Computational Statistics and Data Analysis 184 (2023) 107743

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Weighted least squares model averaging for accelerated failure time models

Qingkai Dong, Binxia Liu, Hui Zhao *

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430070, China

ARTICLE INFO

Article history:
Received 21 July 2022
Received in revised form 7 March 2023
Accepted 13 March 2023
Available online 17 March 2023

Keywords:
Accelerated failure time model
Mallows criterion
Model averaging
Weighted least squares estimation

ABSTRACT

This paper proposes a new model averaging method for the accelerated failure time models with right censored data. A weighted least squares procedure is used to estimate the parameters of candidate models. In this procedure, the candidate models are not required to be nested, and the weights selected by Mallows criterion are not limited to be discrete, which make the proposed method very flexible and general. The asymptotic optimality of the proposed method is proved under some mild conditions. Particularly, it is shown that the optimality remains valid even when the variances of the error terms are estimated and the feasible weighted least squares estimators are averaged. Simulation studies show that the proposed method has better prediction performance than many popular model selection or model averaging methods when all candidate models are misspecified. Finally, an application about primary biliary cirrhosis is provided.

1. Introduction

In survival analysis, the accelerated failure time (AFT) model has received extensive attention and has become an important alternative to Cox models, since it is more natural and direct in describing the covariates effects on the event time than Cox models (Kalbfleisch and Prentice, 2011). Various strategies have been proposed to estimate the parameters in the AFT model, including Miller's estimator (Miller, 1976), Buckley-James estimator (Buckley and James, 1979; Jin et al., 2006), KSV estimator (Koul et al., 1981), and in this paper, the weighted least squares (WLS) estimator (Stute, 1993, 1996; He and Huang, 2003). Compared with other estimators, the WLS estimator has three major advantages. Firstly, it is easy to be carried out because no iterations are required. Next, it has consistency and asymptotic normality under reasonable assumptions. Lastly, comprehensive simulation studies in Bao et al. (2007) show that it performs much better than the other estimators, particularly when the number of covariates is large or the censoring is heavy.

In some practical problems, we need to choose useful covariates from many potential ones. Earlier model selection methods were based on information criteria such as AIC and BIC. Later, regularization methods become popular, including Tibshirani (1996); Fan and Li (2001); Zou (2006); Lv and Fan (2009); Dai et al. (2018). About model selection in the AFT model, there are some methods based on the penalized weighted least squares estimator, such as Huang et al. (2006); Hu and Chai (2013); Cheng et al. (2022). However, when a single model is not overwhelmingly supported by the data, these model selection methods may ignore contributions of other candidate models and suffer from the model selection uncer-

* Corresponding author.
E-mail address: hzhao@zuel.edu.cn (H. Zhao).

https://doi.org/10.1016/j.csda.2023.107743
0167-9473/

tainty and bias problem (Hjort and Claeskens, 2003). More importantly, when the data change, different model selection methods or criteria may lead to different optimal models.

To address these issues and improve prediction accuracy, various model averaging approaches have been proposed by exploiting all information from every candidate model. Inspired by AIC and BIC, Buckland et al. (1997) proposed smoothed AIC (SAIC) and smoothed BIC (SBIC) methods. Hjort and Claeskens (2003) proposed a local misspecification framework to establish properties of model averaging estimators. Hansen (2007) proposed a model averaging estimator with weights selected by minimizing a Mallows criterion. This Mallows model averaging (MMA) estimator asymptotically achieves the smallest possible squared error in the class of model averaging estimators. Wan et al. (2010) modified the conditions of Hansen (2007) by allowing non-nested candidate models and continuous weights. These improvements make the conditions of MMA more natural at the cost of limiting the number of candidate models. Another important model averaging criterion is the Jackknife model averaging (JMA) proposed by Hansen and Racine (2012), which selects the weights by minimizing a cross-validation criterion and has significantly lower MSE than MMA when the errors are heteroskedastic.

In survival analysis, MMA and JMA, the most representative frequentist model averaging criteria, have not been used until recently. Under the proportional hazards model assumption, He et al. (2020) improved the prediction accuracy of the integral intensity function by JMA, and Li et al. (2021) proposed a semiparametric model averaging prediction method to approximate the nonparametric regression function by a weighted sum of low-dimensional nonparametric submodels.

As for the AFT model, Yan et al. (2021) proposed a high dimensional JMA procedure, where the penalized Buckley-James method (Wang et al., 2008) was used to obtain the coefficient estimators. However, the convergence of Buckley-James estimate cannot be guaranteed, and the possible overlap of variables in different candidate models is not considered in Yan et al. (2021). Recently, Liang et al. (2022) proposed another model averaging method based on KSV estimate and MMA criterion. As specified by Bao et al. (2007), in many cases, the effect of KSV estimator is not as good as WLS estimator. Moreover, constructing a linear model for the synthetic response may face the problem of excessive error variance.

Therefore in this paper, we propose the weighted least squares model averaging (WLSMA) method under the AFT model, where the averaging weights are selected by minimizing a MMA criterion. We show that the proposed method has asymptotic optimality in the sense of Li (1986). In particular, as the variances of error terms are unknown in many applications, we also consider the estimation of variance in the Mallows criterion and prove that even when the variances of the error terms are estimated and the feasible weighted least squares estimators are averaged, our method still has asymptotic optimality, which is the most important theoretical property of model averaging method and one of main theoretical contributions of this paper. Besides, our method allows continuous weights, and the variables in each candidate model can be overlapped, which greatly improves the flexibility and applicability of the method. Extensive simulation shows that our WLSMA method outperforms many existing model selection and model averaging methods. In the empirical study of the PBC dataset, WLSMA method has also obtained good prediction accuracy.

The rest of the paper is organized as follows. We begin in Section 2 with the description of some notations, the AFT model and the WLS estimate. In Section 3, we propose our WLSMA method and present the asymptotic optimality of this new method. Sections 4 and 5 report the simulation results and the application in the PBC dataset. Finally, we provide some concluding remarks in Section 6 and outline the proofs of the theorems in the Appendix.

2. Notations and model

Let T and V denote the survival time and censored time, respectively. $\tilde{Y} = \log T$ and $C = \log V$. $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)'$ denotes the covariate matrix for N independent observations, where the dimension of $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots)$ is countably infinite. The AFT model assumes

$$\tilde{Y}_i = \tilde{\mu}_i + e_i = \sum_{j=1}^{\infty} \beta_j x_{ij} + e_i, \quad i = 1, \ldots, N, \tag{1}$$

with $E(e_i|\boldsymbol{x}_i) = 0$ and $E\left(e_i^2|\boldsymbol{x}_i\right) = \sigma^2$. We consider a sequence of linear approximating models $m = 1, \ldots, M$, where the mth model, with any $k_m (> 0)$ regressors belonging to $\boldsymbol{x}_i$, takes the form of

$$\tilde{Y}_i = \sum_{j=1}^{k_m} \beta_{j,m} x_{ij,m} + e_i, \quad i = 1, \ldots, N,$$

which can be rewritten as

$$\tilde{Y} = X_m \boldsymbol{\beta}_m + \boldsymbol{e},$$

where $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_N)'$ and $\boldsymbol{\beta}_m = \left(\beta_{1,m}, \beta_{2,m}, \ldots, \beta_{k_m,m}\right)'$. Here X_m is the corresponding $N \times k_m$ submatrix of X and assumed to be column full rank.

Following Stute (1993), we assume C is independent of X and $\tilde{Y}$, and $P(\tilde{Y} \leqslant C \mid \tilde{Y}, X) = P(\tilde{Y} \leqslant C \mid \tilde{Y})$ for identifiability. When $\tilde{Y}$ is subject to random right censoring, we only observe $(U_i, \boldsymbol{x}_i, \delta_i), i = 1, 2, \ldots, N$, where $U_i = \min(\tilde{Y}_i, C_i)$ and

$\delta = I(\tilde{Y}_i \leqslant C_i)$. Let $F(x)$ and $G(x)$ be the distribution functions of $\tilde{Y}_i$ and C_i, respectively and assume $\sup\{x : F(x) < 1\} \leq \sup\{x : G(x) < 1\}$.

By Minimizing the weighted least squares loss function

$$Q(\boldsymbol{\beta}_m) = \sum_{i=1}^{N} \frac{\delta_i}{1 - G(U_i)} \left(U_i - \boldsymbol{x}'_{i,m}\boldsymbol{\beta}_m\right)^2,$$

we have the WLS estimator

$$\hat{\boldsymbol{\beta}}_m = \left(\sum_{i=1}^{N} a_i \boldsymbol{x}_{i,m}\boldsymbol{x}'_{i,m}\right)^{-1} \left(\sum_{i=1}^{N} a_i \boldsymbol{x}_{i,m} U_i\right), \tag{2}$$

where $a_i = \delta_i/(1 - G(U_i))$ is related to the inverse probability weighting. Obviously, $a_i = 0$ when individual i is censored and $a_i \geq 1$ otherwise. Stute (1993) discussed this weight and its theoretical properties. Simulation results from Stute (1993) and Bao et al. (2007) demonstrate that the WLS estimator outperforms the well-known Miller, Buckley-James, and KSV estimators, especially when the number of covariates increases or the censoring is heavy.

3. The proposed model averaging method

Let n denote the number of uncensored observations in all N observations, and $D = \mathrm{diag}(d_1, \ldots, d_n)$ denote the diagonal matrix consisting of the non-zero elements in $\{a_i\}_{i=1}^N$. Similarly, denote Z_m as the $n \times k_m$ submatrix of X_m composed of the n uncensored individuals' covariates under the mth candidate model.

Since the weight $a_i = 0$ for the censored individuals, the WLS estimator (2) of $\boldsymbol{\beta}_m$ in the mth ($m = 1, \ldots, M$) model can be rewritten as

$$\hat{\boldsymbol{\beta}}_m = \left(Z'_m D Z_m\right)^{-1} Z'_m D Y, \tag{3}$$

where $Y = (Y_1, \ldots, Y_n)'$ is the corresponding uncensored subvector of $\tilde{Y}$. It's not difficult to find that the estimators obtained by using all observations are equivalent to those obtained by using the weighted version of uncensored observations. Let $\mu_i = E(Y_i|\boldsymbol{x}_i)$, then the estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ from the mth candidate model is:

$$\hat{\boldsymbol{\mu}}_m = Z_m \hat{\boldsymbol{\beta}}_m = Z_m \left(Z'_m D Z_m\right)^{-1} Z'_m D Y = P_m Y, \tag{4}$$

where $P_m = Z_m \left(Z'_m D Z_m\right)^{-1} Z'_m D$. Usually the distribution of C is unknown, and we can replace $G(x)$ by the Kaplan-Meier estimator $\hat{G}(x) = 1 - \prod_{U_j \leqslant x} \left[\frac{N-j}{N-j+1}\right]^{1-\delta_j}$. It should be noted that the resulted WLS estimator is different from the ordinary least squares estimator calculated by using only uncensored observations, because the information of censored observations has been used in constructing $\hat{G}(x)$, rather than being directly discarded.

We expect that under the assumption of global model misspecification, averaging the estimators of μ from multiple candidate models would produce a better estimator than any individual model. Let $\boldsymbol{w} = (w_1, \ldots, w_M)'$ be an $M \times 1$ weighting vector from

$$\mathcal{H}_M = \left\{\boldsymbol{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\right\}.$$

The model averaging estimator of $\boldsymbol{\mu}$ is defined as

$$\begin{aligned}\hat{\mu}(\boldsymbol{w}) &= \sum_{m=1}^{M} w_m \hat{\mu}_m \\ &= \sum_{m=1}^{M} w_m Z_m \left(Z'_m D Z_m\right)^{-1} Z'_m D Y \\ &= P(\boldsymbol{w}) Y \end{aligned} \tag{5}$$

for some $\boldsymbol{w} \in \mathcal{H}_M$, where the matrix $P(\boldsymbol{w}) = \sum_{m=1}^{M} w_m Z_m \left(Z'_m D Z_m\right)^{-1} Z'_m D = \sum_{m=1}^{M} w_m P_m$. We would like to choose a weight vector that achieves a small MSE for the fitted model.

