compbiomed.2014.09.002

[18] Xia, X.C., Jiang, B.Y., Li, J.L. and Zhang, W.Y. (2016) Low-Dimensional Confounder Adjustment and High-Dimensional Penalized Estimation for Survival Analysis. *Lifetime Data Analysis*, **22**, 547-569. https://doi.org/10.1007/s10985-015-9350-z

[19] Khan, M.H.R. and Shaw, J.E.H. (2016) Variable Selection for Survival Data with a Class of Adaptive Elastic Net Techniques. *Statistics and Computing*, **26**, 725-741. https://doi.org/10.1007/s11222-015-9555-8

[20] Yang, Y.C., Fasching, P.A. and Tresp, V. (2017) Modeling Progression Free Survival in Breast Cancer with Tensorized Recurrent Neural Networks and Accelerated Failure Time Models. *Proceedings of Machine Learning Research*, **68**, 164-176.

[21] Yue, M., Li, J.L. and Ma, S.G. (2018) Sparse Boosting for High-Dimensional Survival Data with Varying Coefficients. *Statistics in Medicine*, **37**, 789-800. https://doi.org/10.1002/sim.7544

[22] Khan, M.H.R. and Shaw, J.E.H. (2019) Variable Selection for Accelerated Lifetime Models with Synthesized Estimation Techniques. *Statistical Methods in Medical Research*, **28**, 937-952. https://doi.org/10.1177%2F0962280217739522

[23] Zhang, C.H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[24] Stute, W. (1993) Almost Sure Representations of the Product-Limit Estimator for Truncated Datay. *The Annals of Statistics*, **21**, 146-156. https://doi.org/10.1214/aos/1176349019

[25] Zhou, M. (1992) M-Estimation in Censored Linear Models. *Biometrika*, **79**, 837-841.

[26] Satten, G.A. and Datta, S. (2001) The Kaplan-Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average. *The American Statistician* **55**, 207-210.

https://doi.org/10.1198/000313001317098185

[27] Fu, W.J. (1998) Penalized Regressions: The Bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, **7**, 397-416. https://doi.org/10.1080/10618600.1998.10474784

[28] Breheny, P. and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Statistics*, **5**, 232-253. https://doi.org/10.1214/10-AOAS388

[29] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*: *Series B*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[30] Stute, W. (1999) Nonlinear Censored Regression. *Statistica Sinica*, **9**, 1089-1102.

[31] Zhang, C.H. and Huang, J. (2008) The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *The Annals of Statistics*, **36**, 1567-1594. https://doi.org/10.1214/07-AOS520

[32] Huang, J., Breheny, P. and Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, **27**, 481-499. https://doi.org/10.1214/12-STS392

[33] Laurent, B. and Laurent, P. (2000) Adaptive Estimation of a Quadratic Functional by Model Selection. *The Annals of Statistics*, **28**, 1302-1338. https://doi.org/10.1214/aos/1015957395

[34] Yang, G.R., Huang, J. and Zhou, Y. (2014) Concave Group Methods for Variable Selection and Estimation in High-Dimensional Varying Coefficient Models. *Science china Mathematics*, **57**, 2073-2090. https://doi.org/10.1007/s11425-014-4842-y

[35] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[36] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma (2008) Gene Expression-Based Survival Prediction in Lung Adenocarcinoma: A Multi-Site, Blinded Validation Study. *Nature Medicine*, **14**, 822-827. https://doi.org/10.1038/nm.1790

Appendix Proof

Lemma 1. *Let* χ_k^2 *be a random variable with chi-square distribution with k degrees of freedom. For* $t>1$, $P\left(\chi_k^2 \geq kt\right) \leq h(t,k)$, *where* $h(t,k)$ *is defined in* (3.2).

This lemma is a restatement of the exponential inequality for chi-square distributions of [33].

proof of Theorem 3.1. Since $\hat{\beta}^o$ is the oracle least squares estimator, we have $\hat{\boldsymbol{\beta}}_j^o, j \in A_0$ and

$$-\tilde{\boldsymbol{B}}_j'\left(\tilde{Y}-\tilde{\boldsymbol{B}}\hat{\boldsymbol{\beta}}^o\right)/n=0, \ \forall j \in A_0^c$$

