Conditional screening for ultrahigh-dimensional survival data in case-cohort studies

Jing Zhang · Haibo Zhou · Yanyan Liu · Jianwen Cai

Received: 23 July 2020 / Accepted: 5 August 2021 / Published online: 20 August 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
The case-cohort design has been widely used to reduce the cost of covariate measurements in large cohort studies. In many such studies, the number of covariates is very large, and the goal of the research is to identify active covariates which have great influence on response. Since the introduction of sure independence screening, screening procedures have achieved great success in terms of effectively reducing the dimensionality and identifying active covariates. However, commonly used screening methods are based on marginal correlation or its variants, they may fail to identify hidden active variables which are jointly important but are weakly correlated with the response. Moreover, these screening methods are mainly proposed for data under the simple random sampling and can not be directly applied to case-cohort data. In this paper, we consider the ultrahigh-dimensional survival data under the case-cohort design, and propose a conditional screening method by incorporating some important prior known information of active variables. This method can effectively detect hidden active variables. Furthermore, it possesses the sure screening property under some mild regularity conditions and does not require any complicated numerical optimization. We evaluate the finite sample performance of the proposed method via extensive
simulation studies and further illustrate the new approach through a real data set from patients with breast cancer.

**Keywords**  Case-cohort design · Conditional screening · Sure screening property · Survival data · Ultrahigh-dimensional data · Weighted estimating equation

### 1 Introduction

In large epidemiological cohort studies, it is common that some diseases of interest (e.g., cancer, heart disease, HIV infection) have very low incidence. In addition, some exposures can be very expensive to measure and it is not feasible to obtain the measures on all cohort members due to restrictions on resources. To reduce the cost while keeping as much efficiency as possible, Prentice (1986) proposed the case-cohort design, where the expensive covariates are obtained only for a random sample of the full cohort, called the subcohort, as well as the additional cases who have experienced the event of interest during the follow-up period. When covariate dimension $p$ is smaller than sample size $n$, various methods have been proposed for analyzing data under this design, such as the pseudo-likelihood approach (Prentice 1986; Self and Prentice 1988; Kalbfleisch and Lawless 1988), the estimating equation method (Chen and Lo 1999; Chen 2001), the multiple imputation approach (Marti and Chavance 2011; Keogh and White 2013), the maximum likelihood estimation (Scheike and Martinussen 2004; Zeng and Lin 2014), weighted estimating equation approach (Barlow 1994; Borga et al. 2000; Kulich and Lin 2004; Breslow and Wellner 2007; Kang and Cai 2009; Kim et al. 2013), among others.

With the rapid development of biomedical technology, high-dimensional data are frequently collected in large epidemiological studies. The feature of this kind of data is that the covariate dimension $p$ is much larger than sample size $n$. An important purpose of analyzing this type of data is to identify a subset of covariates related to the event of interest and construct the effective models based on the selected covariates. For scenarios where $p$ increases with $n$ at polynomial rate (e.g., $p = n^\alpha$ with $\alpha > 0$), the regularization method has been demonstrated to be an effective dimension reduction method for simple random sampling (SRS) data (e.g., Tibshirani 1996; Fan and Li 2001; Zou 2006; Candes and Tao 2007; Zhang 2010) and has been generalized to high-dimensional data under the case-cohort design recently. For example, Ni et al. (2016) proposed a variable selection procedure by using the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) for scenarios where $p$ increases at a slower rate than $n$. Kim and Ahn (2019) proposed a bi-level variable selection method to select non-zero group and within-group variables for cases where variables have group structure. These methods can select variables and estimate parameters simultaneously, however, the computation inherent in regularization methods makes them involve the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability when the dimension $p$ is ultrahigh in the sense that $p = \exp(n^\alpha)$ with $\alpha > 0$ (Fan et al. 2009).

For SRS data, the feature screening method has achieved great success in dealing with the challenge of ultrahigh-dimensional settings. Various marginal screening
methods have been proposed under different settings, such as linear models (Fan and Lv 2008), generalized linear models (Fan and Song 2010), additive models (Fan et al. 2011), the varying coefficient models (Fan et al. 2014; Liu et al. 2014) and model-free scenarios (e.g., Zhu et al. 2011; Li et al. 2012a, b; He et al. 2013; Chang et al. 2013; Cui et al. 2015; Mai and Zou 2015; Wu and Yin 2015). For censored survival data, several model-based screening methods (e.g., Tibshirani 2009; Zhao and Li 2012; Gorst-Rasmussen and Scheike 2013) and model-free screening methods (e.g., Song et al. 2014; Wu and Yin 2015; Zhang et al. 2017; Zhou and Zhu 2017; Lin et al. 2018; Zhang et al. 2018; Liu et al. 2018; Pan et al. 2019) have been proposed via defining different marginal utilities. Although they are powerful in reducing the dimensionality, they may face some challenges in some situations. For instance, as noted in Fan and Lv (2008), the correlation among covariates heavily influence the marginal utility. When the correlation among covariates is relatively high, the marginal screening methods may fail to retain the hidden active variables which have great influence on response but are weakly correlated with the response. Although some iterative screening methods (e.g., Fan and Lv 2008; Zhu et al. 2011; Zhang et al. 2018; Pan et al. 2019) and forward screening approaches have been proposed to alleviate this problem, the computation speed is relatively slow and the statistical properties are elusive.

In many applications, researchers can obtain some prior information of active variables from previous investigations and experiences. For example, in the breast cancer study (van de Vijver et al. 2002), gene AL080059 has been known to be predictive to patients’ survival time in the literature (Yeung et al. 2005; van’t Veer et al. 2002). Barut et al. (2016) pointed out we can improve the accuracy in variable screening by including such prior knowledge. In view of this thought, they proposed the conditional screening approach for generalized linear models and showed that conditioning helps reducing the correlation among covariates, thus can detect the hidden active variables with higher probability. Hong et al. (2016) further proposed to integrate prior information using data-driven approaches. Hu and Lin (2017) put forward a conditional screening procedure via ranking covariates based on conditional marginal empirical likelihood ratios. Liu and Wang (2018) proposed a screening method based on conditional distance correlation. Hong et al. (2018) developed a conditional screening method for censored data under the proportional hazards model. Liu and Chen (2018) considered the conditional quantile independence screening approach for ultrahigh-dimensional heterogeneous data. Lu and Lin (2020) proposed a model-free conditional screening via conditional distance correlation. Extensive simulation studies showed these conditional screening methods which incorporate important prior information of active variables can provide a powerful means to identify hidden active variables for ultrahigh-dimensional data.

The research on marginal and conditional screening methods has been fruitful for ultrahigh-dimensional SRS data, but to the best of our knowledge, conditional screening method has not been studied for case-cohort data, the existing conditional screening methods can not be directly applied to the case-cohort data due to its special data structure. To fill the gap, we propose a conditional screening method for ultrahigh-dimensional case-cohort data under the framework of Cox proportional hazards model. We construct the marginal hazards regression models for each covariate by including the known important covariates. As some covariates are not fully observed, we build
the weighted estimating equation to obtain the estimators of the parameters. Then we propose the marginal utilities based on the parameter estimates to measure the contribution of each covariate and retain the covariates with top ranked contributions. We refer to it as conditional weighted screening method, in short the CWSIS procedure. As pointed out by Barut et al. (2016), the correlation between covariates can be weakened upon conditioning, so that hidden active covariates have a higher chance to be retained. Therefore, the proposed method enables the detection of hidden active covariates for ultrahigh dimensional survival data under the case-cohort design. Under some reasonable conditions, it enjoys the sure screening property and the ranking consistency. Our research is the first one that focus on conditional screening for ultrahigh dimensional case-cohort data, it can be viewed as an extension of (Hong et al. 2018) from SRS data to case-cohort data. Note that although the ideas are similar, the generalization is quite challenging due to the much more complex structure of case-cohort data, both implementation and the theory will be quite different.

The rest of the article is organized as follows. In Sect. 2, we introduce the model, data and present the details of the CWSIS procedure. In Sect. 3, we establish the theoretical properties of the proposed CWSIS method. Section 4 presents results from simulation studies. A real data set from the breast cancer study is analyzed in Sect. 5. Section 6 provides some remarks and discussions. The regularity conditions and the technical proofs are presented in the Appendix.

2 Conditional screening for case-cohort data

Suppose there are $n$ independent subjects in a cohort study. Let $T_i$ and $C_i$ denote the failure time and censoring time of subject $i$, we only observe $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ due to right-censoring. Let $Z_i = (Z_{i1}, \ldots, Z_{ip})^T$ denote the $p$-dimensional covariate, under the case-cohort design, $Z_i$ is available only on the cases ($\Delta_i = 1$) and the subcohort (a random subset of the full cohort). Let $\xi_i$ be the indicator for subcohort membership, i.e., $\xi_i = 1$ and 0 denote whether or not the $i$th subject in the full cohort is selected into the subcohort. For the selection of subcohort, we consider independent Bernoulli sampling with selection probability $\pi = Pr(\xi_i = 1) \in (0, 1)$. Thus, the observable data for the $i$th subject is $\{X_i, \Delta_i, Z_i, \xi_i\}$ when $\xi_i = 1$ or $\Delta_i = 1$, and $\{X_i, \Delta_i, \xi_i\}$ when $\xi_i = 0$ and $\Delta_i = 0$.

Suppose that the failure time follows the proportional hazards model (Cox 1972), under which the conditional hazard function of $T_i$ given $Z_i$ has the form

$$\lambda(t \mid Z_i) = \lambda_0(t) \exp(\alpha^T Z_i),$$

(1)

where $\lambda_0(t)$ is the unspecified baseline hazard function and $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ is the unknown regression parameter. Assume that the failure time $T_i$ and the censoring time $C_i$ are independent given $Z_i$. In an ultrahigh-dimensional setting, the dimensionality $p$ greatly exceeds sample size $n$ and can be allowed to increase at an exponential rate of $n$. Under the sparsity principle, only a small number of covariates have great influence on the response variable, i.e., $\|\alpha\|$ is much smaller than $p$, where $\|\alpha\|$ denotes the number of nonzero elements of $\alpha$. Assume we have the prior information that a set of covariates
are related to survival time $T$ and the index set is denoted by $C$, $q = |C|$ denotes the number of covariates in $C$. Write $Z_i,C = (Z_{i,j}, j \in C)$, $Z_i,\neg C = (Z_{i,j}, j \notin C)$, $\alpha_C = (\alpha_j, j \in C)$ and $\alpha,\neg C = (\alpha_j, j \notin C)$. Here, $C$ is known, $\alpha_C$ and $\alpha,\neg C$ are unknown. The true hazard function in (1) is equivalent to

$$\lambda(t|Z_i) = \lambda_0(t) \exp(\alpha_C^T Z_i,C + \alpha_{\neg C}^T Z_i,\neg C).$$  \hspace{1cm} (2)$$

Let $A,\neg C = \{ j \notin C : \alpha_j \neq 0 \}$ and $a = |A,\neg C| = \sum_{j \notin C} I(\alpha_j \neq 0)$ be the true set of non-zero coefficients and its cardinality. Our goal is to recover the set $A,\neg C$ as precisely as possible based on data from case-cohort studies. In other words, we want to find a subset of covariates $\hat{A},\neg C$ which satisfies $A,\neg C \subseteq \hat{A},\neg C$.