Consider the squared loss function

$$\begin{aligned}L_n(\boldsymbol{w}) &= \sum_{i=1}^{n} \left(\mu_i - \hat{\mu}_i(\boldsymbol{w})\right)^2 \\ &= \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2, \end{aligned} \tag{6}$$

where $\|\cdot\|$ denote the Euclidean norm. Then the risk function is

$$R_n(\boldsymbol{w}) = E[L_n(\boldsymbol{w})]. \tag{7}$$

Since $\boldsymbol{\mu}$ is unknown, we cannot minimize (6) directly, so we use the estimation of loss function as the criterion to choose the optimal weighting vector. This criterion is in the spirit of Mallows criterion from Hansen (2007), which is defined as

$$C_n(\boldsymbol{w}) = \|Y - \hat{\mu}(\boldsymbol{w})\|^2 + 2\sigma^2 \operatorname{tr}\{P(\boldsymbol{w})\}. \tag{8}$$

With some calculation, we can observe that

$$E[C_n(\boldsymbol{w})] = E[L_n(\boldsymbol{w})] + n\sigma^2.$$

Since $C_n(\boldsymbol{w})$ is an unbiased estimator of the risk function $R_n(\boldsymbol{w})$ plus a term that do not depend on $\boldsymbol{w}$, we choose the weighting vector by minimizing $C_n(\boldsymbol{w})$.

Let $\boldsymbol{v} = (k_1, \ldots, k_M)'$ and $B = \left(\hat{\boldsymbol{\mu}}_1 - Y, \ldots, \hat{\boldsymbol{\mu}}_M - Y\right)$, then $C_n(\boldsymbol{w}) = \boldsymbol{w}'B'B\boldsymbol{w} + 2\sigma^2\boldsymbol{v}'\boldsymbol{w}$. This is a typical constrained quadratic programming problem and can be solved quickly by statistical software. Denote $\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathcal{H}_M}{\arg\min}\, C_n(\boldsymbol{w})$, we then use $\hat{\boldsymbol{\mu}}(\boldsymbol{w}^*)$ as our WLSMA estimator of $\boldsymbol{\mu}$. It should be emphasized that although the loss function we define is the squared loss for those uncensored observations, the estimators $\hat{\boldsymbol{\beta}}_m$ obtained by WLS in each candidate model already take the censored observations into account and minimizing (8) is to find the optimal linear combination of these estimators. In addition, this definition avoids many technical difficulties in proving the optimality theorem. Simulations also show that our WLSMA method works well.

Now we establish the asymptotic optimality of WLSMA. First, we need the following regular conditions:

(C1) Define $\xi_n = \inf_{\boldsymbol{w}\in\mathcal{H}_M} R_n(\boldsymbol{w})$ and $\boldsymbol{w}_m^0$ is an $M \times 1$ unit vector in which the mth element is 1 and the others are 0. For some integer $1 \le J < \infty$ and some positive constant κ such that $E\left(e_i^{4J}|\boldsymbol{x}_i\right) \le \kappa < \infty$, assume

$$M\xi_n^{-2J}\sum_{m=1}^{M}\left(R_n\left(\boldsymbol{w}_m^0\right)\right)^J \to 0.$$

(C2) Let $\lambda_{\max}(\cdot)$ denote the maximum eigenvalue of the matrix, assume that

$$\lim_{n\to\infty}\sup_{\boldsymbol{w}\in\mathcal{H}_M}\lambda_{\max}(P(\boldsymbol{w})) < \infty.$$

(C3) $P(\tilde{Y} \leqslant C) \in [c_0, 1]$, where c_0 is a positive constant.

Condition (C1) imposes a bound on the conditional moments of $\boldsymbol{e}$, which is used in most of the model averaging literature (see Wan et al. (2010); Liu et al. (2016)). It also requires that there is no finite approximating model for which the bias is zero. This is obvious because the real model (1) is not in the candidate model set. (C2) is a mild and natural condition (see, for example, Li (1986); Liu et al. (2016)). (C3) requires enough uncensored observations to ensure the amount of information in the data.

Next, we present the main results of this paper, which demonstrate the asymptotic optimality of the WLSMA estimator under the non-nested set-up described above. The proof of these theorems will be sketched in the appendix.

Theorem 1. *Assume that the regularity conditions (C1)–(C3) hold. Then as* $n \to \infty$,

$$\frac{L_n(\boldsymbol{w}^*)}{\inf_{\boldsymbol{w}\in\mathcal{H}_M} L_n(\boldsymbol{w})} \xrightarrow{p} 1.$$

Theorem 1 implies that the weight vector in WLSMA yields a squared error that is asymptotically identical to that of the infeasible optimal weight vector restricted to $\mathcal{H}_M$. When $G(x)$ is unknown and replaced by $\hat{G}(x)$, define $\hat{d}_i = \delta_i/(1 - \hat{G}(U_i))$ and $\hat{D} = \operatorname{diag}\left(\hat{d}_1, \ldots, \hat{d}_n\right)$. Then (5) becomes

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w}) &= \sum_{m=1}^{M} w_m \hat{\boldsymbol{\mu}}_{\hat{G},m} \\ &= \sum_{m=1}^{M} w_m Z_m \left(Z_m'\hat{D}Z_m\right)^{-1} Z_m'\hat{D}Y \\ &= P_{\hat{G}}(\boldsymbol{w})Y,\end{aligned}$$

and the corresponding formulas (6) and (8) become

$$L_{\hat{G}}(\boldsymbol{w}) = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\|^2, \tag{9}$$

$$C_{\hat{G}}(\boldsymbol{w}) = \|Y - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\|^2 + 2\sigma^2 \operatorname{tr}\{P_{\hat{G}}(\boldsymbol{w})\}. \tag{10}$$

We need three additional conditions and modify Theorem 1 to maintain the optimality.

(C4) $\lim_{n\to\infty} \sum_{i=1}^{n} z_{ij,m}^2/n < \infty$ uniformly in j and m. $\lambda_{\max}\left(\sum_{i=1}^{n} \boldsymbol{z}_{i,m}\boldsymbol{z}_{i,m}'/n\right) < \infty$ and $\lambda_{\min}\left(\sum_{i=1}^{n} \boldsymbol{z}_{i,m}\boldsymbol{z}_{i,m}'/n\right) > 0$ uniformly in n and m.
(C5) $\boldsymbol{\mu}'\boldsymbol{\mu}/n = O(1)$.
(C6) $k_{m^*}^2/n = o(1)$, $k_{m^*}/\xi_n = o_p(1)$ and $k_{m^*}^2/\xi_n = O_p(1)$ as $n \to \infty$, where $k_{m^*} = \max_{1\le m\le M} k_m$.

The bounded conditions in (C4) and (C5) are required to prove Theorem 2. (C6) is a restriction on the number of covariates in the candidate models.

Theorem 2. *Denote* $\hat{\boldsymbol{w}}^* = \arg\min_{\boldsymbol{w}\in\mathcal{H}_M} C_{\hat{G}}(\boldsymbol{w})$. *Assume (C1)–(C6) hold and* $n \to \infty$, *then*

$$\frac{L_{\hat{G}}(\hat{\boldsymbol{w}}^*)}{\inf_{\boldsymbol{w}\in\mathcal{H}_M} L_{\hat{G}}(\boldsymbol{w})} \xrightarrow{p} 1.$$

Furthermore, the variance σ^2 of error terms is usually unknown in many applications and here we replaced it by $\hat{\sigma}^2 = \|Y - \hat{\boldsymbol{\mu}}_{m^*}\|^2/(n - k_{m^*})$ (see Hansen (2007)), then the corresponding formula (10) becomes $\hat{C}_{\hat{G}}(\boldsymbol{w}) = \|Y - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\|^2 + 2\hat{\sigma}^2 \operatorname{tr}\{P_{\hat{G}}(\boldsymbol{w})\}$. The following theorem shows that when σ^2 is replaced by $\hat{\sigma}^2$ and the feasible weighted least squares estimators are averaged, the optimality of the proposed method remains valid.

Theorem 3. *Define* $\tilde{\boldsymbol{w}}^* = \arg\min_{\boldsymbol{w}\in\mathcal{H}_M} \hat{C}_{\hat{G}}(\boldsymbol{w})$ *and suppose that (C1)-(C6) hold. When* $\hat{\sigma}^2 = \|Y - \hat{\boldsymbol{\mu}}_{m^*}\|^2/(n - k_{m^*})$ *and* $n \to \infty$, *we have*

$$\frac{L_{\hat{G}}\left(\tilde{\boldsymbol{w}}^*\right)}{\inf_{\boldsymbol{w}\in\mathcal{H}_M} L_{\hat{G}}(\boldsymbol{w})} \xrightarrow{p} 1.$$

4. Simulation

In the simulation study, the data are generated from the AFT model, $\log(T_i) = \tilde{Y}_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i$, where e_i follows the normal distribution $\mathcal{N}(0, 1)$. The censoring time C_i is generated from $\mathcal{N}(C_0, 2)$. By adjusting the value of C_0, the censoring rate(CR) is about 20%, 35% and 50%. We set $N = 100, 200$ and $p = 100$, and consider different cases about the selection of candidate models and true values of $\boldsymbol{\beta}$.

Case 1 (The nested models): We assume that only the first $\lfloor 3N^{1/3} \rfloor$ covariates could be observed. The mth model uses the first m covariates so that $k_m = m$ and $k_{m^*} = M = \lfloor 3N^{1/3} \rfloor$. When $N = 100$ and 200, $M = 13$ and 17. Here the covariates are generated from a multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$. True coefficients $\beta_j = 1/j^2$ or $\sqrt{2}/j$ for $j = 1, 2, \ldots, p$, so that the true model would not be in the candidate model set.

Case 2 (The non-nested models): We assume that only the first 5 variables could be observed and any combination of them are considered, so $k_{m^*} = 5$ and there would be $M = 2^5 - 1 = 31$ candidate models. The set-ups of $\boldsymbol{\beta}$ and covariates are consistent with Case 1.

Case 3 (True model in candidate set): Let the true value of $\beta_j = 0.7$ for $j = 1, 2, 3, 4, 5$ and $\beta_j = 0$ for $j = 6, \ldots, p$. Among p covariates, the 1st and 3rd are randomly generated by a Poisson distribution with parameter $\lambda = 1$, the 2nd and 4th by a binomial distribution $B(1, 0.1)$, and the 5th to 100th by i.i.d. standard normal distribution. To make the covariates correlated with each other, the covariate matrix is multiplied to an upper diagonal matrix R with diagonal elements equal to 1 and non-diagonal elements equal to $\sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$. The way to construct the set of candidate models is the same as in Case 1 and Case 2. In this setting, the true model is always included in the set of candidate models, either in the nested or non-nested scenarios.

In this section, we compare our WLSMA with other classical model selection or model averaging methods. The following is a brief description of them:

Table 1
Mean of MSEs in Case 1 with $\beta_j = 1/j^2$.