If $\left\|\hat{\boldsymbol{\beta}}_j^o\right\|_2 / \sqrt{M_n} \geq \lambda\gamma$, then by the definition of the MCP, $\dot{\rho}\left(\left\|\hat{\boldsymbol{\beta}}_j^o\right\|_2; \sqrt{M_n}\lambda, \gamma\right)=0$. Since $c_{\min} > 1/\gamma$, the criterion (2.8) is strictly convex. By the Karush-Kuhn-Tucker (KKT) conditions, the equality $\hat{\boldsymbol{\beta}}(\lambda,\gamma)=\hat{\boldsymbol{\beta}}^o$ holds in the intersection of the events

$$\Omega_1(\lambda)=\left\{\max_{j\in A_0}\left\|n^{-1}\tilde{\boldsymbol{B}}_j'\left(\tilde{Y}-\tilde{\boldsymbol{B}}\right)\hat{\boldsymbol{\beta}}^o\right\|_2 / \sqrt{M_n} \leq \lambda\right\}$$

and

$$\Omega_2(\lambda)=\left\{\min_{j\in A_0^c}\left\|\hat{\boldsymbol{\beta}}_j^o\right\|_2 \geq \gamma\lambda\right\}$$

We first bound $1-P\left(\Omega_1(\lambda)\right)$. Let $\hat{\boldsymbol{\beta}}_{A_0^c}=\left(\hat{\boldsymbol{\beta}}_j, j\in A_0^c\right)'$. By (A.1) [34] and using $\tilde{Y}=\tilde{\boldsymbol{B}}_{A_0^c}\boldsymbol{\beta}_{A_0^c}^o+\varepsilon$

$$\hat{\boldsymbol{\beta}}_{A_0^c}^o=\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}_{A_0^c}'\tilde{Y}/n=\boldsymbol{\beta}_{A_0^c}^o+\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}_{A_0^c}'\varepsilon/n$$

It follows that $n^{-1}\tilde{\boldsymbol{B}}_j\left(\tilde{Y}-\tilde{\boldsymbol{B}}\hat{\boldsymbol{\beta}}^o\right)=n^{-1}\tilde{\boldsymbol{B}}_k'\left(I_n-P_{A_0^c}\right)\varepsilon$, where $P_{A_0^c}=n^{-1}\tilde{\boldsymbol{B}}_{A_0^c}\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}_{A_0^c}'$, Because $\tilde{\boldsymbol{B}}_j'\tilde{\boldsymbol{B}}_j=I_{M_n}$, $\left\|\tilde{\boldsymbol{B}}_j\left(I_n-P_{A_0^c}\right)\varepsilon\right\|_2^2/\sigma^2$ is distributed as a χ^2 distribution with M_n degrees of freedom. We have, for $n\lambda^2/\sigma^2 \geq 1$

$$\begin{aligned}
1-P\left(\Omega_1(\lambda)\right) &= P\left(\max_{j\in A_0}\left\|n^{1/2}\tilde{\boldsymbol{B}}_j'\left(I_n-P_{A_0^c}\right)\varepsilon\right\|_2^2/\left(M_n\sigma^2\right) > n\lambda^2/\sigma^2\right) \\
&\leq \sum_{j\in A_0} P\left(\left\|n^{-1/2}\tilde{\boldsymbol{B}}_j'\left(I_n-P_{A_0^c}\right)\varepsilon\right\|_2^2/\left(\sqrt{M_n}\sigma^2\right) > M_n n\lambda^2/\sigma^2\right) \\
&\leq \sum_{j\in A_0} h\left(n\lambda^2/\sigma^2, M_n\right) \qquad (6.1)\\
&\leq (p-q)h\left(n\lambda^2/\sigma^2, M_n\right) \\
&= \eta_{1n}(\lambda)
\end{aligned}$$

where we used lemma 1 in the third line. Now consider $\Omega_2(\lambda)$, Recall $\boldsymbol{\beta}_*^o=\min_{j\in A_0^c}\left\|\boldsymbol{\beta}_j^o\right\|_2$. If $\left\|\hat{\boldsymbol{\beta}}_j^o-\boldsymbol{\beta}_j^o\right\|_2/\sqrt{M_n} \leq \boldsymbol{\beta}_*^o-\gamma\lambda$ for all $j\in A_0^c$, then $\min_{j\in A_0^c}\left\|\hat{\boldsymbol{\beta}}_j^o\right\|_2/\sqrt{M_n} \geq \gamma\lambda$. This implies

$$1-P\left(\Omega_2(\lambda)\right)\le P\left(\max_{j\in A_0^c}\left\|\hat{\boldsymbol{\beta}}_j^o-\boldsymbol{\beta}_j^o\right\|_2\Big/\sqrt{M_n}>\boldsymbol{\beta}_*^o-\gamma\lambda\right)$$