To perform an initial screening procedure, we construct the marginal Cox regression models for each covariate individually, here we also add the known covariates in $C$ to each marginal model. Specifically, for $j \notin C$, the hazard function of $T_i$ given $(Z_i,C, Z_{ij})$ has the form

$$\lambda(t|Z_i,C, Z_{ij}) = \lambda_{j,0}(t) \exp(\beta_{C,j}^T Z_{i,C} + \beta_j Z_{ij}),$$  \hspace{1cm} (3)$$

where $\lambda_{j,0}(t)$ is the unspecified baseline hazard function, and $\beta_{C,j}$ and $\beta_j$ are the unknown regression parameters corresponding to covariates $Z_C$ and $Z_j$ in the marginal Cox model, respectively. Since the covariates can only be observed for the selected subcohort and cases for case-cohort data, we consider the following weighted estimating equation

$$U_j(\beta_{C,j}, \beta_j) = [U_{j,k}(\beta_{C,j}, \beta_j), k \in C \cup \{j\}]^T = 0_{q+1},$$  \hspace{1cm} (4)$$

with

$$U_{j,k}(\beta_{C,j}, \beta_j) = \sum_{i=1}^n \int_0^\tau \left\{ Z_{ik} - \frac{\tilde{S}_{j,k}(\beta_{C,j}, \beta_j, t)}{\tilde{S}_{j,k}(\beta_{C,j}, \beta_j, t)} \right\} dN_i(t) = 0,$$

where $\tilde{S}_{j,k}^{(l)}(\beta_{C,j}, \beta_j, t) = n^{-1} \sum_{i=1}^n Z_{ik}^l w_i(t) Y_i(t) \exp(\beta_{C,j}^T Z_{i,C} + \beta_j Z_{ij})$ for $k \in C \cup \{j\}$ and $l = 0, 1, 2$. Here, we choose the time-varying weight function $w_i(t) = \Delta_i + (1 - \Delta_i)\hat{\xi}_i/\hat{\pi}(t)$, where $\hat{\pi}(t) = \sum_{i=1}^n (1 - \Delta_i)\hat{\xi}_i Y_i(t) / \sum_{i=1}^n (1 - \Delta_i)Y_i(t)$ is a consistent estimator of the true sampling probability $\pi$. Note that $w_i(\cdot)$ weights the $i$th subject by the inverse probability of selection, it equals to 1 for the cases and $\hat{\pi}(t)^{-1}$ for the sampled censored subjects. The maximum marginal pseudo-partial likelihood estimator $(\hat{\beta}_{C,j}, \hat{\beta}_j)$ is defined as the solution to the weighted estimating equation $U_j(\beta_{C,j}, \beta_j) = 0_{q+1}$. Define the information matrix $I_j(\beta_{C,j}, \beta_j) = -\left( \partial U_{j,k}(\beta_{C,j}, \beta_j)/\partial \beta_{l} \right)_{k,l \in C \cup \{j\}}$, which is of $(q + 1)$ dimension. Let

$$\hat{\sigma}_j^2 = [I_j(\beta_{C,j}, \hat{\beta}_j)]^{-1}_{q+1,q+1}$$

be the variance estimate of $\hat{\beta}_j$, i.e., the $(p+1)$th diagonal element of matrix $I_j(\beta_{C,j}, \hat{\beta}_j)$. For $j \notin C$, we define
\[ M_{C,j} = \frac{|\hat{\beta}_j|}{\hat{\sigma}_j}, \]

which serves as the proposed utility measure for the \( j \)th covariate. We rank covariates \( Z_j (j \notin C) \) by the value of \( M_{C,j} \) from the largest to smallest and retain those at the top of the rank list. For a given threshold \( \gamma > 0 \), the selected index set in addition to set \( C \) is given by

\[ \hat{A}_{-C} = \{ j \notin C : M_{C,j} = \frac{|\hat{\beta}_j|}{\hat{\sigma}_j} \geq \gamma \}. \tag{5} \]

In practical applications, we can pre-determine a positive integer \( d_0 \) and define the estimated active set as

\[ \hat{A}_{-C} = \{ j : M_{C,j} \text{ is amongst the first } d_0 \text{ largest of all } M_{C,j} (j \notin C) \}. \]

Similar to Fan and Lv (2008) and other literature related to feature screening, we can choose \( d_0 = \lceil n_{cc} / \log n_{cc} \rceil \), where \( n_{cc} \) denotes the case-cohort sample size.

Similar to the conditional screening procedures of Barut et al. (2016) and Hong et al. (2018), the outstanding advantage of the proposed CWSIS procedure is that it enables the detection of hidden active covariates for ultrahigh dimensional case-cohort data. To demonstrate this merit, we set up an example in a similar way to Barut et al. (2016) and Hong et al. (2018). In particular, the failure time \( T_i \) follows the Cox proportional hazards model

\[ \lambda(t | Z_i) = \lambda_0(t) \exp(\alpha^T Z_i), \]

where \( \lambda_0(t) = 1, \alpha = (1_4^T, -2, 0_{p-5}^T)^T, Z_i \sim N_p(0, \Sigma) \) with \( \Sigma = (\sigma_{ij})_{p \times p}, \sigma_{ii} = 1 \text{ for } i = 1, \ldots, p, \sigma_{ij} = 0.5 \text{ for } i \neq j. \)

By this design, \( Z_5 \) is a hidden active covariate. We consider four different conditioning sets, \( C = \{ \emptyset \}, \{1\}, \{1, 2\}, \{6, 7, 8\}. \) The densities of the proposed screening statistic \( M_{C,j} \) for \( Z_5 \) (hidden active covariate) and \( Z_{11}, \ldots, Z_{2000} \) (inactive covariates) are summarized in Fig. 1. When \( C = \emptyset \), CWSIS is equivalent to the marginal screening approach, the value of \( M_{C,j} \) for \( Z_5 \) is much smaller than the corresponding value of inactive covariates with a high probability. When the conditioning set includes one truly active covariate (\( C = \{1\} \)), the curve for \( Z_5 \) is on the right and there is a clear separation between these two curves. When we include more truly active covariates (\( C = \{1, 2\} \)), this separation becomes larger. We note a very interesting phenomenon that when the conditioning set consists of three inactive covariates (\( C = \{6, 7, 8\} \)), the chance of identifying the hidden variable \( Z_5 \) using CWSIS is still higher than the marginal screening method. This may be due to the correlation between them and the active covariates, such inactive variables can effectively function as surrogates for the active variables, thus conditioning on them can help detect hidden variables. A similar phenomenon was also observed in Barut et al. (2016) and Hong et al. (2018).

### 3 Theoretical property

In this section, we show the CWSIS procedure enjoys the sure screening property and the ranking consistency property, which demonstrate that our CWSIS
procedure tends to rank the active covariates above the inactive ones with high probability, furthermore, all the active covariates survive after screening with probability tending to 1 as \(n \to \infty\). These two properties lay out the theoretical foundation of our CWSIS procedure. Define

\[
S_k^{(l)}(t) = n^{-1} \sum_{i=1}^{n} Z_{ik}^{l} Y_i(t) \lambda(t | Z_{i}),
\]

\[
S_k^{(0)}(t) = n^{-1} \sum_{i=1}^{n} Z_{ik}^{l} Y_i(t) \exp(\beta_{C,j} Z_{i,C} + \beta_j Z_{ij}),
\]

and

\[
s_j^{(l)}(\beta_{C,j}, \beta_j, t) = E\{S_k^{(l)}(t)\},
\]

\[
s_j^{(0)}(\beta_{C,j}, \beta_j, t) = E\{S_k^{(0)}(t)\}
\]

for \(k \in C \cup \{j\}\) and \(l = 0, 1, 2\). Let \((\beta_{C,j}^{0}, \beta_j^{0})^T\) be the solution of the following equation

\[
u_{j,k}(\beta_{C,j}, \beta_j) = u_{j,k}(\beta_{C,j}, \beta_j) = \int_0^\tau \left\{ s_k^{(1)}(t) - \frac{s_j^{(1)}(\beta_{C,j}, \beta_j, t)}{s_j^{(0)}(\beta_{C,j}, \beta_j, t)} s_k^{(0)}(t) \right\} dt = 0.
\]
The regularity conditions are given in Appendix A, under which we establish the following lemmas and theorems.

**Lemma 1** Under conditions C1-C8, $\beta_0^j = 0$ if and only if $\alpha_j = 0$ for all $j \not\in C$.

**Lemma 2** Suppose conditions C1-C8 hold, there exist constants $c_2 > 0$ and $0 < \kappa < 1/2$ such that

$$\min_{j \in \bar{A}_C} |\beta_j^0| \geq c_2 n^{-\kappa}.$$ 

**Lemma 3** Under conditions C1-C8, for any $\epsilon_1 > 0$ and $\epsilon_2 > 0$, there exist positive constant $c_3$ and integer $N$ such that for any $n > N$ and $0 < \kappa < 1/2$,

$$P \left( \max_{j \in \bar{A}_C} |\hat{\beta}_j - \beta_j^0| > c_2 (n^{-\kappa} + \epsilon_1)/2 \right) \leq 2a(q + 1) \exp(-c_3 n^{-2\kappa}) + a\epsilon_2,$$

where $a$ is the size of $A_C$, $q$ is the size of $C$, $c_2$ is the same value in lemma 2.

Lemma 3 shows that the proposed maximum marginal pseudo-partial likelihood estimate $\hat{\beta}_j$ is a consistent estimate of $\beta_0^j$. By lemmas 1 and 3, we indeed can distinguish $Z_j (j \in \bar{A}_C)$ from $Z_j (j \not\in \bar{A}_C)$ by the proposed marginal utility $M_{C,j}$. Theorem 1 states the sure independent screening property of the CWSIS procedure.

**Theorem 1** (The sure screening property) Under conditions C1-C8, for any $0 < \kappa < 1/2$ and $\epsilon_2 > 0$, there exists positive constant $c_3$ such that

$$P \left( A_C \subseteq \hat{A}_C \right) \geq 1 - 2a(q + 1) \exp(-c_3 n^{-2\kappa}) - a\epsilon_2,$$

where $a$ is the size of $A_C$, $q$ is the size of $C$. Furthermore, we have

$$\lim_{n \to \infty} P \left( A_C \subseteq \hat{A}_C \right) = 1.$$

From this theorem, we can see that all active covariates survive after screening with a probability tending to one. The next theorem establishes the ranking consistency property of the proposed method.

**Theorem 2** (The ranking consistency) Under conditions C1-C8, we have

$$P \left( \max_{j \not\in \bar{A}_C} M_{C,j} < \min_{j \in \bar{A}_C} M_{C,j} \right) \to 1$$

when $n \to \infty$.

This lays out the theoretical foundation that our procedure ensures active covariates be ranked prior to the inactive ones with overwhelming probability. The proof of theorems and these lemmas are presented in the “Appendix B”.

 Springer
4 Simulation studies

We examine the finite sample performance of the proposed CWSIS procedure and make comparisons with some existing methods via simulation studies. For brevity, we refer to the feature aberration at survival times screening procedure of Gorst-Rasmussen and Scheike (2013) as FAST-SIS, the principled sure independence screening procedure of Zhao and Li (2012) as P-SIS, the censored rank independence screening of Song et al. (2014) as CRIS. Furthermore, we consider the marginal weighted screening procedure (MWSIS), where we fit the marginal Cox regressions \( \lambda(t | Z_{ij}) = \lambda_{0j}(t) \exp(\beta_j Z_{ij}) \) for each \( Z_{ij} \) and construct the weighted estimating equation to obtain the estimate \( \hat{\beta}_j \), then define the active index set as \( \hat{A} = \{ 1 \leq j \leq p : |\hat{\beta}_j| I_j(\hat{\beta}_j)^{1/2} \geq \gamma \} \), \( I_j(\hat{\beta}_j) \) denotes the information matrix. As the PSIS, FAST and CRIS can only deal with the SRS data, we generate the SRS data with the same sample size as the case-cohort data for PSIS, FAST and CRIS.