Method\CR	20%	35%	50%	20%	35%	50%
	$N = 100$ and $M = 13$			$N = 200$ and $M = 17$		
WLS	0.1868	0.2473	0.3433	0.1323	0.1706	0.2425
KSVMA	0.1179	0.1797	0.2787	0.0650	0.1045	0.2131
WLSMA1	0.0688	0.0803	0.1015	0.0411	0.0478	0.0634
WLSMA2	0.0728	0.0926	0.1270	0.0427	0.0518	0.0745
AIC	0.1155	0.2056	0.3067	0.0824	0.1352	0.2231
BIC	0.0845	0.1123	0.1924	0.0503	0.0681	0.1189
SAIC	0.0764	0.0971	0.1302	0.0508	0.0641	0.0939
SBIC	0.0741	0.0939	0.1261	0.0494	0.0623	0.0912
LASSO	0.0975	0.1438	0.2204	0.0653	0.0890	0.1471
SCAD	0.0992	0.1410	0.2083	0.0755	0.1001	0.1459

- Model selection methods based on AIC and BIC: The mth model's criteria are $\mathrm{AIC}_m = \log\left(\hat{\sigma}_m^2\right) + 2n^{-1}\,\mathrm{tr}\{P_{\hat{G},m}\}$ and $\mathrm{BIC}_m = \log\left(\hat{\sigma}_m^2\right) + n^{-1}\log(n)\,\mathrm{tr}\{P_{\hat{G},m}\}$, $m = 1, \ldots, M$.
- Penalty methods LASSO and SCAD: The penalty likelihood function is $Q\left(\boldsymbol{\beta}_{m^*}\right) + \sum_{j=1}^{k_{m^*}} p\left(\lambda_n, \beta_{j,m^*}\right)$, $m = 1, \ldots, M$. The penalty functions are $p_{LASSO}\left(\lambda_n, \beta_{j,m^*}\right) = \lambda_n|\beta_{j,m^*}|$ and

$$p_{SCAD}\left(\lambda_n, \beta_{j,m^*}\right) = \begin{cases} \lambda_n\left|\beta_{j,m^*}\right|, & \left|\beta_{j,m^*}\right| \le \lambda_n, \\ -\frac{\beta_{j,m^*}^2 - 2\alpha\lambda_n\left|\beta_{j,m^*}\right| + \lambda_n^2}{|2(\alpha-1)|}, & \lambda_n < \left|\beta_{j,m^*}\right| \le \alpha\lambda_n, \\ \frac{(\alpha+1)\lambda_n^2}{2}, & \alpha\lambda_n < \left|\beta_{j,m^*}\right|. \end{cases}$$

  Following Fan and Li (2001), we take $\alpha = 3.7$. The choice of λ_n is crucial to the performance of the penalty method. We use the 'ncvreg' package in R to implement LASSO and SCAD, and the cross-validation method to pick λ_n. As for the range of λ_n, we pick a very small value (e.g. 1×10^{-4}) as a starting point, up to the maximum value that makes all coefficients become zero. Within this range 100 values are taken.
- Model averaging methods based on SAIC and SBIC: The weights of the mth model are:

$$\mathrm{w}_{SAIC,m} = \exp\left(-\mathrm{AIC}_m/2\right) / \sum_{j=1}^{M} \exp\left(-\mathrm{AIC}_j/2\right),$$

$$\mathrm{w}_{SBIC,m} = \exp\left(-\mathrm{BIC}_m/2\right) / \sum_{j=1}^{M} \exp\left(-\mathrm{BIC}_j/2\right), m = 1, \ldots, M.$$

- Mallows model averaging based on KSV estimation: We adopt the model averaging method proposed in this paper, but change the WLS estimation into KSV estimation, that is, let $\hat{\boldsymbol{\beta}}_{KSV,m} = \left(\sum_{i=1}^{N} \boldsymbol{x}_{i,m}\boldsymbol{x}_{i,m}'\right)^{-1}\left(\sum_{i=1}^{N} a_i \boldsymbol{x}_{i,m} U_i\right)$. We named it KSVMA.
- WLS estimation: We also consider the WLS estimation based on all M variables.

In our WLSMA method, another choice of $\hat{\sigma}^2$ is $(Y - \hat{\boldsymbol{\mu}}_{m^*})'\hat{D}(Y - \hat{\boldsymbol{\mu}}_{m^*})/N$ which is recommended by He and Huang (2003). We named it WLSMA2 and the method in Theorem 3 WLSMA1. Evaluation is based on mean squared error (MSE) of $\tilde{\mu}$ defined in (1): $\mathrm{MSE} = \frac{1}{N}\left\|\hat{\tilde{\boldsymbol{\mu}}} - \tilde{\boldsymbol{\mu}}\right\|^2$, which is a common criterion to measure predictive uncertainty. It evaluates the predictive performance of these methods for all observations. For our WLSMA method, $\hat{\tilde{\boldsymbol{\mu}}} = \sum_{m=1}^{M} \tilde{w}_m^* X_m \hat{\boldsymbol{\beta}}_{\hat{G},m}$, where $\hat{\boldsymbol{\beta}}_{\hat{G},m}$ is calculated by plugging $\hat{G}(x)$ into (3). We report the mean of MSEs of 100 replications. Box-plots of the MSEs are also displayed. These results are shown in Figs. 1–6 and Tables 1–6.

From Fig. 1 and Table 1, it can be seen that as far as MSE is concerned, the proposed WLSMA methods outperform the other methods, where WLSMA1 is the best under all settings, and WLSMA2 is close to WLSMA1. Model averaging methods SAIC and SBIC also perform well under most settings. Besides, when the censoring rate is low, the performance of most methods is relatively close, but when it increases, WLSMA methods are obviously better. WLSMA1 tends to have the smallest quartile deviation and lower box position.

Fig. 2 and Table 2 also display the advantage of WLSMA1 and WLSMA2 over the other methods, but the performance of using WLS estimation on all observable variables is also good. In addition, it is easy to see that the KSVMA method crashes under these settings. We believe that this may be due to the slower decay rate of β with j. Under this setting, the performance of each method decreases.

Figs. 3–6 and Tables 3–6 show the results of Case 2 and Case 3. The conclusion is almost consistent with Case 1, that is, WLSMA has the best performance in most settings, especially in high censoring rates cases.

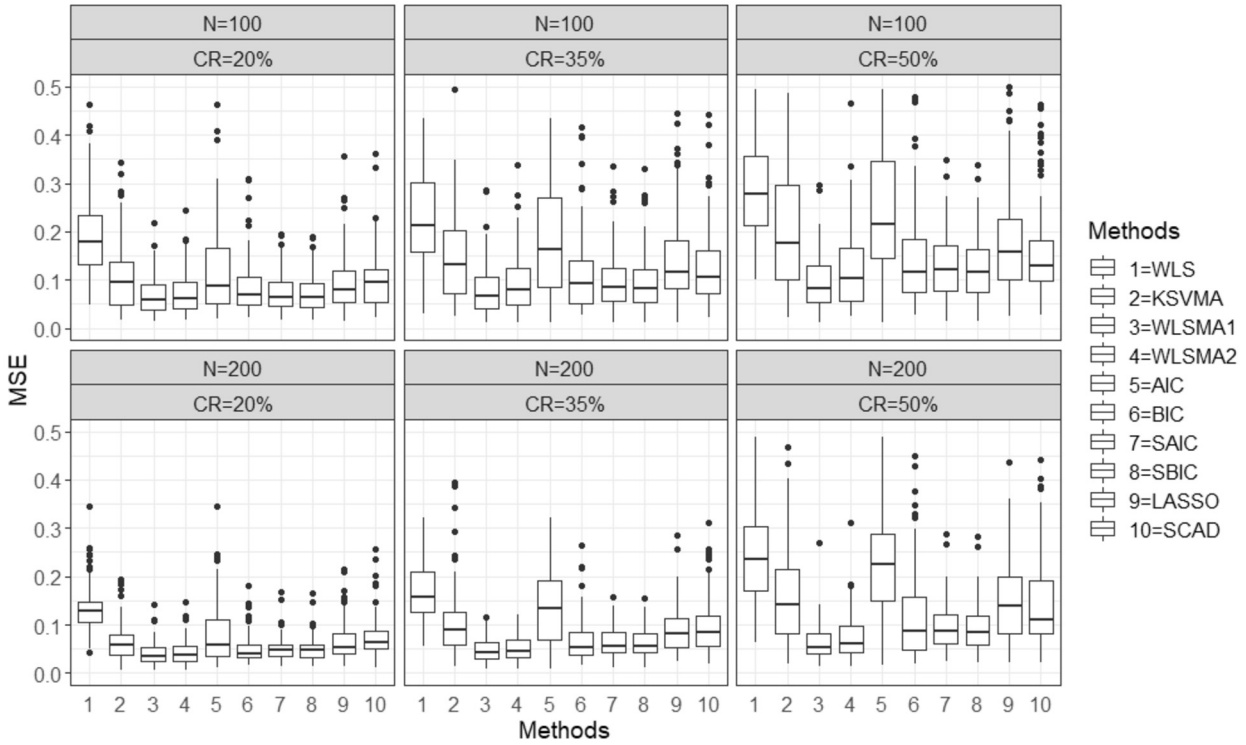

Fig. 1. Box-plots of MSEs in Case 1 with $\beta_j = 1/j^2$.

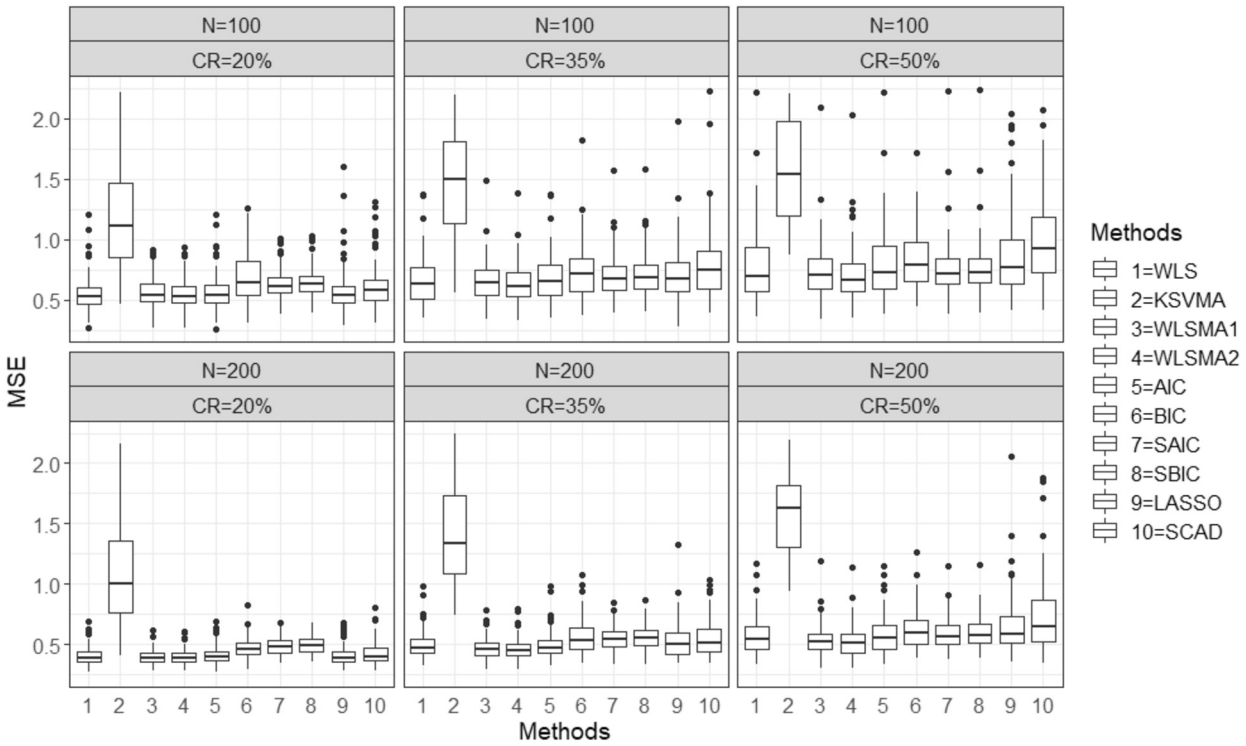

Fig. 2. Box-plots of MSEs in Case 1 with $\beta_j = \sqrt{2}/j$.

Table 2
Mean of MSEs in Case 1 with $\beta_j = \sqrt{2}/j$.