Let $\tilde{\boldsymbol{B}}_j$ be a $M_n\times M_nq$ matrix with a $M_n\times M_n$ identity matrix I_{M_n} in the pth block and 0's elsewhere. Then $n^{1/2}\left(\hat{\boldsymbol{\beta}}_j^o-\boldsymbol{\beta}_j^o\right)=n^{-1/2}\tilde{\boldsymbol{B}}_j\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon$. Note that

$$\left\|n^{-1/2}\tilde{\boldsymbol{B}}_j\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon\right\|_2\le\left\|\tilde{\boldsymbol{B}}_j\right\|_2\left\|\Sigma_{A_0^c}^{-1/2}\right\|_2\left\|n^{-1/2}\Sigma_{A_0^c}^{-1/2}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon\right\|_2\le c_1^{-1/2}\left\|n^{-1/2}\Sigma_{A_0^c}^{-1/2}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon\right\|_2$$

and $\left\|n^{-1/2}\Sigma_{A_0^c}^{-1/2}\tilde{B}'_{A_0^c}\varepsilon\right\|_2^2\Big/\sigma^2$ id distributed as a χ distribution with q degrees of freedom. Therefore, similar to $\eta_{1n}(\lambda)$, we have, for $c_1n\left(\boldsymbol{\beta}_*^o-\gamma\lambda\right)\Big/\sigma^2>1$,

$$\begin{aligned}1-P\left(\Omega_2(\lambda)\right)&=P\left(\max_{j\in A_0^c}n^{-1/2}\left\|\tilde{\boldsymbol{B}}_j\Sigma_{A_0^c}^{-1}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon\right\|_2\Big/\sqrt{M_n}>\sqrt{n}\left(\boldsymbol{\beta}_*^o-\gamma\lambda\right)\right)\\&\le P\left(\max_{j\in A_0^c}\left\|n^{-1/2}\Sigma_{A_0^c}^{-1/2}\tilde{\boldsymbol{B}}'_{A_0^c}\varepsilon\right\|_2^2\Big/\left(M_n\sigma^2\right)>c_1n\left(\boldsymbol{\beta}_*^o-\gamma\lambda\right)^2\Big/\sigma^2\right)\\&\le qh\left(c_1n\left(\boldsymbol{\beta}_*^o-\gamma\lambda\right)^2\Big/\sigma^2,M_n\right)\\&=\eta_{2n}(\lambda)\end{aligned}\tag{6.2}$$

Combining $\eta_{1n}(\lambda)$ and $\eta_{2n}(\lambda)$, we have

$$P\left(\hat{\boldsymbol{\beta}}(\lambda,\gamma)\ne\hat{\boldsymbol{\beta}}^o\right)\le1-P\left(\Omega_1(\lambda)\right)+1-P\left(\Omega_2(\lambda)\right)\le\eta_{1n}(\lambda)+\eta_{2n}(\lambda)$$

Since $\hat{f}(x)=\boldsymbol{B}\hat{\boldsymbol{\beta}}$, we can obtain $P\left(\hat{f}\ne\hat{f}^o\right)\le\eta_{1n}(\lambda)+\eta_{2n}(\lambda)$. This completes the proof.

For any $Q\subset\{1,\cdots,p\}$ and $m\ge1$, define

$$\zeta(v;m,B)=\max\left\{\frac{\left\|\left(P_A-P_B\right)v\right\|_2}{(mn)^{1/2}}:Q\subseteq A\subseteq\{1,\cdots,p\},d_A=m+d_B\right\}$$

Lemma 2. *Suppose* $\xi n\lambda^2>\sigma^2M_n$. *We have*

$$P\left(2\sqrt{c_2M_n}\zeta\left(\tilde{Y};m,A_0^c\right)>\lambda\right)\le(p-q)^m\frac{e^m}{m^m}\exp\left(-m\xi n\lambda^2/16\right)$$

proof. For any $A\supseteq A_0^c$. We have $\left(P_a-P_{A_0^c}\right)\tilde{\boldsymbol{B}}_{A_0^c}\hat{\boldsymbol{\beta}}_{A_0^c}=0$. Thus

$$\left(P_A-P_{A_0^c}\right)\tilde{Y}=\left(P_A-P_{A_0^c}\right)\left(\tilde{\boldsymbol{B}}_{A_0^c}\boldsymbol{\beta}_{A_0^c}+\varepsilon\right)=\left(P_A-P_{A_0^c}\right)\varepsilon$$