We consider the survival data generated from the Cox proportional hazards model and employ the independent Bernoulli sampling to generate the subcohort. We consider full cohort sample size \( n = 500, 1000, \) and the number of covariates \( p = 2000, 4000 \). As the incidence rate for case-cohort studies is usually very low or moderate, we consider the failure rate of 20% for \( n = 500, 5\% \) and 10% for \( n = 1000 \). We consider the noncase-to-case ratio of 1 : 1, thus the sample size of the case-cohort data in our simulation studies equals to 100, 200. For each configuration, we repeat 500 simulations and employ three evaluation criteria (Li et al. 2012b). The first one is the minimum model size to include all active predictors, denoted by \( S \). We present the median and interquartile range (IQR) of \( S \) out of 500 replications. The second one is the selection proportion that each important variable is selected into the model with a given model size \( d_0 \), denoted by \( P_e \). The third one is the selection proportion that all important variables are selected into the model with a given model size \( d_0 \), denoted by \( P_{a} \). An effective screening procedure is expected to yield \( S \) close to the true minimum model size and both \( P_e \) and \( P_{a} \) close to one. Here, we choose \( d_0 = [n_{cc}/\log n_{cc}] \) (Fan and Lv 2008), \( n_{cc} \) is the case-cohort sample size and \( [x] \) denotes the integer part of \( x \).

**Example 1** \( T_i \) are generated from the Cox proportional hazards model

\[
\lambda(t | Z_i) = \lambda_0(t) \exp(\alpha^T Z_i),
\]

where \( \lambda_0(t) = 1, \alpha = (1^T, -2, 0^T_{p-5})^T, Z_i \sim N_p(0, \Sigma) \) with \( \Sigma = (\sigma_{ij})_{p \times p}, \sigma_{ii} = 1 \) for \( i = 1, \ldots, p, \sigma_{ij} = 0.5 \) for \( i \neq j \). The censoring time \( C_i \sim \text{Unif}(0, \tau) \), the constant \( \tau \) represents the end time of the study and is used to control the failure rate.

**Example 2** We consider the same model as example 1, with \( \alpha = (10, 0^T_{p-2}, 1)^T \), i.e., only \( Z_1 \) and \( Z_p \) are active covariates. The first \( (p - 1) \) covariates \( (Z_1, \ldots, Z_{(p-1)}) \sim N_{p-1}(0, \Sigma) \) with \( \Sigma = (\sigma_{ij})_{(p-1) \times (p-1)} \), where \( \sigma_{ii} = 1 \) for \( i = 1, \ldots, (p-1), \sigma_{ij} = \rho \) for \( i \neq j \). We vary the value of \( \rho \) to be 0, 0.3, 0.7, with a larger \( \rho \) yielding a higher collinearity. The last covariate \( Z_p \sim N(0, 1) \).

We compute the absolute correlation between the survival time \( T \) and each covariate \( Z_j \) \( (j = 1, \ldots, p) \) for \( p = 2000 \) through the inverse probability weighting scheme.
### Example 1 \((FR = 20\%)\)

| Active Covariates | Hidden Active Covariates | Inactive Covariates |
|-------------------|--------------------------|---------------------|
| Density           | Density                  | Density             |
| 0.00              | 0.05                     | 0.10                |
| 0.15              | 0.20                     | 0.25                |

### Example 2 \((FR = 10\%)\)

| Active Covariates | Hidden Active Covariates | Inactive Covariates |
|-------------------|--------------------------|---------------------|
| Density           | Density                  | Density             |
| 0.00              | 0.05                     | 0.10                |
| 0.15              | 0.20                     | 0.25                |

### Example 3 \((FR = 5\%)\)

| Active Covariates | Hidden Active Covariates | Inactive Covariates |
|-------------------|--------------------------|---------------------|
| Density           | Density                  | Density             |
| 0.00              | 0.05                     | 0.10                |
| 0.15              | 0.20                     | 0.25                |

**Fig. 2** Absolute correlation of the survival time and the covariates for \(p = 2000\)

and further summarize the marginal correlation in three groups: the active covariates \((Z_1, \ldots, Z_4\) for example 1 and \(Z_1\) for example 2), the hidden active covariates \((Z_5\) for example 1 and \(Z_p\) for example 2), and the inactive covariates \((Z_6, \ldots, Z_p\) for example 1 and \(Z_2, \ldots, Z_{(p-1)}\) for example 2). Figures 2 and 3 depict the distribution of the absolute correlation for these three groups, from which we can see the marginal signal strength of hidden active covariates are weaker than the inactive covariates. Therefore, the marginal screening methods MWSIS, PSIS, FAST and CRIS are difficult to identify the hidden active covariates. The proposed conditional screening method CWSIS is an
ideal alternative. In our simulations, we simply choose $Z_1$ as the conditional covariate. In practice, if we have no useful prior information about active covariates, we can choose those covariates which have higher marginal signal strength as the conditional set (Barut et al. 2016; Lu and Lin 2020). To have a fair comparison, we add one (the number of conditional covariate in our examples) to $S$ for the proposed conditional screening method CWSIS.

The simulation results for $S$, $P_e$ and $P_a$ are summarized in Tables 1 and 2. By observing the values of $P_e$ for $Z_5$ in example 1 and $Z_p$ in example 2, we can conclude
Table 1 The median and interquartile range (IQR) of $S$, the selection proportions $P_e$ and $P_a$ among 500 replications for example 1

| $p$  | $n$  | FR (%) | $n_c$ | Method | Median | IQR | $P_e$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $P_a$ |
|------|------|--------|-------|--------|--------|-----|-------|-------|-------|-------|-------|-------|------|
| 2000 | 1000 | 5      | 50    | PSIS   | 1849   | 469 | 0.130 | 0.106 | 0.094 | 0.102 | 0.016 | 0.000 |
|      |      |        |       | FAST   | 1843   | 446 | 0.116 | 0.108 | 0.106 | 0.104 | 0.016 | 0.000 |
|      |      |        |       | CRIS   | 1398   | 764 | 0.290 | 0.270 | 0.288 | 0.250 | 0.220 | 0.084 |
|      |      |        |       | MWSIS  | 2000   | 2   | 0.354 | 0.342 | 0.352 | 0.342 | 0.016 | 0.000 |
|      |      |        |       | CWSIS  | 447    | 746 | –     | 0.318 | 0.350 | 0.330 | 0.476 | 0.018 |
|      | 1000 | 10     | 100   | PSIS   | 1998   | 18  | 0.488 | 0.456 | 0.436 | 0.462 | 0.000 | 0.000 |
|      |      |        |       | FAST   | 1998   | 19  | 0.474 | 0.444 | 0.422 | 0.450 | 0.000 | 0.000 |
|      |      |        |       | CRIS   | 1721   | 379 | 0.050 | 0.032 | 0.020 | 0.038 | 0.004 | 0.000 |
|      |      |        |       | MWSIS  | 2000   | 0   | 0.784 | 0.804 | 0.774 | 0.794 | 0.000 | 0.000 |
|      |      |        |       | CWSIS  | 69     | 172 | –     | 0.790 | 0.768 | 0.810 | 0.760 | 0.356 |
|      | 500  | 20     | 100   | PSIS   | 2000   | 1   | 0.686 | 0.706 | 0.706 | 0.668 | 0.002 | 0.000 |
|      |      |        |       | FAST   | 2000   | 1   | 0.654 | 0.654 | 0.694 | 0.622 | 0.002 | 0.000 |
|      |      |        |       | CRIS   | 1720   | 405 | 0.054 | 0.054 | 0.044 | 0.054 | 0.002 | 0.000 |
|      |      |        |       | MWSIS  | 2000   | 0   | 0.812 | 0.832 | 0.840 | 0.798 | 0.000 | 0.000 |
|      |      |        |       | CWSIS  | 47     | 168 | –     | 0.828 | 0.852 | 0.806 | 0.764 | 0.442 |
Table 1 continued

| $p$ | $n$ | FR (%) | $n_c$ | Method  | Median | IQR  | $P_e$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $P_a$ |
|-----|-----|--------|-------|---------|--------|------|-------|-------|-------|-------|-------|-------|-------|
| 4000 | 1000 | 5      | 50    | PSIS    | 3747   | 716  | 0.384 | 0.364 | 0.376 | 0.380 | 0.002 | 0.000 |
|      |      |        |       | FAST    | 3748   | 744  | 0.368 | 0.352 | 0.378 | 0.362 | 0.002 | 0.000 |
|      |      |        |       | CRIS    | 3133   | 1100 | 0.022 | 0.018 | 0.014 | 0.022 | 0.000 | 0.000 |
|      |      |        |       | MWSIS   | 4000   | 3    | 0.670 | 0.700 | 0.710 | 0.702 | 0.000 | 0.000 |
|      |      |        |       | CWSIS   | 908    | 1477 | –     | 0.720 | 0.680 | 0.734 | 0.688 | 0.252 |
|      | 1000 | 10     | 100   | PSIS    | 3995   | 46   | 0.384 | 0.364 | 0.376 | 0.380 | 0.002 | 0.000 |
|      |      |        |       | FAST    | 3995   | 48   | 0.368 | 0.352 | 0.378 | 0.362 | 0.002 | 0.000 |
|      |      |        |       | CRIS    | 3363   | 795  | 0.022 | 0.018 | 0.014 | 0.022 | 0.000 | 0.000 |
|      |      |        |       | MWSIS   | 4000   | 0    | 0.670 | 0.700 | 0.710 | 0.702 | 0.000 | 0.000 |
|      |      |        |       | CWSIS   | 136    | 389  | –     | 0.720 | 0.680 | 0.734 | 0.688 | 0.252 |
|      | 500  | 20     | 100   | PSIS    | 4000   | 2    | 0.600 | 0.608 | 0.578 | 0.630 | 0.000 | 0.000 |
|      |      |        |       | FAST    | 4000   | 2    | 0.582 | 0.592 | 0.574 | 0.582 | 0.000 | 0.000 |
|      |      |        |       | CRIS    | 3447   | 871  | 0.036 | 0.050 | 0.038 | 0.024 | 0.000 | 0.000 |
|      |      |        |       | MWSIS   | 4000   | 0    | 0.770 | 0.732 | 0.730 | 0.766 | 0.000 | 0.000 |
|      |      |        |       | CWSIS   | 86     | 277  | –     | 0.770 | 0.784 | 0.806 | 0.746 | 0.350 |