Method\CR	20%	35%	50%	20%	35%	50%
	$N = 100$ and $M = 13$			$N = 200$ and $M = 17$		
WLS	0.5516	0.6588	0.7755	0.4061	0.4917	0.5720
KSVMA	1.5043	2.6591	4.1299	1.2137	2.4616	3.6381
WLSMA1	0.5597	0.6494	0.7426	0.4012	0.4751	0.5414
WLSMA2	0.5475	0.6311	0.7143	0.3965	0.4655	0.5276
AIC	0.5705	0.6734	0.7909	0.4109	0.4973	0.5773
BIC	0.6756	0.7321	0.8595	0.4775	0.5561	0.6159
SAIC	0.6294	0.6983	0.7577	0.4867	0.5483	0.5925
SBIC	0.6408	0.7090	0.7674	0.4946	0.5561	0.5999
LASSO	0.5756	0.7292	0.9076	0.4081	0.5315	0.6857
SCAD	0.6135	0.7873	1.0207	0.4350	0.5621	0.7515

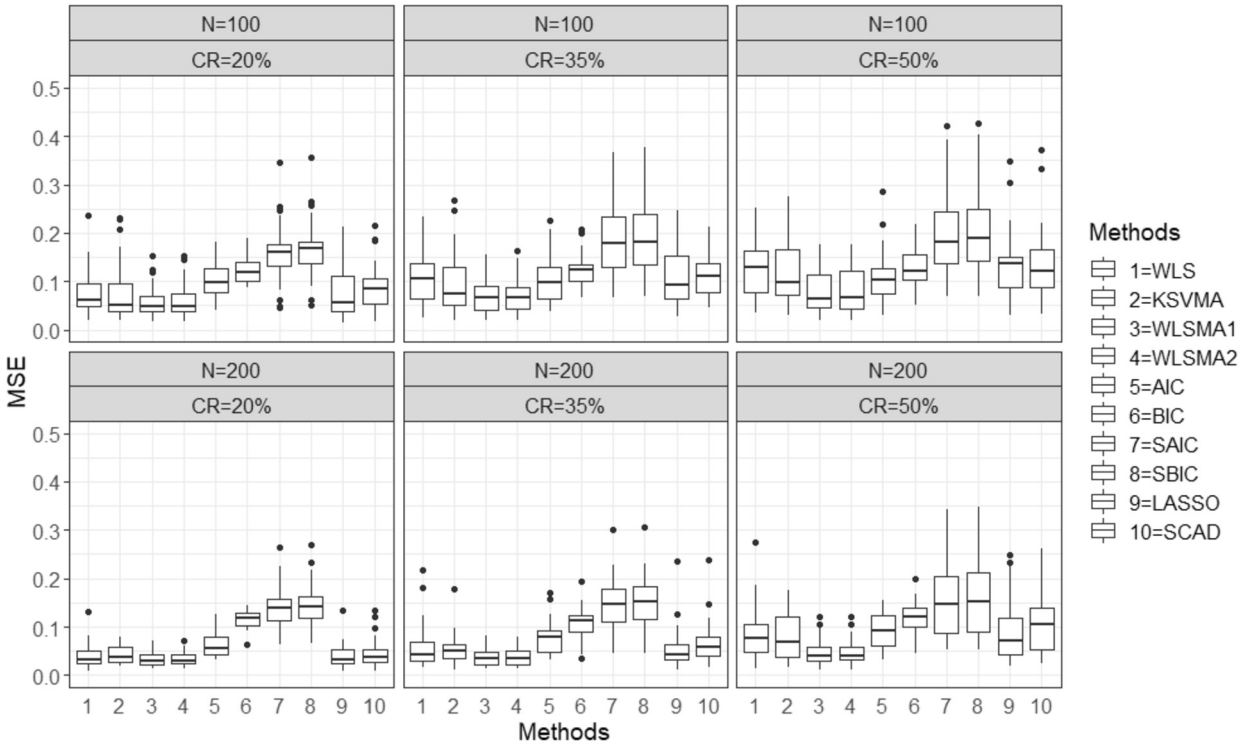

Fig. 3. Box-plots of MSEs in Case 2 with $\beta_j = 1/j^2$.

Table 3
Mean of MSEs in Case 2 with $\beta_j = 1/j^2$.

Method\CR	20%	35%	50%	20%	35%	50%
	$N = 100$ and $M = 31$			$N = 200$ and $M = 31$		
WLS	0.0788	0.1072	0.1261	0.0397	0.0576	0.0871
KSVMA	0.0821	0.1173	0.1216	0.0439	0.0530	0.0781
WLSMA1	0.0588	0.0686	0.0783	0.0339	0.0375	0.0481
WLSMA2	0.0609	0.0696	0.0794	0.0340	0.0372	0.0481
AIC	0.0991	0.1018	0.1106	0.0623	0.0784	0.0938
BIC	0.1238	0.1251	0.1469	0.1157	0.1091	0.1194
SAIC	0.1583	0.1793	0.1967	0.1398	0.1446	0.1559
SBIC	0.1652	0.1857	0.2028	0.1446	0.1489	0.1559
LASSO	0.0759	0.1075	0.1351	0.0401	0.0566	0.0885
SCAD	0.0889	0.1166	0.1383	0.0458	0.0680	0.1074

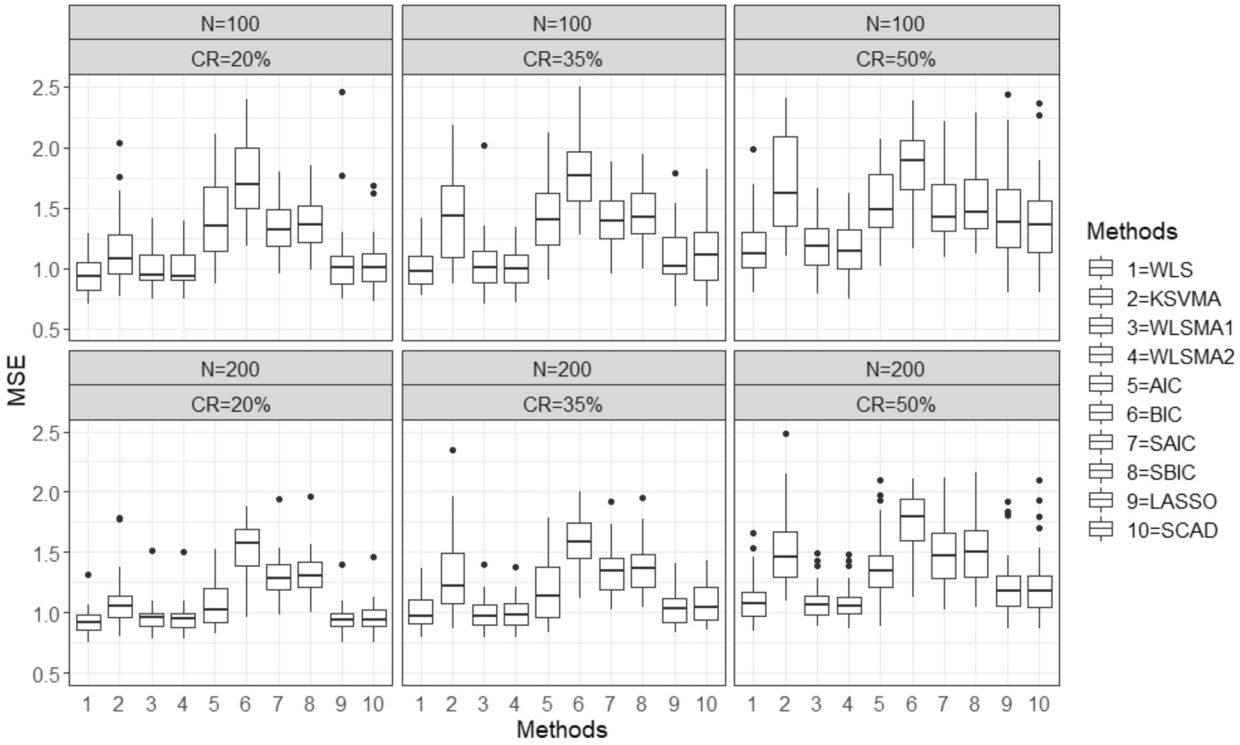

Fig. 4. Box-plots of MSEs in Case 2 with $\beta_j = \sqrt{2}/j$.

Table 4
Mean of MSEs in Case 2 with $\beta_j = \sqrt{2}/j$.

Method\CR	20%	35%	50%	20%	35%	50%
	$N = 100$ and $M = 31$			$N = 200$ and $M = 31$		
WLS	0.9987	1.0771	1.1524	0.9289	0.9940	1.0613
KSVMA	1.3399	1.4930	1.9220	1.0358	1.4020	1.6006
WLSMA1	0.9944	1.0870	1.1611	0.9530	1.0083	1.0225
WLSMA2	0.9912	1.0817	1.1441	0.9528	0.9796	1.0167
AIC	1.4850	1.5677	1.6331	1.1036	1.2386	1.2746
BIC	1.7596	1.9085	2.0435	1.6229	1.6349	1.6589
SAIC	1.3203	1.4215	1.5667	1.2764	1.3835	1.4173
SBIC	1.3538	1.4574	1.6073	1.2975	1.4064	1.4402
LASSO	1.0544	1.1923	1.5132	0.9429	1.0472	1.2032
SCAD	1.0619	1.2222	1.5105	0.9443	1.0368	1.1689

Table 5
Mean of MSEs in Case 3 with non-nested candidates.

Method\CR	20%	35%	50%	20%	35%	50%
	$N = 100$			$N = 200$		
WLS	0.2046	0.3151	0.3978	0.1525	0.2237	0.1501
KSVMA	0.8900	1.4777	2.4417	0.5890	1.0616	0.6302
WLSMA1	0.1151	0.1749	0.2349	0.0742	0.1087	0.0668
WLSMA2	0.1187	0.1846	0.2541	0.0767	0.1144	0.0696
AIC	0.1595	0.2823	0.3679	0.1121	0.1991	0.1126
BIC	0.1072	0.2061	0.3026	0.0722	0.1259	0.0706
SAIC	0.1669	0.2342	0.2960	0.1179	0.1603	0.1043
SBIC	0.1730	0.2393	0.3011	0.1209	0.1626	0.1065
LASSO	0.1895	0.3488	0.5243	0.1361	0.2343	0.1282
SCAD	0.2042	0.4038	0.5749	0.0995	0.2161	0.1025

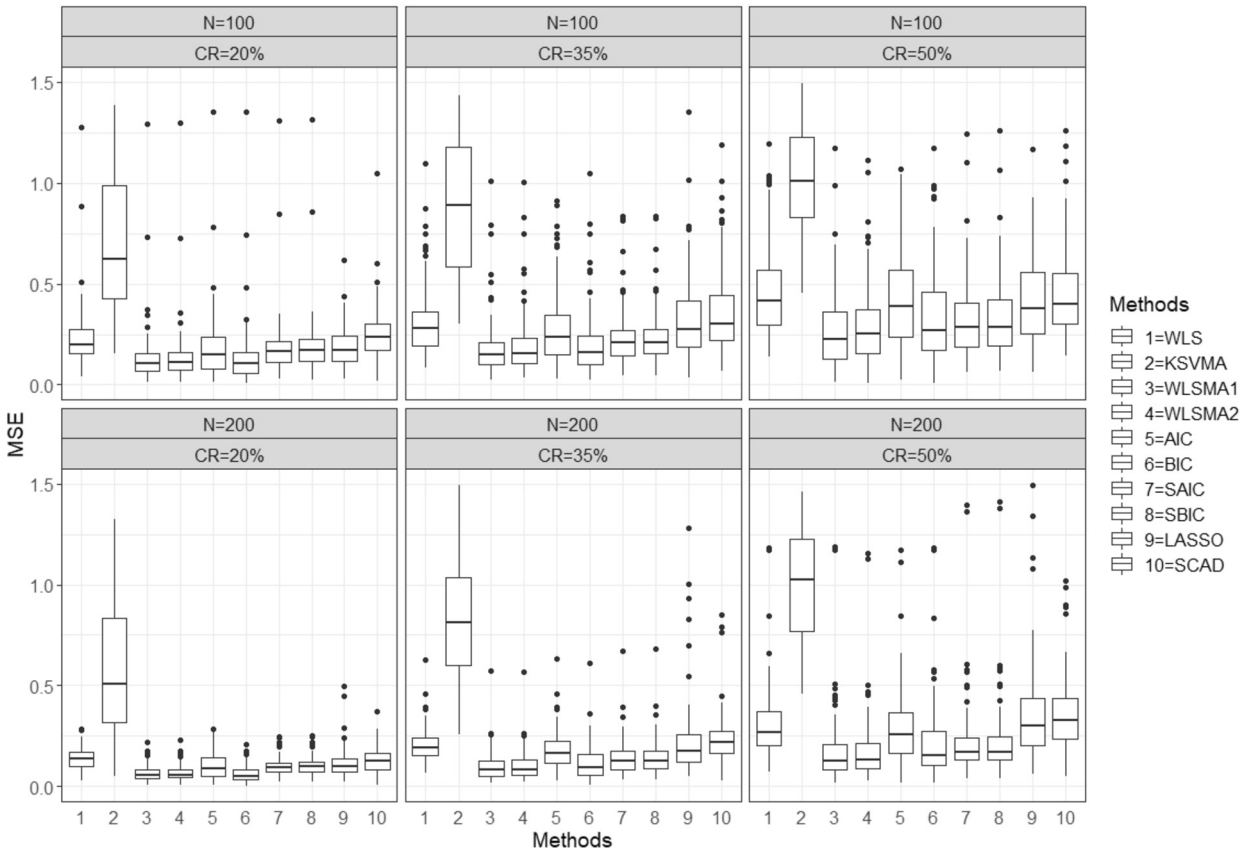

Fig. 5. Box-plots of MSEs in Case 3 with nested candidates.