Therefore,

$$P\left(2\sqrt{c_2M_n}\zeta\left(\tilde{Y};m,A_0^c\right)>\lambda\right)=P\left(\max_{A\supseteq A_0^c,|A|=m+q}\left\|\left(P_A-P_{A_0^c}\right)\varepsilon\right\|_2^2\Big/\sigma^2>\xi mn\lambda^2\right)$$

Since P_A-P_B is a projection matrix, $\left\|\left(P_A-P_{A_0^c}\right)\varepsilon\right\|_2^2\Big/\sigma^2\sim\chi_{m_A}^2$, where $m_A=\sum_{j\in A-A_0^c,A\supseteq A_0^c}M_n\le mM_n$. Since there $\binom{p-q}{m}$ are ways to choose A from $\{1,\cdots,p\}$, we have

$$P\left(2\sqrt{c_2M_n}\zeta\left(\tilde{Y};m,A_0^c\right)>\lambda\right)\le\binom{p-q}{m}P\left(\chi_{mM_n}^2>\xi mn\lambda^2\right).$$

This and Lemma A.2 imply that

$$P\left(2\sqrt{c_2 M_n}\,\zeta\left(\tilde{Y};m,A_0^c\right)>\lambda\right)\le\binom{p-q}{m}h\left(\xi n\lambda^2/M_n,mM_n\right)\le(p-q)^m\frac{e^m}{m^m}h\left(\xi n\lambda^2/M_n,mM_n\right) \tag{6.3}$$

here we used the inequality $\binom{p-q}{m}\le(p-q)^m\frac{e^m}{m^m}$, this completes the proof.

Define I as any set that satisfies

$$A_0^c\cup\left\{j:\left\|\hat{\boldsymbol{\beta}}_j\right\|_2\neq 0\right\}\subseteq I$$

$$\subseteq A_0^c\cup\left\{j:n^{-1}\tilde{\boldsymbol{B}}'\left(\tilde{Y}-\tilde{\boldsymbol{B}}\hat{\boldsymbol{\beta}}\right)=\dot{\rho}\left(\left\|\hat{\boldsymbol{\beta}}_j\right\|_2;\sqrt{M_n}\lambda,\gamma\right)\sqrt{M_n}\hat{\boldsymbol{\beta}}_j/\left\|\hat{\boldsymbol{\beta}}_j\right\|_2\right\}$$

Lemma 3. *Suppose that* $\tilde{\boldsymbol{B}}$ *satisfies that* $SRC\left(q^*,c_1,c_2\right)$, $q^*\ge\left(K_*+1\right)m_n q$, *and* $\gamma\ge c_1^{-1}\sqrt{4+\overline{c}}$. *Let* $m_*=K_*q$. *Then for any* $\tilde{Y}\in\mathcal{R}^n$ *with* $\lambda\ge 2\sqrt{c_2 M_n}\,\zeta\left(\tilde{Y};m_*,A_0^c\right)$, *we have* $|I|\le\left(K_*+1\right)q$.

proof This lemma can be proved along the line of the proof of Lemma 1 of [23] and is omitted. *proof of Theorem* 3.2. By Lemma 3, in the event

$$2\sqrt{c_2 M_n}\,\zeta\left(\tilde{\boldsymbol{Y}};m_*,A_0^c\right)\le\lambda \tag{6.4}$$

we have $|I|\le\left(K_*+1\right)q$, Thus in the event (6.4), the original model with p groups reduces a model with at most $\left(K_*+1\right)q$ groups, in this reduced model, the condition of Theorem 3.2 implies that the conditions of Theorem 3.2. By Lemma 2,

$$P\left(2\sqrt{c_2 M_n}\,\zeta\left(\tilde{Y};m_*,A_0^c\right)\le\lambda\right)\le\eta_{3n}\left(\lambda\right) \tag{6.5}$$

Therefore, combining (6.5) and Theorem 3.1, we have $P\left(\hat{\boldsymbol{\beta}}\left(\lambda,\gamma\right)\neq\hat{\boldsymbol{\beta}}^o\right)\le\eta_{1n}\left(\lambda\right)+\eta_{2n}\left(\lambda\right)+\eta_{3n}\left(\lambda\right)$, since $\hat{f}\left(x\right)=\boldsymbol{B}\hat{\boldsymbol{\beta}}$, we can obtain $P\left(\hat{f}\neq\hat{f}^o\right)\le\eta_{1n}\left(\lambda\right)+\eta_{2n}\left(\lambda\right)+\eta_{3n}\left(\lambda\right)$. This proves Theorem 3.2.