$n$, the sample size of the full cohort; $p$, the number of covariates; FR, the failure rate; $n_c$, the average number of cases; CWSIS: the proposed conditional screening method; MWSIS: the marginal weighted screening procedure; PSIS: the screening procedure of Zhao and Li (2012); FAST: the screening procedure of Gorst-Rasmussen and Scheike (2013); CRIS: the screening procedure of Song et al. (2014)
Table 2  The median and interquartile range (IQR) of $S$, the selection proportions $P_e$ and $P_a$ among 500 replications for example 2

| $n$ | FR | $\rho$ | $n_c$ | Method | $\bar{P}_e$ | Median | IQR | $X_1$ | $X_p$ | $\bar{P}_a$ | Median | IQR | $X_1$ | $X_p$ | $\bar{P}_a$ |
|-----|----|-----|-----|------|----------|--------|-----|-----|-----|----------|--------|-----|-----|-----|----------|
| 500 | 20%| 0   | 100 | PSIS | 578      | 979    | 1.00| 0.092| 0.092| 1279     | 2088   | 1.00| 0.066| 0.066|
|     |    |     |     | FAST | 594      | 975    | 1.00| 0.090| 0.090| 1286     | 2054   | 1.00| 0.068| 0.068|
|     |    |     |     | CRIS | 841      | 1005   | 1.00| 0.032| 0.032| 1683     | 1920   | 0.998| 0.014| 0.014|
|     |    |     |     | MWSIS| 424      | 951    | 1.00| 0.104| 0.104| 936      | 1819   | 1.00| 0.088| 0.088|
|     | 0.3| 100 |     | CWSIS| 2       | 0      | –   | 1.000| 1.000| 2        | 0      | –   | 1.000| 1.000|
|     |    |     |     | PSIS | 1973     | 131   | 1.00| 0.000| 0.000| 3958     | 302    | 1.00| 0.000| 0.000|
|     |    |     |     | FAST | 1971     | 138   | 1.00| 0.000| 0.000| 3958     | 311    | 1.00| 0.000| 0.000|
|     |    |     |     | CRIS | 1278     | 1192  | 0.998| 0.022| 0.022| 2795     | 2220   | 0.998| 0.012| 0.012|
|     |    |     |     | MWSIS| 1997     | 33    | 1.00| 0.000| 0.000| 3993     | 62     | 1.00| 0.002| 0.002|
|     | 0.7| 100 |     | CWSIS| 2       | 0      | –   | 1.000| 1.000| 2        | 0      | –   | 1.000| 1.000|
|     |    |     |     | PSIS | 2000     | 0     | 0.380| 0.000| 0.000| 4000     | 0      | 0.278| 0.000| 0.000|
|     |    |     |     | FAST | 2000     | 0     | 1.000| 0.000| 0.000| 4000     | 0      | 1.000| 0.000| 0.000|
|     |    |     |     | CRIS | 1945     | 460   | 0.990| 0.008| 0.008| 3923     | 848    | 0.978| 0.004| 0.004|
|     |    |     |     | MWSIS| 2000     | 0     | 0.676| 0.000| 0.000| 4000     | 0      | 0.558| 0.000| 0.000|
|     |    |     |     | CWSIS| 2       | 0      | –   | 1.000| 1.000| 2        | 0      | –   | 1.000| 1.000|
| $n$ | FR | $\rho$ | $n_c$ | Method | $p = 2000$ | $p = 4000$ |
|-----|----|-------|------|--------|------------|------------|
|     |     |       |      |        | Median | IQR | $P_x$ | $X_1$ | $X_p$ | $P_a$ | Median | IQR | $P_x$ | $X_1$ | $X_p$ | $P_a$ |
| 1000 | 10% | 0     | 100  | PSIS   | 664    | 1033 | 1.000 | 0.064 | 0.064 | 1385 | 2072 | 1.000 | 0.038 | 0.038 |
|      | FAST| 684   | 1024 | 1.000 | 0.064 | 0.064 | 1376 | 2031 | 1.000 | 0.036 | 0.036 |
|      | CRIS| 937   | 982  | 0.920 | 0.014 | 0.008 | 1938 | 2140 | 0.816 | 0.002 | 0.000 |
|      | MWSIS| 315  | 795  | 1.000 | 0.170 | 0.170 | 599  | 1766 | 1.000 | 0.116 | 0.116 |
|      | CWSIS| 2    | 0    | –     | 0.998 | 0.998 | 2    | 0    | –     | 0.998 | 0.998 |
| 0.3  | 100 | PSIS  | 1928 | 374   | 0.964 | 0.002 | 0.002 | 3870 | 710  | 0.946 | 0.000 | 0.000 |
|      | FAST| 1926 | 382  | 1.000 | 0.002 | 0.002 | 3863 | 678  | 1.000 | 0.000 | 0.000 |
|      | CRIS| 1233 | 1119 | 0.884 | 0.034 | 0.030 | 2427 | 2355 | 0.810 | 0.024 | 0.012 |
|      | MWSIS| 1999 | 15   | 1.000 | 0.000 | 0.000 | 3998 | 63   | 1.000 | 0.000 | 0.000 |
|      | CWSIS| 2    | 0    | –     | 1.000 | 1.000 | 2    | 0    | –     | 0.998 | 0.998 |
| 0.7  | 100 | PSIS  | 2000 | 0     | 0.042 | 0.000 | 0.000 | 4000 | 0    | 0.016 | 0.000 | 0.000 |
|      | FAST| 2000 | 0    | 0.996 | 0.000 | 0.000 | 4000 | 0    | 0.994 | 0.000 | 0.000 |
|      | CRIS| 1737 | 930  | 0.794 | 0.028 | 0.024 | 3451 | 1921 | 0.710 | 0.024 | 0.010 |
|      | MWSIS| 2000 | 0    | 0.208 | 0.000 | 0.000 | 4000 | 0    | 0.150 | 0.000 | 0.000 |
|      | CWSIS| 2    | 0    | –     | 1.000 | 1.000 | 2    | 0    | –     | 0.998 | 0.998 |
Table 2  continued

| $n$  | FR  | $\rho$ | $n_c$ | Method   | $p = 2000$ | $p = 4000$ |
|------|-----|--------|-------|----------|------------|------------|
|      |     |        |       |          | Median     | Median     |
|      |     |        |       |          | IQR $X_1$  | IQR $X_p$  |
|      |     |        |       |          | $P_e$      | $P_a$      |
| 1000 | 5%  | 0      | 50    | PSIS     | 1075       | 2010       |
|      |     |        |       |          | 984        | 2274       |
|      |     |        |       |          | 0.254      | 0.266      |
|      |     |        |       |          | 0.006      | 0.008      |
|      |     |        |       |          | 0.002      | 0.006      |
|      |     |        |       | FAST     | 931        | 1771       |
|      |     |        |       |          | 1022       | 2138       |
|      |     |        |       |          | 0.990      | 1.000      |
|      |     |        |       |          | 0.008      | 0.002      |
|      |     |        |       |          | 0.008      | 0.002      |
|      |     |        |       | CRIS     | 1332       | 2364       |
|      |     |        |       |          | 1002       | 1826       |
|      |     |        |       |          | 0.562      | 0.436      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       | MWSIS    | 520        | 1082       |
|      |     |        |       |          | 983        | 2023       |
|      |     |        |       |          | 1.000      | 1.000      |
|      |     |        |       |          | 0.042      | 0.030      |
|      |     |        |       |          | 0.042      | 0.030      |
|      |     |        |       | CWSIS    | 2          | 102        |
|      |     |        |       |          | 1          | 2          |
|      |     |        |       |          | –          | –          |
|      |     |        |       |          | 0.936      | 0.882      |
|      |     |        |       |          | 0.936      | 0.882      |
|      | 0.3 | 50     |       | PSIS     | 1678       | 3459       |
|      |     |        |       |          | 667        | 1476       |
|      |     |        |       |          | 0.080      | 0.046      |
|      |     |        |       |          | 0.002      | 0.004      |
|      |     |        |       |          | 0.000      | 0.002      |
|      |     |        |       | FAST     | 1580       | 3249       |
|      |     |        |       |          | 825        | 1642       |
|      |     |        |       |          | 0.958      | 0.976      |
|      |     |        |       |          | 0.002      | 0.002      |
|      |     |        |       |          | 0.002      | 0.002      |
|      |     |        |       | CRIS     | 1501       | 2592       |
|      |     |        |       |          | 1077       | 2036       |
|      |     |        |       |          | 0.568      | 0.448      |
|      |     |        |       |          | 0.006      | 0.004      |
|      |     |        |       |          | 0.004      | 0.000      |
|      |     |        |       | MWSIS    | 1981       | 3971       |
|      |     |        |       |          | 155        | 204        |
|      |     |        |       |          | 0.984      | 0.956      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       | CWSIS    | 2          | 2          |
|      |     |        |       |          | 0          | 1          |
|      |     |        |       |          | –          | –          |
|      |     |        |       |          | 0.936      | 0.890      |
|      |     |        |       |          | 0.936      | 0.890      |
|      | 0.7 | 50     |       | PSIS     | 2000       | 4000       |
|      |     |        |       |          | 2          | 1          |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       | FAST     | 2000       | 4000       |
|      |     |        |       |          | 2          | 1          |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       | CRIS     | 1721       | 3093       |
|      |     |        |       |          | 1037       | 2247       |
|      |     |        |       |          | 0.590      | 0.462      |
|      |     |        |       |          | 0.010      | 0.008      |
|      |     |        |       |          | 0.002      | 0.002      |
|      |     |        |       | MWSIS    | 2000       | 4000       |
|      |     |        |       |          | 0          | 0          |
|      |     |        |       |          | 0.020      | 0.006      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       |          | 0.000      | 0.000      |
|      |     |        |       | CWSIS    | 2          | 2          |
|      |     |        |       |          | 0          | 1          |
|      |     |        |       |          | –          | –          |
|      |     |        |       |          | 0.942      | 0.900      |
|      |     |        |       |          | 0.942      | 0.900      |

$n$, the sample size of the full cohort; $p$, the number of covariates; FR, the failure rate; $n_c$, the average number of cases; $\rho$, the correlation coefficient of covariates; CWSIS: the proposed conditional screening method; MWSIS: the marginal weighted screening procedure; PSIS: the screening procedure of Zhao and Li (2012); FAST: the screening procedure of Gorst-Rasmussen and Scheike (2013); CRIS: the screening procedure of Song et al. (2014)
that the proposed CWSIS procedure can detect the hidden active covariates with high probabilities, while the other four methods MWSIS, PSIS, FAST and CRIS fail to select them. In example 2, $\rho$ equals to 0, 0.3, and 0.7, with a larger $\rho$ yielding a higher collinearity. The proposed method CWSIS performs well even with high collinearity, while the other four methods do not behave well even when $\rho = 0$ and the performance deteriorates with the increasing value of $\rho$. As expected, CWSIS needs a smaller model size to possess the sure screening property in all settings. Larger case-cohort sample size and higher failure rate are associated with better performance. In particular, larger cohort sample size can handle rare disease situations better.

To assess the performance of the proposed method in the settings that are similar to the real data, we further consider $n = 300$ and the failure rate of 25% for example 2, the remaining setups are kept the same as before. Here, we also consider the unweighted conditional screening method NCWSIS which does not adopt the weight function and simply treat the case-cohort data as SRS data, and the conditional screening method CSMPLE in Hong et al. (2018). Since the method CSMPLE in Hong et al. (2018) is proposed for SRS data, it can not be directly used to handle the case-cohort data, we generate the SRS data with the same sample size as the case-cohort data for CSMPLE. The simulation results for $S$, $P_e$ and $P_a$ are summarized in Table 3, from which we can see that the proposed method can detect the hidden active covariates with high probabilities and delivers its distinctive advantages for all the considered settings. By comparing the results of NCWSIS, CSMPLE and CWSIS, we can conclude that the performance of the conditional screening method is improved by including the case-cohort weight. Moreover, the proposed conditional screening procedure based on case-cohort design is more accurate in selecting the active covariates than the conditional screening based on a SRS of the same size as the case-cohort sample. For example, when $p = 2000$ and $\rho = 0.7$, the value of $P_a$ is only 0.460 for CSMPLE, while the corresponding value of the proposed method CWSIS equals to 1.