Table 6
Mean of MSEs in Case 3 with non-nested candidates.

Method\CR	20% $N = 100$	35%	50%	20% $N = 200$	35%	50%
WLS	0.1112	0.1772	0.2433	0.0526	0.0935	0.1351
KSVMA	0.7827	1.0383	1.5549	0.4780	0.8329	1.3864
WLSMA1	0.1418	0.1921	0.2696	0.0965	0.0876	0.1251
WLSMA2	0.1454	0.2002	0.2441	0.0912	0.0889	0.1432
AIC	0.3280	0.4356	0.4853	0.0572	0.1165	0.1612
BIC	0.8704	0.9454	1.0227	0.4081	0.4825	0.5695
SAIC	0.3535	0.4340	0.4949	0.2698	0.3205	0.3797
SBIC	0.3923	0.4747	0.5377	0.2924	0.3438	0.4048
LASSO	0.1542	0.2866	0.4323	0.0652	0.1311	0.2459
SCAD	0.1600	0.2920	0.4348	0.0659	0.1264	0.2717

Next, to investigate the effect of covariate correlations on the prediction accuracy, we fix $\beta_j = 1/j^2$ and the censoring rate at 35%, N at 200 for all three cases, let the covariates correlation parameter ρ increase from 0.2 to 0.9 and record the prediction performance of WLSMA and the best results of the other eight methods. Results are shown in Fig. 7. For Case 1 (nested) and 2 (non-nested), it is clear that covariate correlations have minor effect on prediction, and WLSMA dominates other methods. Similar conclusions can be drawn from Case 3, except for some situations where the correlations are small.

Following one reviewer's suggestion, we study the robustness of WLSMA to the model structure misspecification. For this, we fit the data from a Cox's model with our AFT model averaging method to explore the generalization ability of our method. Specifically, generate the survival time T randomly from a Cox's model with hazard function $\lambda(t|\boldsymbol{x}) = \lambda_0 \exp\{-\sum_{j=1}^{p} \beta_j x_j\}$, where $\lambda_0 = 0.2$ and the covariates are generated from a zero-mean multivariate normal with covariance $\sigma_{ij} = 0.8^{|i-j|}$. The censoring variable C_i is still generated from $\mathcal{N}(C_0, 2)$. We control the censoring rate by adjusting C_0 and fix N at 200. However, since the Cox's model models hazard of event occurrence rather than survival time, the estimate of the mean of individual survival time is unavailable. In this case, we calculate the value of $C_n(\boldsymbol{w})$ as a substitute for MSE

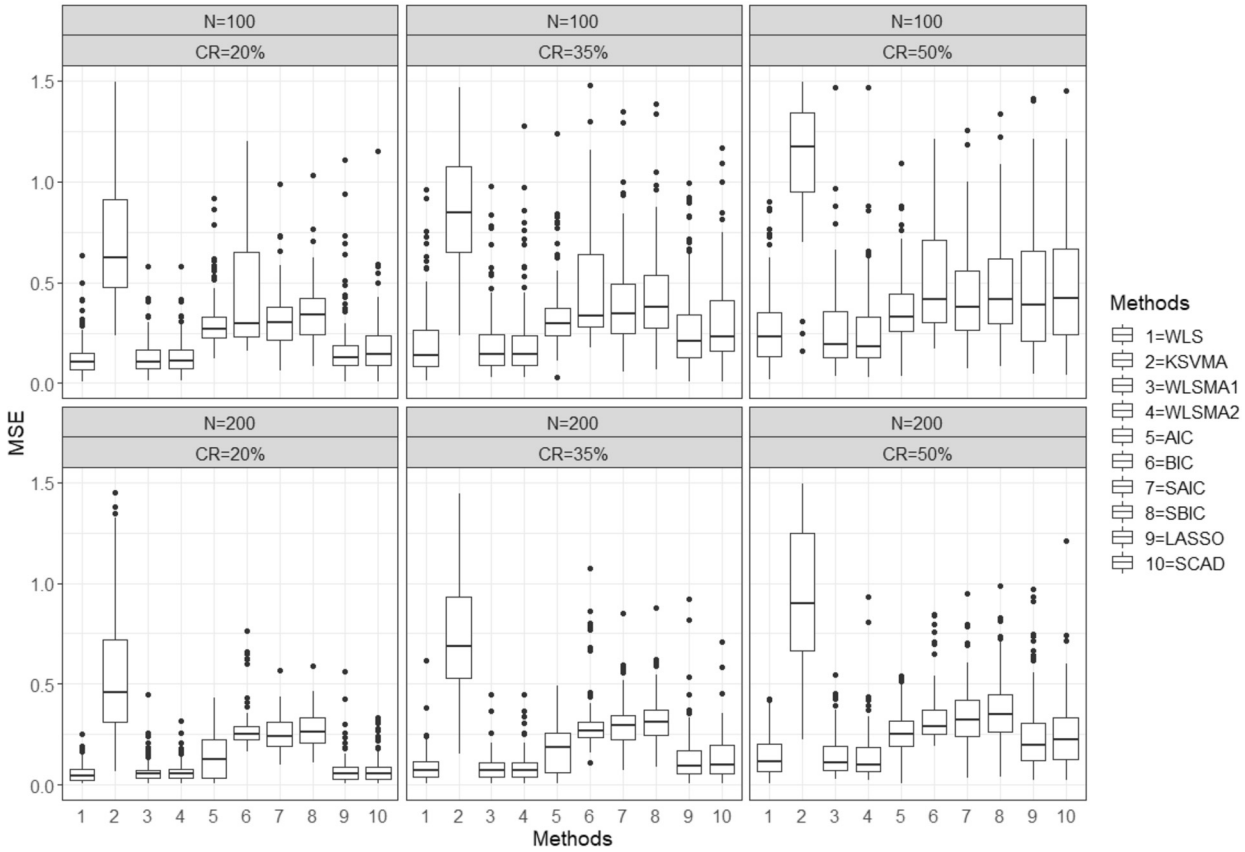

Fig. 6. Box-plots of MSEs in Case 3 with non-nested candidates.

Table 7
Means of $C_n(\boldsymbol{w})$ for Cox's model misspecified.

Method\CR	20%	35%	50%
WLS	45.7495	40.8451	34.9752
KSVMA	37.9957	55.6680	66.8343
WLSMA1	15.6698	14.2429	15.0889
WLSMA2	16.1551	14.2010	13.5421
AIC	22.6365	20.0556	17.0071
BIC	19.6554	17.0345	14.5335
SAIC	25.6805	22.9276	19.8533
SBIC	25.0765	22.3875	19.3988
LASSO	42.1024	36.9338	30.3962
SCAD	36.1555	32.2750	26.0745

to measure the predictive performance. For simplicity, only the results of Case 2 are shown here, where $\beta_j = \sqrt{2}/j$. It is shown in Table 7 and Fig. 8 that WLSMA still dominates most other methods.

5. Applications

In this section, we will evaluate the prediction performance of the proposed WLSMA method in a real dataset. The Mayo Clinic has established a dataset of 424 patients with primary biliary cirrhosis (PBC), which includes complete data on 17 covariates from 276 patients. The survival time of interest is the days between registration and death. Patients who underwent liver transplantation or were still alive at the end of the study were considered right censored. Since we can never know the real model, the assumption of model misspecification is quite reasonable in the empirical analysis. Following Huang et al. (2006)'s practice for data preprocessing, we first take log transformations to the covariates alkphos, bili, chol, copper, platelet, protime, ast, and trig and then standardize all continuous covariates. Discrete variables such as trt, sex, ascites, hepato, spiders, edema, and stage remain unchanged. A more detailed account of the PBC data can be found in Dickson et al. (1989). This dataset can be obtained from the package "survival" in R.

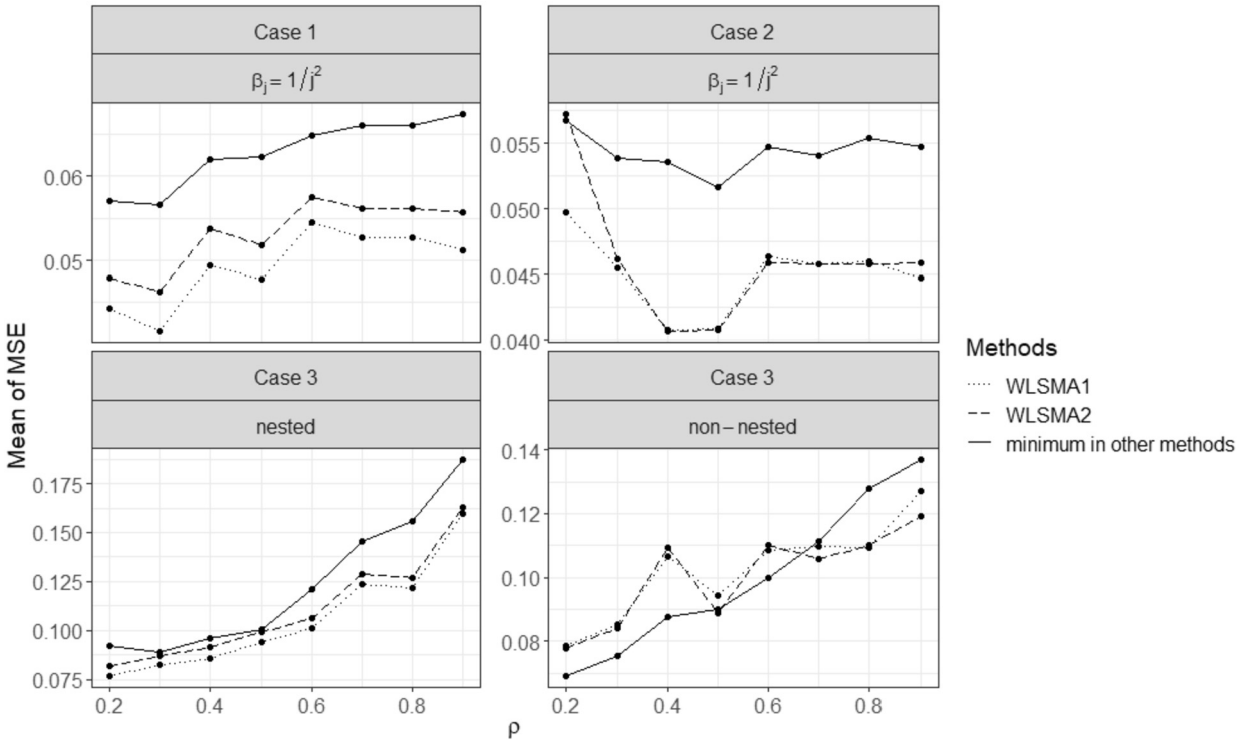

Fig. 7. Means of MSE with respect to ρ in three cases.