5 Application to breast cancer data

As an illustration, we apply the proposed CWSIS method to the breast cancer data (van de Vijver et al. 2002), with 295 female patients who have primary invasive breast carcinoma. For each patient, the expressions of 24885 genes were profiled on cDNA arrays from all tumors. A set of 4919 candidate genes were selected after initial screening using the Rosetta error model (van’t Veer et al. 2002). By excluding the individuals with missing values, we have 289 subjects with 4919 candidate genes. The median observed time was 7.23 years (ranging from 0.05 to 18.34 years). During the follow-up, 78 patients died of breast cancer and the other 211 patients were still alive, which led to the failure rate of 26.99%. Of the 289 patient samples, 60 samples overlapped with the 78 training samples from van’t Veer et al. (2002), we use these 60 samples as the testing set and the case-cohort samples as our training set. The details of these two sets are summarized in Table 4. The interest of the study is to identify genes that have great influence on patients’ overall survival rate.
Table 3. The median and interquartile range (IQR) of $S$, the selection proportions $P_e$ and $P_a$ among 500 replications for example 2 with $n = 300$ and FR=25%.

| $\rho$ | $n_c$ | Method | $p=2000$ |  | $p=4000$ |  |
|---|---|---|---|---|---|---|
| 0 | 75 | PSIS | 671 | 1040 | 1.000 | 0.062 | 0.062 | 1226 | 1920 | 1.000 | 0.032 | 0.032 |
|  |  | FAST | 711 | 1035 | 1.000 | 0.068 | 0.068 | 1221 | 1927 | 1.000 | 0.034 | 0.034 |
|  |  | CRIS | 854 | 1064 | 1.000 | 0.026 | 0.026 | 1626 | 2057 | 0.998 | 0.012 | 0.012 |
|  |  | MWSIS | 599 | 1057 | 1.000 | 0.058 | 0.058 | 1014 | 1900 | 1.000 | 0.040 | 0.040 |
|  |  | NCWSIS | 5 | 53 | – | 0.706 | 0.706 | 8 | 115 | – | 0.648 | 0.648 |
|  |  | CSAMPLE | 7 | 68 | – | 0.674 | 0.674 | 12 | 108 | – | 0.608 | 0.608 |
|  |  | CWSIS | 2 | 0 | – | 1.000 | 1.000 | 2 | 0 | – | 0.994 | 0.994 |

| 0.3 | 75 | PSIS | 1960 | 236 | 1.000 | 0.000 | 0.000 | 3917 | 485 | 1.000 | 0.000 | 0.000 |
|  |  | FAST | 1959 | 236 | 1.000 | 0.000 | 0.000 | 3915 | 516 | 1.000 | 0.000 | 0.000 |
|  |  | CRIS | 1339 | 1216 | 0.998 | 0.014 | 0.014 | 2718 | 2467 | 0.994 | 0.020 | 0.020 |
|  |  | MWSIS | 1987 | 106 | 1.000 | 0.000 | 0.000 | 3966 | 225 | 1.000 | 0.000 | 0.000 |
|  |  | NCWSIS | 3 | 60 | – | 0.712 | 0.712 | 4 | 57 | – | 0.704 | 0.704 |
|  |  | CSAMPLE | 10 | 65 | – | 0.648 | 0.648 | 16 | 223 | – | 0.574 | 0.574 |
|  |  | CWSIS | 2 | 0 | – | 1.000 | 1.000 | 2 | 0 | – | 1.000 | 1.000 |

| 0.7 | 75 | PSIS | 2000 | 0 | 0.596 | 0.000 | 0.000 | 4000 | 0 | 0.578 | 0.000 | 0.000 |
|  |  | FAST | 2000 | 0 | 1.000 | 0.000 | 0.000 | 4000 | 0 | 1.000 | 0.000 | 0.000 |
|  |  | CRIS | 1948 | 399 | 0.990 | 0.002 | 0.002 | 3921 | 1035 | 0.978 | 0.006 | 0.006 |
|  |  | MWSIS | 2000 | 0 | 0.610 | 0.000 | 0.000 | 4000 | 0 | 0.556 | 0.000 | 0.000 |
|  |  | NCWSIS | 2 | 40 | – | 0.732 | 0.732 | 2 | 38 | – | 0.736 | 0.736 |
|  |  | CSAMPLE | 44 | 204 | – | 0.460 | 0.460 | 74 | 606 | – | 0.388 | 0.388 |
|  |  | CWSIS | 2 | 0 | – | 1.000 | 1.000 | 2 | 0 | – | 1.000 | 1.000 |

$n$, the sample size of the full cohort; $p$, the number of covariates; FR, the failure rate; $n_c$, the average number of cases; $\rho$, the correlation coefficient of covariates; PSIS: the screening procedure of Zhao and Li (2012); FAST: the screening procedure of Gorst-Rasmussen and Scheike (2013); CRIS: the screening procedure of Song et al. (2014); MWSIS: the marginal weighted screening procedure; NCWSIS: the unweighted conditional screening method; CSAMPLE: the conditional screening method of Hong et al. (2018); CWSIS: the proposed conditional screening method.
Table 4  Summary of the breast cancer data

| Dataset | Num | Min | Max    | Median | Fail(%) |
|---------|-----|-----|--------|--------|--------|
| Train   | 289 | 0.055 | 18.341 | 7.225  | 26.99  |
| Test    | 60  | 0.712 | 15.352 | 7.606  | 38.33  |

Train, the training set; Test, the testing set; Num, the number of patients; Min, the minimum observed survival time; Max, the maximum observed survival time; Median, the median of observed survival time; Fail, the failure rate

Table 5  The results of selected important genes for the breast cancer data using the regularization methods

| Name                  | Est. | Name                  | Est. | Name                  | Est. |
|-----------------------|------|-----------------------|------|-----------------------|------|
| Contig58368.RC        | 0.392| Contig58368.RC        | 0.516| Contig58368.RC        | 0.515|
| NM.014889             | 0.277| NM.014889             | 0.446| NM.014889             | 0.445|
| NM.005689             | 0.201| NM.005689             | 0.329| NM.005689             | 0.329|
| NM.013332             | 0.199| NM.013290             | 0.326| NM.013290             | 0.325|
| Contig63649.RC        | 0.178| AL080059              | 0.312| AL080059              | 0.312|
| NM.013290             | 0.172| NM.013332             | 0.256| NM.013332             | 0.256|
| AL080059              | 0.168| Contig63649.RC        | 0.249| Contig63649.RC        | 0.249|
| NM.002916             | 0.140| NM.002916             | 0.204| NM.002916             | 0.206|
| NM.012291             | 0.102|                      |      |                      |      |
| Contig31288.RC        | 0.083|                      |      |                      |      |
| Contig38288.RC        | 0.049|                      |      |                      |      |
| NM.003376             | 0.017|                      |      |                      |      |
| NM.001673             | 0.014|                      |      |                      |      |

Name: the name for selected genes; Est.: the corresponding estimated value of the coefficient for selected genes

We illustrate the proposed method by identifying genes that have great influence on patients’ overall survival rate based on data from a case-cohort sample. Specifically, we select the subcohort by independent Bernoulli sampling with the selection probability $\pi = 0.37$, which results in about the same number of cases and noncases. The subcohort has 111 subjects and the final case-cohort sample has 155 subjects. Gene AL080059 has been known to be predictive to patients’ survival time in the literature (Yeung et al. 2005; van’t Veer et al. 2002), we use it as the conditional variable in the proposed procedure. The screening methods are usually considered as an initial step to reduce the dimensionality and then followed with some model-based regularization methods. In particular, we first apply the proposed CWSIS procedure to reduce the dimension from $p = 4919$ to $\lceil 155/\log(155) \rceil = 31$ and then utilize different regularization methods LASSO, SCAD and MCP to select the significant ones among these 31 genes under the framework of the Cox proportional hazards regression, the tuning parameter was selected by the 10-fold cross-validation. We summarize the name and the corresponding estimated value of the coefficient for selected genes in Table 5, from
which we can see that genes Contig58368.RC, NM.014889, NM.005689, NM.013290, AL080059, NM.013332, Contig63649.RC and NM.002916 were all selected by the LASSO, SCAD and MCP methods, indicating that these eight genes could be associated with patients’ survival rate. Moreover, genes Contig58368.RC, NM.014889 and NM.005689 were ranked at the first three position, which means that these three genes may have great influence on patients’ survival rate.

To evaluate the predictive accuracy of CWSIS, we further compute the $C$-statistic estimator (Uno et al. 2011). For comparison, we also apply the MWSIS and NCWSIS procedures to analyze this data. In particular, we first apply these three screening methods to reduce the dimension to $\lceil 155/\log(155) \rceil = 31$, then perform the LASSO penalization to further remove some irrelevant covariates, with the tuning parameter selected by the 10-fold cross-validation. We obtain the risk score for each subject by using the final model selected by LASSO and further compute the corresponding concordance statistic ($C$-statistic) (Uno et al. 2011) in the testing set. The standard deviations (SD) of $C$-statistic are obtained from perturbation resampling 1000 times. The corresponding values of $C$-statistic and SD (the values in the parenthesis) are 0.862 (0.059), 0.796 (0.078), 0.802 (0.053) for CWSIS, MWSIS, NCWSIS procedures, respectively. According to Uno et al. (2011), the larger the $C$-statistic is, the stronger predictive power the method possesses. We can conclude that the proposed CWSIS method performs reasonably well for ultrahigh-dimensional survival data under the case-cohort design and delivers a favorable performance in terms of prediction.

We also consider $d_n = n/2, n/3, n/4$ when analyzing this data and summarize the results in the supplementary material, from which we can see that the selected genes under different cut-offs are highly consistent. Furthermore, we compute the $C$-statistic estimator for CWSIS, MWSIS, NCWSIS procedures under these three cases. From the results in the supplementary material we can make similar conclusion to that with $d_n = n/\log(n)$.

6 Conclusion

For ultrahigh-dimensional survival data under the case-cohort design, we propose a conditional screening procedure CWSIS by incorporating the prior information of active covariates. This method enables the detection of hidden active covariates, which is an outstanding advantage compared with the marginal screening procedures. Moreover, the proposed procedure does not require any complicated numerical optimization and is computationally efficient. Theoretically, it enjoys the sure screening property and ranking consistency property under some mild regularity conditions. In the development of the theoretical properties, we adopt the conditional linear expectation and conditional linear covariance, which are proposed in Hong et al. (2018) and are useful to specify the regularity conditions.

There are some issues that deserve further considerations. First, the proposed method requires the prior information of active covariates, sometimes it may be difficult to obtain such useful information. Hong et al. (2016) proposed a data-driven method to obtain the conditional set for generalized linear models. How to develop a data-driven conditional screening method for survival data under the case-cohort is an
interesting question. Furthermore, when we have prior knowledge of active covariates, how to balance it with the information extracted from the given data merits further investigation. Second, under our design, the subcohort is selected by independent Bernoulli sampling. When the subcohort is selected by simple random sampling without replacement, our method also works, although more complicated arguments would be needed to develop the theoretical properties. Moreover, when some covariates are available for all cohort members, we can consider the stratified case-cohort design based on those covariates. Third, we can consider to propose more efficient screening methods which incorporating more complex prior knowledge, such as the network structure or the spatial information of the covariates.

Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s10985-021-09531-7.

Acknowledgements  This work is funded in part by the U.S. National Institute of Health Grants (P01CA142538, P42ES031007, P30ES010126), the National Natural Science Foundation of China grants (Nos. 11971362, 11901581, 11771366).

Declarations

Conflict of interest The authors declare that we have no conflict of interest.

Appendix

Appendix A: regularity conditions

Let \( S_T(t|Z_i) = \exp\{-\Lambda_0(t) \exp(\alpha^T Z_i)\} \) and \( S_C(t|Z_i) = P(C_i > t|Z_i) \) denote the survival functions of \( T_i \) and \( C_i \), \( F_T(t|Z_i) = 1 - S_T(t|Z_i) \), \( \Lambda_0(t) = \int_0^t \lambda_0(s)ds \) denotes the cumulative baseline hazard function. For any vector \( v = (v_1, \ldots, v_p) \in \mathbb{R}^p \), let \( \|v\|_d = \sqrt{\sum_{j=1}^p |v_j|^d} \) be the \( L_d \) norm. For any random variables \( \zeta : \Omega \rightarrow \mathbb{R}^d \), \( \zeta_1 : \Omega \rightarrow \mathbb{R}^{d_1} \), \( \zeta_2 : \Omega \rightarrow \mathbb{R}^{d_2} \) and \( \eta : \Omega \rightarrow \mathbb{R}^p \), the conditional linear expectation of \( \zeta \) given \( \eta \) is defined as \( E^*(\zeta|\eta) = E(\zeta) + B^T(\eta - E(\eta)) \), where \( B = \arg\min_{B \in \mathbb{R}^{d \times p}} E\{[\zeta - E(\zeta) - D^T(\eta - E(\eta))]^2|\eta]\}. \) The conditional linear covariance between \( \zeta_1 \) and \( \zeta_2 \) given \( \eta \) is defined as \( Cov^*(\zeta_1, \zeta_2|\eta) = E^*[\{\zeta_1 - E^*(\zeta_1|\eta)]\{\zeta_2 - E^*(\zeta_2|\eta)]|\eta]\}. \) The properties of \( E^*(\zeta|\eta) \) and \( Cov^*(\zeta_1, \zeta_2|\eta) \) are presented in “Appendix B”. The regularity conditions listed below are imposed throughout our discussions.

C1. For each \( j \notin C \) and \( k \in C \cup \{j\} \), there exists a neighborhood \( B_j \) of \( (\beta_{C,j}^0, \beta_j^0)^T \) such that

\[
\sup_{t \in [0, \tau], (\beta_{C,j}, \beta_j)^T \in B_j} \| S_{j,k}^{(l)}(\beta_{C,j}, \beta_j, t) - s_{j,k}^{(l)}(\beta_{C,j}, \beta_j, t) \|_2 \to 0
\]

in probability as \( n \to \infty \) (\( l = 0, 1 \)), \( s_{j,k}^{(l)}(\beta_{C,j}, \beta_j, t) \) is bounded away from zero on \( B_j \times [0, \tau] \), \( s_{j,k}^{(l)}(\beta_{C,j}, \beta_j, t) \) are bounded on \( B_j \times [0, \tau] \).
C2. For all \( j = 1, \ldots, p \), \( \int_0^T \lambda_{j,0}(t) dt < \infty \) and \( E\{Y(\tau)\} > 0 \).

C3. The covariates \( Z_j (j = 1, \ldots, p) \) are independent of time and bounded by a constant \( L_0 \). Furthermore, \( E(Z_j) = 0 \) for all \( j \in \{1, \ldots, p\} \).

C4. All \( Z_j, j \in A_C \) are independent of all \( Z_j, j \notin A_C \) given \( Z_C \).

C5. There exists a constant \( L_1 \) such that \( \|\alpha\|_1 < L_1 \) and \( \|B_j\|_1 < L_1 \).

C6. There exist constants \( C_1 \geq 0 \) and \( 0 < \kappa < 1/2 \) such that \( \min_{j \in A_C} |E[Cov^*(Z_j, P(\delta = 1|Z)|Z_C)]| \leq C_1 n^{-\kappa} \).

C7. There exists a constant \( L > 0 \) such that \( \|\hat{\beta}_{C,j} - \beta_j\|_2 \leq L \|\beta_j\|_2 \) for all \( j \notin C \).

C8. Let \( \tilde{n} = \sum_{i=1}^n \xi_i \) denote the sample size of subcohort, then \( \tilde{n}/n \) converges to the constant \( \pi \in (0, 1) \).

Conditions C1 and C2 are common assumptions in survival analysis (Andersen and Gill 1982; Fleming and Harrington 1991). Condition C3 assumes the covariates are bounded, which is similar to the partial orthogonality assumption of the covariates. Condition C5 controls the total effect size of the covariates, it is reasonable under the sparsity principle. Condition C6 is a typical assumption which has been widely used in the literature of feature screening, such as condition 3 in Fan and Lv (2008), condition 2 in Song et al. (2014), conditions 2 and 5 in Wu and Yin (2015), etc. Condition C7 is a mild assumption which holds in many situations. Condition C8 is a common assumption on the case-cohort design.

Appendix B: lemmas and theoretic proofs

Let \( \beta_{C,0} \) be the solution of the equations \( u_C(\beta_C) = [u_{j,k}(\beta_C, 0), k \in C]^T = 0_q \). Define \( v_j(\beta_{C,0}, \beta_j) = u_{j,j}(\beta_{C,0}, \beta_j) - \sum_{k \in C} b_k u_{j,k}(\beta_{C,0}, \beta_j) \), where vector \( b_C = [b_k, k \in C]^T \) such that \( E^*[Z_j|Z_C] = \sum_{k \in C} b_k Z_k \). As a preparation, we first introduce some lemmas.

Lemma 4 Let \( \xi = (\xi_1, \ldots, \xi_n) \) be a random vector containing \( n - \tilde{n} \) ones and \( n - \tilde{n} \) zeros, with each permutation equally likely. Let \( B_i(t) (i = 1, \ldots, n) \) be independent and identically distributed real-valued random processes on \([0, \tau]\) with \( E\{B_i(t)\} = \mu_B(t) \), \( var(B_i(\tau)) \) converges weakly in \( l^\infty[0, \tau] \) to a zero-mean Gaussian process and therefore \( n^{-1/2} \sum_{i=1}^n \xi_i (B_i(t) - \mu_B(t)) \) converges in probability to zero uniformly in \( t \).

This Lemma is the same as Lemma A1 of Kang and Cai (2009).

Lemma 5 Given that \( \xi \) is independent of \( \Delta \) and \( Y(t) \), \( n^{1/2} \tilde{\pi}^{-1}(t) - \pi^{-1} \) converges weakly to a zero-mean Gaussian process.

This lemma is extracted from lemma A3 of Ni et al. (2016).
Lemma 6 For independent random variables $Y_1, \ldots, Y_n$ with bounded ranges $[-M, M]$ and zero mean,

$$P \left( |Y_1 + \ldots + Y_n| > y \right) \leq 2 \exp \left( -\frac{1}{2} \frac{y^2}{V + M y/3} \right)$$

for $V \geq Var(Y_1 + \ldots + Y_n)$.

This lemma is extracted from lemma 2.2.9 of van der Vaart and Wellner (1996).

Lemma 7 Let $\xi, \xi_1, \xi_2$ and $\eta$ be any four random variables in the probability space $(\Omega, F, P)$, the following properties hold for the conditional linear expectation $E^*(\cdot | \eta)$ given $\eta$:

1. $E^*(\xi | \eta) = E(\xi) + \text{Cov}(\xi, \eta) \text{Var}(\eta)^{-1} \{ \eta - E(\eta) \}$;
2. $E^*(\eta | \eta) = \eta$;
3. For any matrices $A_1$ and $A_2$, $E^*(A_1 \xi_1 + A_2 \xi_2 | \eta) = A_1 E^*(\xi_1 | \eta) + A_2 E^*(\xi_2 | \eta)$;
4. $E^*[E^*(\xi | \eta)] = E[E^*(\xi | \eta)] = E[\xi]$.

This lemma is extracted from proposition 2 of Hong et al. (2018).

Lemma 8 The conditional linear covariance has the following properties:

1. $\text{Cov}^*(\xi_1, \xi_2 | \eta) = 0 \iff E^*(\xi_1 \xi_2 | \eta) = E^*(\xi_1 | \eta) E^*(\xi_2 | \eta)$;
2. $E[\text{Cov}^*(\xi_1, \xi_2 | \eta)] = \text{Cov}(\xi_1, \xi_2) - \text{Cov}(\xi_1, \eta) \text{Var}(\eta)^{-1} \text{Cov}(\eta, \xi_2)$;
3. For any increasing function $h(\cdot) : R \to R$ and random variable $\xi : \Omega \to R$, we have $\text{Cov}^*(h(\xi), \xi | \eta) \geq 0$.

This lemma is extracted from proposition 3 of Hong et al. (2018).

Proof of Lemma 1

Proof We first relate $\beta_j^0$ to $E[\text{Cov}^*[Z_j, P(\delta = 1 | Z)]Z_C]$], then by condition C6, we relate it to $\alpha_j$. For any $j \notin C$ and $k \in C$, straightforward calculations entail that

$$s_k^l(t) = E[Z_k^l \lambda_0(t) \exp(\alpha^T Z) S T S C]$$ and $s_{j,k}^{(l)}(\beta_{C,j}, \beta_j, \tau) = E[Z_k^l \exp(\beta_{C,j}^T Z + Z_j \beta_j) S T S C] (l = 0, 1, 2)$, then

$$u_{j,k}(\beta_{C,j}, \beta_j) = \int_0^\tau E \left[ Z_k - \frac{E[Z_k \exp(\beta_{C,j}^T Z + Z_j \beta_j) S T S C]}{E[\exp(\beta_{C,j}^T Z + Z_j \beta_j) S T S C]} \right] \lambda_0(t) \exp(\alpha^T Z) S T S C \, dt.$$

By the definition, we have

$$v_j(\beta_{C,j}, \beta_j) = u_{j,j}(\beta_{C,j}, \beta_j) - \sum_{k \in C} b_k u_{j,k}(\beta_{C,j}, \beta_j) \equiv F_{1,j}(\beta_{C,j}, \beta_j) - F_{2,j}(\beta_{C,j}, \beta_j).$$

Springer
where

\[
F_1(\beta_{c,j}, \beta_j) = \int_0^T E\{ (Z_j - \sum_{k \in c} b_k Z_k) \lambda_0(t) \exp(\alpha^T Z) S_T S_C \} dt \\
= \int_0^T E\{ [Z_j - E^*(Z_j|Z_C)] \lambda_0(t) \exp(\alpha^T Z) S_T S_C \} dt \\
= E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)],
\]

and

\[
F_2(\beta_{c,j}, \beta_j) = \int_0^T \left[ E\{ Z_j \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} / E\{ \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} \right] \\
- \sum_{k \in c} b_k \left[ E\{ Z_k \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} / E\{ \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} \right] \\
\times E\left\{ \lambda_0(t) \exp(\alpha^T Z) S_T S_C \right\} dt \\
= \int_0^T E\{ [Z_j - E^*(Z_j|Z_C)] \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} / E\{ \exp(\mathbf{Z}_C^T \beta_{c,j} + Z_j \beta_j) S_T S_C \} \\
\times E\left\{ \lambda_0(t) \exp(\alpha^T Z) S_T S_C \right\} dt.
\]

By the definition of \((\beta_{c,j}^0, \beta_j^0)\), we have \(u_{j,k}(\beta_{c,j}^0, \beta_j^0) = 0\) for any \(k \in C \cup \{j\}\). For \(k \in C \cup \{j\}\), \(v_{j,k}(\beta_{c,j}^0, \beta_j^0) = u_{j,k}(\beta_{c,j}^0, \beta_j^0) - \sum_{k \in C} b_k u_{j,k}(\beta_{c,j}^0, \beta_j^0) = 0\), \(F_2(\beta_{c,j}^0, \beta_j^0) = F_1(\beta_{c,j}^0, \beta_j^0) = E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)]\). When \(\alpha_j = 0\), \(E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)] = 0\), thus \(F_2(\beta_{c,j}^0, \beta_j^0) = 0\). Because of \(F_2(\beta_{c,0}, 0) = 0\), \(v_j(\beta_{c,0}, 0) = E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)] - F_2(\beta_{c,0}, 0) = \mathbf{0}_{q+1}\). By the uniqueness of the solution of \(v_j(\beta_C, \beta)\), we have \(\beta_j^0 = 0\).