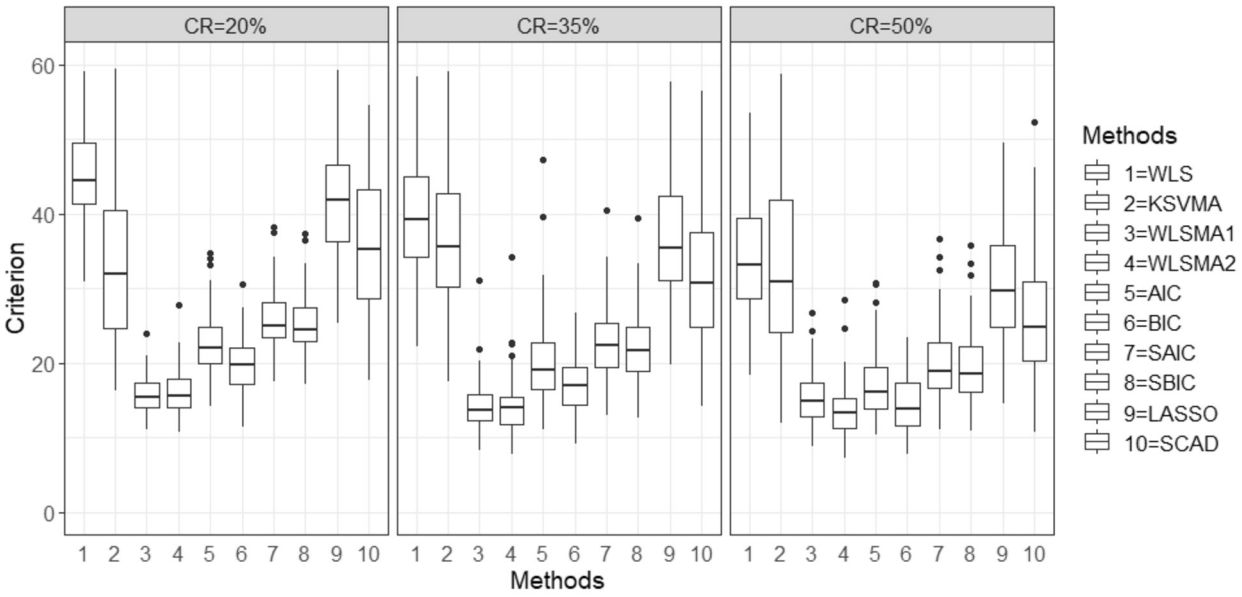

Fig. 8. Box-plots of $C_n(\boldsymbol{w})$ for Cox's model misspecified.

As for the construction of candidate models, we adopt two strategies. One is to use the solution path of the LARS algorithm proposed by Efron et al. (2004) to construct nested candidate models. The second is to refer to Huang et al. (2006)'s conclusion and use the six significant covariates they found to construct $2^6 - 1 = 63$ non-nested candidate models.

We randomly divide 276 observations into 70% training set and 30% validation set. The data in the training set are used to estimate the parameters. Then the prediction performance of WLSMA method is compared with other methods mentioned in the simulation section. Since the survival time of censored individuals is unknown, we use the mean squared error of

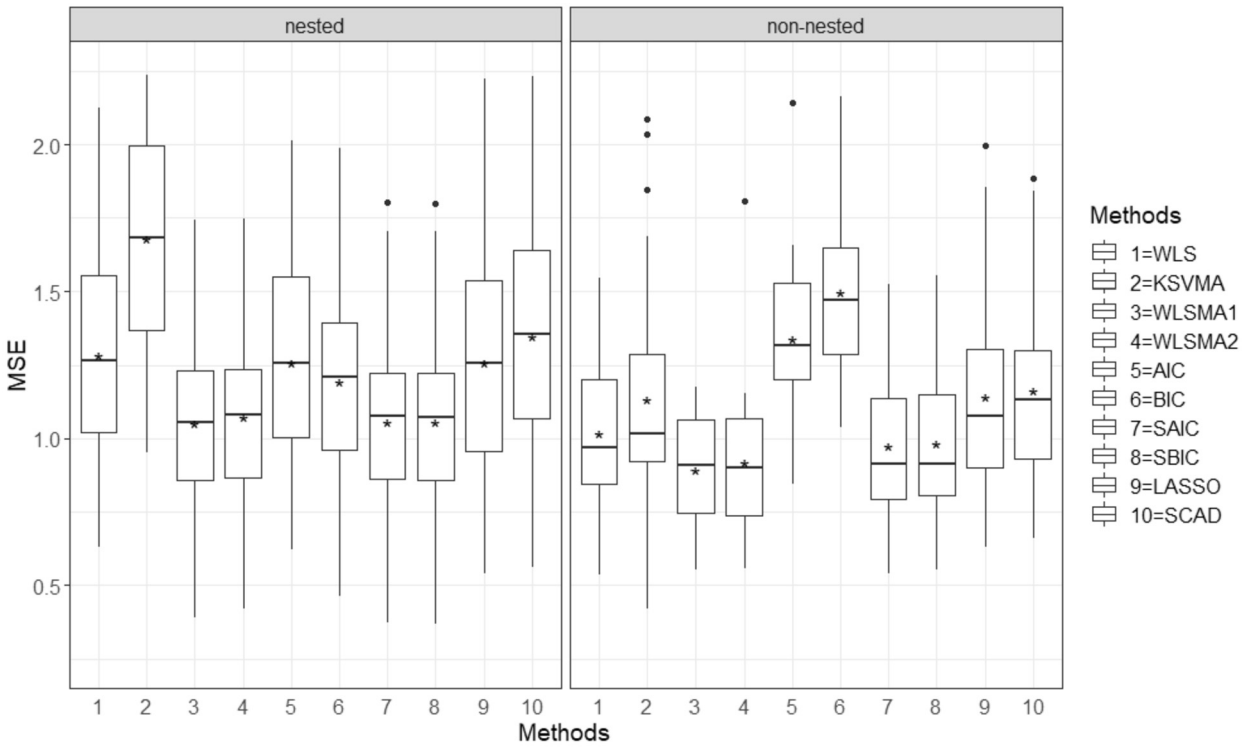

Fig. 9. Box-plots for MSEs of PBC dataset. The stars are the mean of MSEs and the points are the outliers.

uncensored individuals in the validation set to measure the prediction performance. The same experiment is repeated 100 times. We show the box-plots and the mean of MSEs of repeated experiments.

As can be seen from Fig. 9, when using nested candidate models, the prediction performance of WLSMA, SAIC, and SBIC model average methods is relatively close and better than other methods. When using non-nested candidate model, the prediction performance of WLSMA method is better than that of other methods.

6. Discussion

In order to overcome model selection uncertainty and improve prediction accuracy, we propose a WLS model average method based on the Mallows criterion for the AFT model with right censored data in this paper. Simulation results demonstrate the good performance of the proposed WLSMA method. In addition, the asymptotic optimality is also proved under certain mild conditions.

Note that although the proposed method does not require nested candidate models, the construction of candidate models is still a challenging problem. In empirical research, it is difficult to obtain an optimal ranking of variables in advance to construct nested candidate models. Moreover, when there are too many covariates, especially in the high-dimensional case ($p > n$), the possible combination of covariates will be a big computational burden to the prediction. In addition to ranking variables using the solution path of the LARS algorithm mentioned in this paper, various correlation coefficients can also be used to measure the importance of covariates to independent variables. A screening step prior to the model averaging procedure is also desirable.

In the preceding sections, the focus has been on the prediction of survival time. The asymptotic properties of the model averaging estimator of some parameters need further investigation in future research. Moreover, the Mallows model averaging method developed here is for right censored data under the AFT model assumption. It may be extended to other survival models or other types of censored data, such as the interval-censored data.

Appendix A

Proof of Theorem 1. As $R_n(\boldsymbol{w})$ can be written as $\|(I - P(\boldsymbol{w}))\boldsymbol{\mu}\|^2 + \sigma^2 \operatorname{tr}\{P(\boldsymbol{w})'P(\boldsymbol{w})\}$, we have

$$\begin{aligned} &\|(I - P(\boldsymbol{w}))\boldsymbol{\mu}\|^2 \le R_n(\boldsymbol{w}), \\ &\sigma^2 \operatorname{tr}\{P(\boldsymbol{w})'P(\boldsymbol{w})\} \le R_n(\boldsymbol{w}). \end{aligned} \tag{A.1}$$

Also we observe that

$$C_n(\boldsymbol{w}) = L_n(\boldsymbol{w}) + \|\boldsymbol{e}\|^2 + 2\langle \boldsymbol{e}, (I - P(\boldsymbol{w}))\boldsymbol{\mu}\rangle + 2\left(\sigma^2 \operatorname{tr}\{P(\boldsymbol{w})\} - \langle \boldsymbol{e}, P(\boldsymbol{w})\boldsymbol{e}\rangle\right).$$

Omit $\|\boldsymbol{e}\|^2$ and denote the last two terms on the right as $l_n(\boldsymbol{w})$, we can also write $\hat{\boldsymbol{w}}$ as $\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{H}_M}{\arg\min}(L_n(\boldsymbol{w}) + l_n(\boldsymbol{w}))$.

To prove Theorem 1, it is sufficient to show the following three conclusions hold:

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_M} \lambda_{\max}\left(P(\boldsymbol{w})P(\boldsymbol{w})'\right) < \infty, \tag{A.2}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} |l_n(\boldsymbol{w})/R_n(\boldsymbol{w})| \xrightarrow{p} 0, \tag{A.3}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} |L_n(\boldsymbol{w})/R_n(\boldsymbol{w}) - 1| \xrightarrow{p} 0. \tag{A.4}$$

First we prove (A.2). Define $Q_m = D^{1/2}Z_m\left[\left(D^{1/2}Z_m\right)' D^{1/2}Z_m\right]^{-1}\left(D^{1/2}Z_m\right)'$ which is an idempotent matrix. Then we can observe that

$$\begin{aligned}\lambda_{\max}\left(P(\boldsymbol{w})P(\boldsymbol{w})'\right)^{1/2} &= \left\|P(\boldsymbol{w})'\right\|_1 \le \sum_m w_m \left\|DZ_m\left(Z_m'D^{-1}Z_m\right)^{-1}Z_m'\right\|_1 \\ &= \sum_m w_m \left\|D^{1/2}Q_mD^{-1/2}\right\|_1,\end{aligned}$$

where $\|\cdot\|_1$ denotes the Banach norm. Using the submultiplicity of the Banach norm and the property of idempotent matrix yield

$$\begin{aligned}\left\|D^{1/2}Q_mD^{-1/2}\right\|_1^2 &\le \left\|D^{1/2}\right\|_1^2 \|Q_m\|_1^2 \left\|D^{-1/2}\right\|_1^2 \\ &= \lambda_{\max}(D)\,\lambda_{\max}(Q_m)\,\lambda_{\max}(D^{-1}) \\ &\le \left(\max_i d_i\right)\left(\min_i d_i\right)^{-1}.\end{aligned}$$

Since the diagonal elements of D are bounded, we have

$$\lambda_{\max}\left(P(\boldsymbol{w})P(\boldsymbol{w})'\right)^{1/2} \le \left(\max_i d_i\right)\left(\min_i d_i\right)^{-1} \le C_0,$$

where $C_l (l = 0, 1, 2, \dots)$ is a constant. Hence, (A.2) is proved.