When \(\alpha_j \neq 0\), by condition C6, we have \(F_2(\beta_{c,j}^0, \beta_j^0) = E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)] \geq c_1 n^{-c}\). This implies that \(F_2(\beta_{c,j}^0, \beta_j^0)\) and \(E[\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)]\) are both nonzero and have the same signs since they are equal. Specifically, \(P(\delta = 1|Z)\) is the probability of occurrence of the event and \(S_T S_C = P(X > t|Z)\) represents the probability at risk at time \(t\). For any \(t\), we have

\[
\frac{\partial P(\delta = 1|Z)}{\partial Z_j} \times \frac{\partial P(X > t|Z)}{\partial Z_j} \leq 0.
\]

By lemma 8, \(\text{Cov}^*(Z_j, P(\delta = 1|Z)|Z_C)\) and \(\text{Cov}^*(Z_j, S_T S_C|Z_C)\) have the opposite signs unless they are zero. This further implies that
\[ F_{2j}(\beta_{C,0}, 0) = \int_0^\tau E\left\{ \frac{E\{\exp(\mathbf{Z}_\tau^T \beta_{C,0})\text{Cov}^*(Z_j, S_T S_C|\mathbf{Z}_C)\}}{E\{\exp(\mathbf{Z}_0^T \beta_{C,0}) S_T S_C\}} \times E\left\{ \lambda_0(t) \exp(\alpha^T \mathbf{Z}) S_T S_C \right\} dt, \]

and \( E[\text{Cov}^*(Z_j, P(\delta = 1|\mathbf{Z})|\mathbf{Z}_C)] \) have opposite signs unless they are equal to zero. So \( F_{2j}(\beta_{C,0}, 0) \neq F_{2j}(\beta_{C,j}^0, \beta_j^0) \), therefore, \( \beta_j^0 \neq 0. \)

**Proof of Lemma 2**

**Proof** By lemma 1, for any \( j \in A_{\cdot C} \), we have \( \beta_j^0 \neq 0 \). By Taylor expansion, there exists \( \tilde{\beta}_j \in (0, \beta_j^0) \) such that

\[ |v_j(\beta_{C,j}^0, 0)| = |v_j(\beta_{C,j}^0, \beta_j^0) - v_j(\beta_{C,j}^0, 0)| = \left| \frac{\partial v_j}{\partial \beta_j}(\beta_{C,j}^0, \tilde{\beta}_j)|\beta_j^0) \right|. \]

By the proof of lemma 1, \( v_j(\beta_{C,j}^0, \beta_j) = E[\text{Cov}^*(Z_j, P(\delta = 1|\mathbf{Z})|\mathbf{Z}_C)] - F_{2j}(\beta_{C,j}^0, \beta_j) \). Given \( \beta_{C,j}^0 \), consider \( F_{2j}(\beta_{C,j}^0, \beta_j) \) as a function of \( \beta_j \), then

\[ \frac{\partial F_{2j}(\beta_{C,j}^0, \beta_j)}{\partial \beta_j} = \int_0^\tau H_j(\beta_{C,j}^0, \beta_j, t) E\left\{ \lambda_0(t) \exp(\alpha^T \mathbf{Z}) S_T S_C \right\} dt \]

\[ = E\left\{ \int_0^\tau H_j(\beta_{C,j}^0, \beta_j, t) S_C dF_T(t|\mathbf{Z}) \right\}, \]

where

\[ H_j(\beta_{C,j}^0, \beta_j, t) = \frac{E[Z_j(Z_j - E^*(Z_j|\mathbf{Z}_C)) \exp(\mathbf{Z}_C^T \beta_{C,j}^0 + Z_j \beta_j) S_T S_C]}{E\{\exp(\mathbf{Z}_C^T \beta_{C,j}^0 + Z_j \beta_j) S_T S_C\}} \]

\[ - \frac{E[Z_j E^*(Z_j|\mathbf{Z}_C)) \exp(\mathbf{Z}_C^T \beta_{C,j}^0 + Z_j \beta_j) S_T S_C] E[Z_j \exp(\mathbf{Z}_C^T \beta_{C,j}^0 + Z_j \beta_j) S_T S_C]}{[E\{\exp(\mathbf{Z}_C^T \beta_{C,j}^0 + Z_j \beta_j) S_T S_C\}]^2}. \]

By condition C3, \( |Z_j| \leq L_0 \), then \( \sup_{\beta_j} |H_j(\beta_{C,j}^0, \beta_j, t)| \leq 2L_0^2 \). So

\[ \left| \frac{\partial v_j}{\partial \beta_j}(\beta_{C,j}^0, \tilde{\beta}_j) \right| \leq \sup_{\beta_j} \left| \frac{\partial F_{2j}(\beta_{C,j}^0, \beta_j)}{\partial \beta_j} \right| \leq 2L_0^2 |E[E(S_C(T)|\mathbf{Z})]| \leq 2L_0^2. \]

By the proof in lemma 1, \( F_{2j}(\beta_{C,j}^0, 0) \) and \( E[\text{Cov}^*(Z_j, P(\delta = 1|\mathbf{Z})|\mathbf{Z}_C)] \) have opposite signs, combining it with condition C6,

\[ |v_j(\beta_{C,j}^0, 0)| = |E[\text{Cov}^*(Z_j, P(\delta = 1|\mathbf{Z})|\mathbf{Z}_C)]| + |F_{2j}(\beta_{C,j}^0, 0)| \geq c_1 n^{-k}. \]
So
\[ |\beta_j^0| = \left| \frac{\partial v_j}{\partial \beta_j} (\beta^0_{Cj}, \hat{\beta}_j) \right|^{-1} \left| v_j (\beta^0_{Cj}, 0) \right| \geq (2L_0^2)^{-1}c_1n^{-\kappa}. \]

Taking \( c_2 = 0.5L_0^{-2}c_1 \), we have
\[ \min_{j \in A_C} |\beta_j^0| \geq c_2n^{-\kappa}, \]
which completes the proof. \( \square \)

**Proof of Lemma 3**

**Proof** Denote \( \hat{U}_j(\beta_{Cj}, \beta_j) = n^{-1}U_j(\beta_{Cj}, \beta_j). \) By the definition of \((\hat{\beta}_{Cj}, \hat{\beta}_j)^T\), we have
\[ \| \hat{U}_j(\beta_{Cj}, \hat{\beta}_j) - \hat{U}_j(\beta^0_{Cj}, \beta^0_j) \| = \| \hat{U}_j(\beta^0_{Cj}, \beta^0_j) \|. \]

For any \( j \notin C \) and \( k \in C \cup \{j\} \), using the similar method of Lin and Wei (1989), by lemmas 4 and 5, we can obtain that
\[ \hat{U}_j(\beta^0_{Cj}, \beta^0_j) = n^{-1} \sum_{i=1}^n W_{i,j}(\beta^0_{Cj}, \beta^0_j) + o_p(1), \]
where \( W_{i,j}(\beta^0_{Cj}, \beta^0_j) \) (\( i = 1, \ldots, n \)) are independent, \( E\{W_{i,j}(\beta^0_{Cj}, \beta^0_j)\} = 0 \) and \( W_{i,j}(\beta^0_{Cj}, \beta^0_j) = [W_{i,j,k}(\beta^0_{Cj}, \beta^0_j), k \in C \cup \{j\}]^T \) with
\[ W_{i,j,k}(\beta^0_{Cj}, \beta^0_j) \]
\[ = \int_0^\tau \left[ Z_{ik} - \frac{E\{Z_{ik} \exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})STSC\}}{E\{\exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})STSC\}} \right] dN_i(t) \]
\[ - \int_0^\tau \frac{Y_i(t) \exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})}{E\{\exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})STSC\}} Z_{ik} \]
\[ - \frac{E\{Z_{ik} \exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})STSC\}}{E\{\exp(\beta^0_{Cj}Z_{i,C} + \beta^0_jZ_{ij})STSC\}} E\{dN_i(t)\}. \]

Let \( E_n \) denote the empirical measure, we can write
\[ \hat{U}_j(\beta^0_{Cj}, \beta^0_j) = E_n[W_{i,j}(\beta^0_{Cj}, \beta^0_j)] + o_p(1). \]

For any given \( i, j, k \), by conditions C1, C3, C5, there exists a constant \( L_2 \) such that \( |W_{i,j,k}(\beta^0_{Cj}, \beta^0_j)| \leq L_2 \), by the fact that \( E[W_{i,j,k}(\beta^0_{Cj}, \beta^0_j)] = 0 \), we have
\[ \text{Var}[W_{i,j,k}(\beta_{C,j}^0, \beta_j^0)] = E[[W_{i,j,k}(\beta_{C,j}^0, \beta_j^0)]^2] \leq L_2^2. \]

By lemma 6, for any \( t > 0, j \notin C \) and \( k \in C \cup \{ j \} \), we have

\[
P \left( \left| E_n(W_{i,j,k}(\beta_{C,j}^0, \beta_j^0)) \right| > \frac{t}{n} \right) \leq 2 \exp \left( -\frac{t^2}{2 n L_2^2 + L_2 t / 3} \right). \]

By Bonferroni inequality, we have

\[
P \left( \| E_n(W_{i,j}(\beta_{C,j}^0, \beta_j^0)) \|_2 > \frac{t(q+1)}{n} \right) \leq 2(q+1) \exp \left( -\frac{t^2}{2 n L_2^2 + L_2 t / 3} \right). \]

As \( \| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) - E_n(W_{i,j}(\beta_{C,j}^0, \beta_j^0)) \|_2 = o_p(1) \), for any \( \epsilon_1 > 0 \) and \( \epsilon_2 > 0 \), there exists \( N_1 \) such that for any \( n > N_1 \), we have

\[
P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) - E_n(W_{i,j}(\beta_{C,j}^0, \beta_j^0)) \right\|_2 > L c_2 \epsilon_1 / 2 \right) < \epsilon_2.
\]

Taking \( t = \frac{c_2 L n^{1-k}}{(q+1)} > 0 \), then \( \frac{t(q+1)}{n} = \frac{c_2 L n^{-k}}{2} \). By Triangle inequality and Bonferroni inequality, we have

\[
P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) \right\|_2 > L c_2 (n^{-k} + \epsilon_1) / 2 \right)
\leq P \left( \left\| E_n(W_{i,j}(\beta_{C,j}^0, \beta_j^0)) \right\|_2 > L c_2 n^{-k} / 2 \right)
+ P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) - E_n(W_{i,j}(\beta_{C,j}^0, \beta_j^0)) \right\|_2 > L c_2 \epsilon_1 / 2 \right)
\leq 2(q+1) \exp \left( -\frac{1}{2 n L_2^2 + L_2 c_2 L n^{1-k} / 6(q+1)} \right) + \epsilon_2.
\]

Taking \( N = \max\{(L_2/3)^{1/k}, N_1\} \), then for any \( n > N, n^{-k} < 3/L_2 \), so we have

\[
P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) \right\|_2 > L c_2 (n^{-k} + \epsilon_1) / 2 \right) \leq 2(q+1) \exp \left( -c_3 n^{1-2k} \right) + \epsilon_2,
\]

where \( c_3 = \frac{c_2^2 L^2}{8 L_2^2(q+1)^2 + 4c_2 L(q+1)} \). By condition C7, we have

\[
P \left( \left| \hat{\beta}_j - \beta_j^0 \right| > c_2 (n^{-k} + \epsilon_1) / 2 \right)
\leq P \left( \left\| \hat{\beta}_{C,j} - (\beta_{C,j}^0, \beta_j^0)^T \right\|_2 > c_2 (n^{-k} + \epsilon_1) / 2 \right)
\leq P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) - \hat{U}_j(\beta_{C,j}^0, \beta_j^0) \right\|_2 > L c_2 (n^{-k} + \epsilon_1) / 2 \right)
= P \left( \left\| \hat{U}_j(\beta_{C,j}^0, \beta_j^0) \right\|_2 > L c_2 (n^{-k} + \epsilon_1) / 2 \right)
\leq 2(q+1) \exp \left( -c_3 n^{1-2k} \right) + \epsilon_2.
\]
Then we have

\[
P \left( \max_{j \in \mathcal{A}_- \cap \mathcal{C}} |\hat{\beta}_j - \beta_j^0| > c_2(n^{-\kappa} + \epsilon_1)/2 \right) \leq 2a(q + 1) \exp(-c_3 n^{-1-2\kappa}) + a\epsilon_2,
\]

where \( a = |\mathcal{A}_- \cap \mathcal{C}| = \sum_{j \notin \mathcal{C}} I(\alpha_j \neq 0) \) is the size of \(|\mathcal{A}_- \cap \mathcal{C}| \).