Instead of proving (A.3), it is sufficient to prove:

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} |\langle \boldsymbol{e}, (I - P(\boldsymbol{w}))\boldsymbol{\mu}\rangle|/R_n(\boldsymbol{w}) \xrightarrow{p} 0, \tag{A.5}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} \left|\sigma^2 \operatorname{tr}\{P(\boldsymbol{w})\} - \langle \boldsymbol{e}, P(\boldsymbol{w})\boldsymbol{e}\rangle\right|/R_n(\boldsymbol{w}) \xrightarrow{p} 0. \tag{A.6}$$

For (A.5), we observe that for any $\delta > 0$,

$$\begin{aligned}&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M} |\langle \boldsymbol{e}, (I - P(\boldsymbol{w}))\boldsymbol{\mu}\rangle|/R_n(\boldsymbol{w}) > \delta\right\} \\ &\le C_1\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left\|(I - P\left(\boldsymbol{w}_m^0\right))\boldsymbol{\mu}\right\|^{2J} \\ &\le C_1\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left(R_n\left(\boldsymbol{w}_m^0\right)\right)^J,\end{aligned}$$

where the first inequality comes from the proof of Theorem 1 in Wan et al. (2010) and the last inequality comes from (A.1). When (C1) holds, we obtain (A.5). Similarly, for (A.6),

$$
\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M}\left|\sigma^2\operatorname{tr}\{P(\boldsymbol{w})\}-\langle\boldsymbol{e},P(\boldsymbol{w})\boldsymbol{e}\rangle\right|/R_n(\boldsymbol{w})>\delta\right\}\\
&\le C_2\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left[\operatorname{tr}\{P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)\}\right]^J\\
&\le C_2\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left(R_n\left(\boldsymbol{w}_m^0\right)\right)^J.
\end{aligned}
$$

Thus, (A.6) is obtained from (C1). To prove (A.4), since we have

$$
L_n(\boldsymbol{w})=R_n(\boldsymbol{w})+\|P(\boldsymbol{w})\boldsymbol{e}\|^2-\sigma^2\operatorname{tr}\{P(\boldsymbol{w})'P(\boldsymbol{w})\}-2\langle(I-P(\boldsymbol{w}))\boldsymbol{\mu},P(\boldsymbol{w})\boldsymbol{e}\rangle,
$$

we only need to show that

$$
\sup_{\boldsymbol{w}\in\mathcal{H}_M}\frac{\left|\|P(\boldsymbol{w})\boldsymbol{e}\|^2-\sigma^2\operatorname{tr}\{P(\boldsymbol{w})'P(\boldsymbol{w})\}\right|}{R_n(\boldsymbol{w})}\xrightarrow{p}0, \tag{A.7}
$$

$$
\sup_{\boldsymbol{w}\in\mathcal{H}_M}\frac{|\langle(I-P(\boldsymbol{w}))\boldsymbol{\mu},P(\boldsymbol{w})\boldsymbol{e}\rangle|}{R_n(\boldsymbol{w})}\xrightarrow{p}0. \tag{A.8}
$$

For (A.7), since $\|P(\boldsymbol{w})\boldsymbol{e}\|^2=\langle e,P(\boldsymbol{w})'P(\boldsymbol{w})e\rangle$, then by the same argument as for proving (A.5) and (A.6), we have

$$
\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M}\left|\|P(\boldsymbol{w})e\|^2-\sigma^2\operatorname{tr}\{P(\boldsymbol{w})'P(\boldsymbol{w})\}\right|/R_n(\boldsymbol{w})>\delta\right\}\\
&\le C_3\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left[\operatorname{tr}\{P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)\}\right]^J.
\end{aligned}
$$

Note that $P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)$ is a real symmetric matrix and (A.2), we have

$$
\begin{aligned}
&\operatorname{tr}\{P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)\}\\
&\le\lambda_{\max}\left(P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)\right)\operatorname{tr}\{P\left(\boldsymbol{w}_m^0\right)'P\left(\boldsymbol{w}_m^0\right)\}\\
&\le C_4R_n(\boldsymbol{w}_{\boldsymbol{m}}^{\boldsymbol{0}}).
\end{aligned}
$$

Hence, (A.7) is obtained from (C1). Similarly by $\langle(I-P(\boldsymbol{w}))\boldsymbol{\mu},P(\boldsymbol{w})\boldsymbol{e}\rangle=\langle P(\boldsymbol{w})'(I-P(\boldsymbol{w}))\boldsymbol{\mu},\boldsymbol{e}\rangle$, we have

$$
\begin{aligned}
&P\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_M}|\langle(I-P(\boldsymbol{w}))\boldsymbol{\mu},P(\boldsymbol{w})\boldsymbol{e}\rangle|/R_n(\boldsymbol{w})>\delta\right\}\\
&\le C_5\delta^{-2J}\xi_n^{-2J}\sum_{m=1}^{M}\left\|P(\boldsymbol{w})'(I-P(\boldsymbol{w}))\boldsymbol{\mu}\right\|^{2J}.
\end{aligned}
$$

Then by

$$
\begin{aligned}
\left\|P(\boldsymbol{w})'(I-P(\boldsymbol{w}))\boldsymbol{\mu}\right\|^2&\le\lambda_{\max}\left(P(\boldsymbol{w})'P(\boldsymbol{w})\right)\|(I-P(\boldsymbol{w}))\boldsymbol{\mu}\|^2\\
&\le C_6R_n(\boldsymbol{w}),
\end{aligned}
$$

(A.8) holds. Then we complete the proof of Theorem 1. □

Before giving the proof of Theorem 2, we need two lemmas:

Lemma 1. *Assume (C3)-(C6) hold. Then we have*

$$
\sup_i\left|\hat{d}_i-d_i\right|=O_p\left(\frac{1}{\sqrt{n}}\right) \tag{A.9}
$$

$$
\sup_{\boldsymbol{w}\in\mathcal{H}_M}\left\|\hat{\boldsymbol{\mu}}(\boldsymbol{w})-\hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|=O_p(\sqrt{k_{m^*}}) \tag{A.10}
$$

$$\sup_m \left\| \left(\frac{1}{n} Z_m' D Z_m\right)^{-1} - \left(\frac{1}{n} Z_m' \hat{D} Z_m\right)^{-1} \right\|_1 = O_p(\frac{k_{m^*}}{\sqrt{n}}) \tag{A.11}$$

$$\sup_m \left\| \left(\frac{1}{n} Z_m' \hat{D} Z_m\right)^{-1} \right\|_1 = O_p(1). \tag{A.12}$$

These are conclusions from Zhou (1992) and Liu et al. (2016). Their assumptions are similar to ours so we can get these conclusions directly. The details are omitted here.

Lemma 2. *Assume (C3)-(C6) hold. Then*

$$\sup_m \left\| P_m - P_{\hat{G},m} \right\|^2 = O_p(\frac{k_{m^*}^2}{n}). \tag{A.13}$$

Proof of Lemma 2.

$$\begin{aligned}
&\sup_m \left\| P_m - P_{\hat{G},m} \right\|^2 \\
=&\sup_m \left\| Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' D - Z_m \left(Z_m' \hat{D} Z_m\right)^{-1} Z_m' \hat{D} \right\|^2 \\
=&\sup_m \left\| Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' D - Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' \hat{D} + Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' \hat{D} - Z_m \left(Z_m' \hat{D} Z_m\right)^{-1} Z_m' \hat{D} \right\|^2 \\
\le&\sup_m \left\| Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' (D - \hat{D}) \right\|^2 + \sup_m \left\| Z_m \left[\left(Z_m' D Z_m\right)^{-1} - \left(Z_m' \hat{D} Z_m\right)^{-1} \right] Z_m' \hat{D} \right\|^2 .
\end{aligned}$$

According to (A.9), (A.11), (A.12), (C5), (C6) and the inequality $\|AB\| \le \|A\|_1 \|B\|$, we have:

$$\begin{aligned}
&\sup_m \left\| Z_m \left(Z_m' D Z_m\right)^{-1} Z_m' (D - \hat{D}) \right\|^2 \\
\le&\sup_m \left\| \left(\frac{1}{n} Z_m' \hat{D} Z_m\right)^{-1} \right\|_1^2 \sup_m \left\| \frac{1}{n} Z_m Z_m' \right\|^2 \max_i \left\| d_i - \hat{d}_i \right\|^2 \\
=&O_p(1) O_p(1) O_p(\frac{1}{n}) \\
=&O_p(\frac{1}{n})
\end{aligned}$$

and

$$\begin{aligned}
&\sup_m \left\| Z_m \left[\left(Z_m' D Z_m\right)^{-1} - \left(Z_m' \hat{D} Z_m\right)^{-1} \right] Z_m' \hat{D} \right\|^2 \\
\le&\sup_m \left\| \left(\frac{1}{n} Z_m' D Z_m\right)^{-1} - \left(\frac{1}{n} Z_m' \hat{D} Z_m\right)^{-1} \right\|_1^2 \sup_m \left\| \frac{1}{n} Z_m Z_m' \right\|^2 \max_i \{\hat{d}_i^2\} \\
=&O_p(\frac{k_{m^*}^2}{n}) O_p(1) O_p(1) \\
=&O_p(\frac{k_{m^*}^2}{n}).
\end{aligned}$$

Then we complete the proof of Lemma 2. □

Proof of Theorem 2. We can write $C_{\hat{G}}(\boldsymbol{w})$ as:

$$C_{\hat{G}}(\boldsymbol{w}) = L_{\hat{G}}(\boldsymbol{w}) + \|\boldsymbol{e}\|^2 + 2\langle e, (I - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\rangle + 2\left(\sigma^2 \operatorname{tr}\{P_{\hat{G}}(\boldsymbol{w})\} - \langle \boldsymbol{e}, P_{\hat{G}}(\boldsymbol{w})\boldsymbol{e}\rangle\right).$$

Based on the proof of Theorem 1, we only need to show:

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} \left| L_{\hat{G}}(\boldsymbol{w}) / R_n(\boldsymbol{w}) - 1 \right| \xrightarrow{p} 0, \tag{A.14}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} |\langle \boldsymbol{e}, (I - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\rangle| / R_n(\boldsymbol{w}) \xrightarrow{p} 0, \tag{A.15}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_M} \left|\sigma^2 \operatorname{tr}\{P_{\hat{G}}(\boldsymbol{w})\} - \langle \boldsymbol{e}, P_{\hat{G}}(\boldsymbol{w})\boldsymbol{e}\rangle\right| / R_n(\boldsymbol{w}) \xrightarrow{p} 0. \tag{A.16}$$

We first consider (A.14). Note that

$$\begin{aligned}
&\left|\frac{L_{\widehat{G}}(\boldsymbol{w})}{R_n(\boldsymbol{w})} - 1\right| \\
=&\left|\frac{L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})} - 1 + \frac{L_{\widehat{G}}(\boldsymbol{w}) - L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})}\right| \\
\le&\left|\frac{L_n(\boldsymbol{w})}{R_n(\boldsymbol{w})} - 1\right| + \left|\frac{\left\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{\widehat{G}}(\boldsymbol{w})\right\|^2 - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2}{R_n(\boldsymbol{w})}\right|.
\end{aligned} \tag{A.17}$$

By (A.4), we have already known that the first term of (A.17) converges to 0 in probability. For the second term,

$$\begin{aligned}
&\left|\frac{\left\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|^2 - \left\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right\|^2}{R_n(\boldsymbol{w})}\right| \\
=&\left|\frac{2(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}))'(\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})) + \left\|\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|^2}{R_n(\boldsymbol{w})}\right|
\end{aligned}$$

Since (A.4) and (A.10),

$$\begin{aligned}
&\left|\frac{(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w}))'(\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w}))}{R_n(\boldsymbol{w})}\right| \\
\le&\left(\frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\|^2}{R_n(\boldsymbol{w})}\right)^{1/2}\left(\frac{\left\|\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|^2}{R_n(\boldsymbol{w})}\right)^{1/2} \\
=&o_p(1),
\end{aligned}$$

(A.14) holds. Next we prove (A.15). Note that we have (A.1) and (A.5), we only need to follow the proof of (A.5) and show

$$\frac{\left\|(I - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\right\|^2}{R_n(\boldsymbol{w})} = O_p(1). \tag{A.18}$$