**Proof of Theorem 1**

**Proof** By the definition of \( \hat{\mathcal{A}}_1 \) and condition \( C_7 \), there exists a positive constant \( c_4 \) such that

\[
P (\mathcal{A}_- \cap \mathcal{C} \subseteq \hat{\mathcal{A}}_1) = P \left( \min_{j \in \mathcal{A}_- \cap \mathcal{C}} |\hat{\beta}_j|/\hat{\sigma}_j \geq \gamma \right) \geq 1 - P \left( \min_{j \in \mathcal{A}_- \cap \mathcal{C}} |\hat{\beta}_j| < n^{-1/2} c_4 \gamma \right).
\]

Following lemma 2, for any \( j \in \mathcal{A}_- \cap \mathcal{C} \), we have \( |\beta_j^0 - \hat{\beta}_j| \geq |\beta_j^0| - |\hat{\beta}_j| \geq c_2 n^{-\kappa} - |\hat{\beta}_j| \).

Suppose \( \min_{j \in \mathcal{A}_- \cap \mathcal{C}} |\hat{\beta}_j| < n^{-1/2} c_4 \gamma \), then \( \max_{j \in \mathcal{A}_- \cap \mathcal{C}} |\beta_j^0 - \hat{\beta}_j| \geq c_2 n^{-\kappa} - n^{-1/2} c_4 \gamma \).

If we have \( \gamma < c_2 (n^{-\kappa} - \epsilon_1) n^{1/2}/(2c_4) \), we can obtain

\[
P \left( \min_{j \in \mathcal{A}_- \cap \mathcal{C}} |\hat{\beta}_j| < n^{-1/2} c_4 \gamma \right) < P \left( \max_{j \in \mathcal{A}_- \cap \mathcal{C}} |\beta_j^0 - \hat{\beta}_j| \geq c_2 (n^{-\kappa} + \epsilon_1)/2 \right).
\]

Then \( P (\mathcal{A}_- \cap \mathcal{C} \subseteq \hat{\mathcal{A}}_1) \geq 1 - 2a(q + 1) \exp(-c_3 n^{-1-2\kappa}) - a\epsilon_2 \). Let \( n \to \infty \), for any \( \epsilon_2 > 0 \), we have \( \lim_{n \to \infty} P (\mathcal{A}_- \cap \mathcal{C} \subseteq \hat{\mathcal{A}}_1) \geq 1 - a\epsilon_2 \), the right side of the above equation does not depend on \( n \) any more. Taking \( \epsilon_2 \to 0 \), we have \( \lim_{n \to \infty} P (\mathcal{A}_- \cap \mathcal{C} \subseteq \hat{\mathcal{A}}_1) = 1 \).

**Proof of Theorem 2**

**Proof** For any \( j \in \mathcal{A}_- \cap \mathcal{C} \), we have \( \alpha_j \neq 0 \). From lemma 1, we know that \( |\beta_j^0| > 0 \). Similarly, we have \( |\beta_j^0| = 0 \) if \( j \notin \mathcal{A}_- \cap \mathcal{C} \). As \( \hat{\beta}_j \) is a consistent estimator of \( \beta_j^0 \) and \( M_{\mathcal{C}, j} = |\hat{\beta}_j|/\hat{\sigma}_j \), we can easily conclude that \( P (\max_{j \notin \mathcal{A}_- \cap \mathcal{C}} M_{\mathcal{C}, j} < \min_{j \in \mathcal{A}_- \cap \mathcal{C}} M_{\mathcal{C}, j}) \to 1 \) when \( n \to \infty \), which completes the proof of theorem 2.

**References**

Andersen PK, Gill RD (1982) Cox’s regression model for counting processes: a large sample study. Ann Statist 10:1100–1120

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50:1064–1072

Barut E, Fan J, Verhasselt A (2016) Conditional sure independence screening. J Am Stat Assoc 111:1266–1277

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J (2000) Exposure stratified case-cohort designs. Lifetime Data Anal 6:39–58

Breslow NE, Wellner JA (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. Scand J Stat 34:86–102
Candes E, Tao T (2007) The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. Ann Stat 35:2313–2351
Chang J, Tang CY, Wu Y (2013) Marginal empirical likelihood and sure independence feature screening. Ann Stat 41:2123–2148
Chen K (2001) Generalized case-cohort sampling. J R Stat Soc B 63:791–809
Chen K, Lo SH (1999) Case-cohort and case-control analysis with Cox’s model. Biometrika 86:755–764
Cox DR (1972) Regression models and life-tables. J R Stat Soc B 34:187–220
Cui H, Li R, Zhong W (2015) Model-free feature screening for ultrahigh dimensional discriminant analysis. J Am Stat Assoc 110:630–641
Fan J, Feng Y, Song R (2011) Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Am Stat Assoc 106:544–557
Fan J, Feng Y, Wu Y (2010) High-dimensional variable selection for Cox’s proportional hazards model. In: Borrowing strength: theory powering applications: a Festschrift for Lawrence D. Brown, Institute of Mathematical Statistics 6:70–86
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 70:849–911
Fan J, Ma Y, Dai W (2014) Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. J Am Stat Assoc 109:1270–1284
Fan J, Samworth R, Wu Y (2009) Ultrahigh dimensional feature selection: beyond the linear model. J Mach Learn Res 10:2013–2038
Fan J, Song R (2010) Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat 38:3567–3604
Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York
Gorst-Rasmussen A, Scheike T (2013) Independent screening for single-index hazard rate models with ultrahigh dimensional features. J R Stat Soc B 75:217–245
He X, Wang L, Hong HG (2013) Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann Stat 41:342–369
Hong HG, Kang J, Li Y (2018) Conditional screening for ultra-high dimensional covariates with survival outcomes. Lifetime Data Anal 24:45–71
Hong HG, Wang L, He X (2016) A data-driven approach to conditional screening of high-dimensional variables. Stat 5:200–212
Hu Q, Lin L (2017) Conditional sure independence screening by conditional marginal empirical likelihood. Ann Inst Stat Math 69:63–96
Kalbfleisch JD, Lawless JF (1988) Likelihood analysis of multi-state models for disease incidence and mortality. Stat Med 7:149–160
Kang S, Cai J (2009) Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika 96:887–901
Keogh RH, White IR (2013) Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. Stat Med 32:4021–4043
Kim S, Ahn WK (2019) Bi-level variable selection for case-cohort studies with group variables. Stat Methods Med Res 28:3404–3414
Kim S, Cai J, Lu W (2013) More efficient estimators for case-cohort studies. Biometrika 100:695–708
Kulich M, Lin D (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc 99:832–844
Li G, Peng H, Zhang J, Zhu L (2012a) Robust rank correlation based screening. Ann Stat 40:1846–1877
Li R, Zhong W, Zhu L (2012b) Feature screening via distance correlation learning. J Am Stat Assoc 107:1129–1139
Lin DY, Wei LJ (1989) The robust inference for the Cox proportional hazards model. J Am Stat Assoc 84:1074–1078
Lin Y, Liu X, Hao M (2018) Model-free feature screening for high-dimensional survival data. Sci China Math 61:1617–1636
Liu Y, Chen XL (2018) Quantile screening for ultra-high-dimensional heterogeneous data conditional on some variables. J Stat Comput Sim 88:329–342
Liu J, Li R, Wu R (2014) Feature selection for varying coefficient models with ultrahigh-dimensional covariates. J Am Stat Assoc 109:266–274

Springer
Liu Y, Wang Q (2018) Model-free feature screening for ultrahigh-dimensional data conditional on some variables. Ann Inst Stat Math 70:283–301
Liu Y, Zhang J, Zhao X (2018) A new nonparametric screening method for ultrahigh-dimensional survival data. Comput Stat Data Anal 119:74–85
Lu J, Lin L (2020) Model-free conditional screening via conditional distance correlation. Stat Pap 61:225–244
Mai Q, Zou H (2015) The fused Kolmogorov filter: a nonparametric model-free screening method. Ann Stat 43:1471–1497
Marti H, Chavance M (2011) Multiple imputation analysis of case-cohort studies. Stat Med 30:1595–1607
Ni A, Cai J, Zeng D (2016) Variable selection for case-cohort studies with failure time outcome. Biometrika 103:547–562
Pan W, Wang X, Xiao W, Zhu H (2019) A generic sure independence screening procedure. J Am Stat Assoc 114:928–937
Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11
Scheike TH, Martinussen T (2004) Maximum likelihood estimation for Cox’s regression model under case-cohort sampling. Scand J Stat 31:283–293
Self SG, Prentice R (1988) Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Stat 16:64–81
Song R, Lu W, Ma S, Jeng XJ (2014) Censored rank independence screening for high-dimensional survival data. Biometrika 101:799–814
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288
Tibshirani R (2009) Univariate shrinkage in the Cox model for high dimensional data. Stat Appl Genet Mol Biol 8:1–18
Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30:1105–1117
van de Vijver MJ, He YD, van Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ (2002) A gene-expression signature as a predictor of survival in breast cancer. New Engl J Med 347:1999–2009
van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes. Springer, New York
van Veer LJ, van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
Wu Y, Yin G (2015) Conditional quantile screening in ultrahigh-dimensional heterogeneous data. Biometrika 102:65–76
Yeung KY, Bumgarner RE, Raftery AE (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21:2394–2402
Zeng D, Lin DY (2014) Efficient estimation of semiparametric transformation models for two-phase cohort studies. J Am Stat Assoc 109:371–383
Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942
Zhang J, Liu Y, Wu Y (2017) Correlation rank screening for ultrahigh-dimensional survival data. Comput Stat Data Anal 108:121–132
Zhang J, Yin G, Liu Y, Wu Y (2018) Censored cumulative residual independent screening for ultrahigh-dimensional survival data. Lifetime Data Anal 24:273–292
Zhao SD, Li Y (2012) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. J Mult Anal 105:397–411
Zhou T, Zhu L (2017) Model-free feature screening for ultrahigh dimensional censored regression. Stat Comput 27:947–961
Zhu LP, Li L, Li R, Zhu LX (2011) Model-free feature screening for ultrahigh-dimensional data. J Am Stat Assoc 106:1464–1475
Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.