Since

$$\begin{aligned}
&\frac{\left\|(I - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\right\|^2}{R_n(\boldsymbol{w})} \\
\le&\frac{\|(I - P(\boldsymbol{w}))\boldsymbol{\mu}\|^2 + \left\|(P(\boldsymbol{w}) - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\right\|^2}{R_n(\boldsymbol{w})} \\
\le&1 + \frac{\left\|(P(\boldsymbol{w}) - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{\mu}\right\|^2}{\xi_n} \\
=&1 + \frac{\left\|(P(\boldsymbol{w}) - P_{\hat{G}}(\boldsymbol{w}))(Y - \boldsymbol{e})\right\|^2}{\xi_n} \\
\le&1 + \frac{\left\|\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|^2 + \left\|(P(\boldsymbol{w}) - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{e}\right\|^2}{\xi_n}
\end{aligned}$$

According to (A.10), (A.13), (C1) and (C6), it follows that

$$\frac{\left\|\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \hat{\boldsymbol{\mu}}_{\hat{G}}(\boldsymbol{w})\right\|^2}{\xi_n} = o_p(1)$$

and

$$\frac{\left\|(P(\boldsymbol{w}) - P_{\hat{G}}(\boldsymbol{w}))\boldsymbol{e}\right\|^2}{\xi_n} = \frac{O_p(\frac{k_{m^*}^2}{n})O_p(n)}{\xi_n} = O_p(1).$$

Then we get (A.18) and therefore (A.15) holds. For (A.16), we also follow the proof of (A.6). All we need to prove is that for any m,

$$\frac{\mathrm{tr}\{P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)' P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)\}}{R_n\left(\boldsymbol{w}_m^0\right)} = O_p(1). \tag{A.19}$$

Let $\hat{Q}_m = \hat{D}^{1/2} Z_m \left[\left(\hat{D}^{1/2} Z_m\right)' \hat{D}^{1/2} Z_m\right]^{-1} \left(\hat{D}^{1/2} Z_m\right)'$. It is easy to observe that

$$\begin{aligned}
&\mathrm{tr}\{P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)' P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)\} \\
=&\mathrm{tr}\{\hat{D}^{1/2}\hat{Q}_m\hat{D}^{-1}\hat{Q}_m\hat{D}^{1/2}\} \\
\le&\left(\max_i \hat{d}_i\right)\left(\min_i \hat{d}_i\right)^{-1}\mathrm{tr}\{\hat{Q}_m\} \\
=&O_p(k_m).
\end{aligned}$$

By (C6), we further obtain

$$\sup_m \frac{\mathrm{tr}\{P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)' P_{\hat{G}}\left(\boldsymbol{w}_m^0\right)\}}{R_n\left(\boldsymbol{w}_m^0\right)} = O_p(\frac{k_{m^*}}{\xi_n}).$$

Then, (A.19) holds and therefore (A.16) holds. Then we complete the proof of Theorem 2. □

Proof of Theorem 3. Notice that $\hat{C}_{\hat{G}}(\boldsymbol{w}) = C_{\hat{G}}(\boldsymbol{w}) + 2(\sigma^2 - \hat{\sigma}^2)\,\mathrm{tr}\{P_{\hat{G}}(\boldsymbol{w})\}$. Hence, from the result of Theorem 2, it suffices to prove that

$$\sup_{w\in\mathcal{H}_n}\left|\hat{\sigma}^2 - \sigma^2\right|\mathrm{tr}\{P_{\hat{G}}(\boldsymbol{w})\}/R_n(\boldsymbol{w}) \xrightarrow{p} 0.$$

Note that

$$\begin{aligned}
&\sup_{w\in\mathcal{H}_n}\frac{\mathrm{tr}\{P_{\hat{G}}(\boldsymbol{w})\}}{R_n(\boldsymbol{w})}\left|\hat{\sigma}^2 - \sigma^2\right| \\
\le&\frac{k_{m^*}}{\xi_n}\left|\hat{\sigma}^2 - \sigma^2\right| \\
=&\frac{k_{m^*}}{\xi_n}\left|\frac{Y'\left(I - P_{\hat{G},m^*}\right)'\left(I - P_{\hat{G},m^*}\right)Y}{n - k_{m^*}} - \sigma^2\right| \\
\le&\frac{k_{m^*}}{n - k_{m^*}}\frac{\boldsymbol{\mu}'\left(I - P_{\hat{G},m^*}\right)'\left(I - P_{\hat{G},m^*}\right)\boldsymbol{\mu}}{\xi_n} + \frac{2k_{m^*}\left|\boldsymbol{\mu}'\left(I - P_{\hat{G},m^*}\right)'\left(I - P_{\hat{G},m^*}\right)\boldsymbol{e}\right|}{\xi_n\left(n - k_{m^*}\right)} \\
&+\frac{k_{m^*}\left|\boldsymbol{e}'\left(I - P_{\hat{G},m^*}\right)'\left(I - P_{\hat{G},m^*}\right)\boldsymbol{e} - \sigma^2\left(n - k_{m^*}\right)\right|}{\xi_n\left(n - k_{m^*}\right)}.
\end{aligned} \tag{A.20}$$

The proof to show (A.20) converges to 0 in probability follows the proof of Theorem 2 in Wan et al. (2010). To be specific, from (C1) and (A.1) we have

$$\frac{\boldsymbol{\mu}'\left(I - P_{\hat{G},m^*}\right)'\left(I - P_{\hat{G},m^*}\right)\boldsymbol{\mu}}{\xi_n} \to 0. \tag{A.21}$$

Also we notice that

$$\begin{aligned}
\|\left(I - P_{\hat{G},m^*}\right)\boldsymbol{\mu}\|^2 &\le \lambda_{\max}^2\left(I - P_{\hat{G},m^*}\right)\|\boldsymbol{\mu}\|^2 \\
&\le \left(\lambda_{\max}(I) + \lambda_{\max}\left(P_{\hat{G},m^*}\right)\right)^2\|\boldsymbol{\mu}\|^2 \\
&= C_7\|\boldsymbol{\mu}\|^2.
\end{aligned}$$

Then by using (A.21) and conditions (C5) and (C6), we obtain that as $n\to\infty$,

$$\frac{k_{m^*}}{n-k_{m^*}}\frac{\boldsymbol{\mu}'\left(I-P_{\hat{G},m^*}\right)'\left(I-P_{\hat{G},m^*}\right)\boldsymbol{\mu}}{\xi_n} \le\left[\frac{k_{m^*}^2}{n-k_{m^*}}\frac{\boldsymbol{\mu}'\left(I-P_{\hat{G},m^*}\right)'\left(I-P_{\hat{G},m^*}\right)\boldsymbol{\mu}}{\xi_n^2}\frac{C_7\|\boldsymbol{\mu}\|^2}{n-k_{m^*}}\right]^{1/2}\to 0. \tag{A.22}$$

Moreover, using the same techniques as Wan et al. (2010), we observe that, for any $\delta>0$, as $n\to\infty$,

$$P\left\{\frac{2k_{m^*}\left|\boldsymbol{\mu}'\left(I-P_{\hat{G},m^*}\right)'\left(I-P_{\hat{G},m^*}\right)\boldsymbol{e}\right|}{\xi_n\left(n-k_{m^*}\right)}>\delta\right\} \le\frac{C_8 4k_{m^*}^2\boldsymbol{\mu}'\left(I-P_{\hat{G},m^*}\right)'\left(I-P_{\hat{G},m^*}\right)\boldsymbol{\mu}}{\delta^2\xi_n^2\left(n-k_{m^*}\right)^2}\to 0, \tag{A.23}$$

and

$$P\left\{\frac{k_{m^*}\left|\boldsymbol{e}'\left(I-P_{\hat{G},m^*}\right)'\left(I-P_{\hat{G},m^*}\right)\boldsymbol{e}-\sigma^2\left(n-k_{m^*}\right)\right|}{\xi_n\left(n-k_{m^*}\right)}>\delta\right\} \le\frac{C_9k_{m^*}^2\left(n-k_{m^*}\right)}{\delta^2\xi_n^2\left(n-k_{m^*}\right)^2}\to 0. \tag{A.24}$$

Consequently, by combining (A.22)-(A.24), we obtain that (A.20) converges to 0 in probability. Then we complete the proof of Theorem 3. □

References

Bao, Y., He, S., Mei, C., 2007. The Koul–Susarla–Van Ryzin and weighted least squares estimates for censored linear regression model: a comparative study. Comput. Stat. Data Anal. 51 (12), 6488–6497.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53 (2), 603–618.

Buckley, J., James, I., 1979. Linear regression with censored data. Biometrika 66 (3), 429–436.

Cheng, C., Feng, X., Huang, J., Jiao, Y., Zhang, S., 2022. ℓ_0-regularized high-dimensional accelerated failure time model. Comput. Stat. Data Anal. 107430.

Dai, L., Chen, K., Sun, Z., Liu, Z., Li, G., 2018. Broken adaptive ridge regression and its asymptotic properties. J. Multivar. Anal. 168 (C), 334–351.

Dickson, E.R., Grambsch, P.M., Fleming, T.R., Fisher, L.D., Langworthy, A., 1989. Prognosis in primary biliary cirrhosis: model for decision making. Hepatology 10 (1), 1–7.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.

Hansen, B.E., Racine, J.S., 2012. Jackknife model averaging. J. Econom. 167 (1), 38–46.

He, B., Liu, Y., Wu, Y., Yin, G., Zhao, X., 2020. Functional martingale residual process for high-dimensional Cox regression with model averaging. J. Mach. Learn. Res. 21, 1–37.

He, S., Huang, X., 2003. Central limit theorem of linear regression model under right censorship. Sci. China Ser. A, Math. 46 (5), 600–610.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Am. Stat. Assoc. 98 (464), 879–899.

Hu, J., Chai, H., 2013. Adjusted regularized estimation in the accelerated failure time model with high dimensional covariates. J. Multivar. Anal. 122 (C), 96–114.

Huang, J., Ma, S., Xie, H., 2006. Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 62 (3), 813–820.

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

Kalbfleisch, J.D., Prentice, R.L., 2011. The Statistical Analysis of Failure Time Data. John Wiley and Sons.

Koul, H., Susarla, V., Van Ryzin, J., 1981. Regression analysis with randomly right-censored data. Ann. Stat. 9 (6), 1276–1288.

Li, J., Yu, T., Lv, J., Lee, M.L.T., 2021. Semiparametric model averaging prediction for lifetime data via hazards regression. J. R. Stat. Soc., Ser. C 70 (5), 1187–1209.

Li, K.C., 1986. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Ann. Stat. 14 (3), 1101–1112.

Liang, Z.Q., Chen, X.L., Zhou, Y.Q., 2022. Mallows model averaging estimation for linear regression model with right censored data. Acta Math. Appl. Sin. Engl. Ser. 38 (1), 5–23.

Liu, Q., Okui, R., Yoshimura, A., 2016. Generalized least squares model averaging. Econom. Rev. 35 (8–10), 1692–1752.

Lv, J., Fan, Y., 2009. A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. 37 (6A), 3498–3528.

Miller, R., 1976. Least square regression with censored data. Biometrika 63 (3), 449–464.

Stute, W., 1993. Consistent estimation under random censorship when covariables are present. J. Multivar. Anal. 45 (1), 89–103.

Stute, W., 1996. Distributional convergence under random censorship when covariables are present. Scand. J. Stat. 23 (4), 461–471.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B 58 (1), 267–288.
Wan, A.T., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. J. Econom. 156 (2), 277–283.
Wang, S., Nan, B., Zhu, J., Beer, D.G., 2008. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics 64 (1), 132–140.
Yan, X., Wang, H., Wang, W., Xie, J., Ren, Y., Wang, X., 2021. Optimal model averaging forecasting in high-dimensional survival analysis. Int. J. Forecast. 37 (3), 1147–1155.
Zhou, M., 1992. Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. Ann. Stat. 20 (2), 1002–1021.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101 (476), 1418–1429.