Annals of the Institute of Statistical Mathematics (2022) 74:515–537
https://doi.org/10.1007/s10463-021-00807-1

The variable selection by the Dantzig selector for Cox's proportional hazards model

Kou Fujimori[1]

Received: 22 October 2019 / Revised: 23 May 2021 / Accepted: 30 July 2021 /
Published online: 31 August 2021

Abstract
The proportional hazards model proposed by D. R. Cox in a high-dimensional and sparse setting is discussed. The regression parameter is estimated by the Dantzig selector, which will be proved to have the variable selection consistency. This fact enables us to reduce the dimension of the parameter and to construct asymptotically normal estimators for the regression parameter and the cumulative baseline hazard function.

Keywords Proportional hazards model · Dantzig selector · Variable selection · The Breslow estimator

1 Introduction

The proportional hazards model, which was proposed by Cox (1972), is one of the most commonly used models for survival analysis. In a fixed-dimensional setting, i.e., the case where the number of covariates p is fixed, Andersen and Gill (1982) proved that the maximum partial likelihood estimator for the regression parameter has the consistency and the asymptotic normality. Besides, they discussed the asymptotic property of the Breslow estimator for the cumulative baseline hazard function.

Recently, many researchers are interested in a high-dimensional and sparse setting for a regression parameter, that is, the case where $p = p_n \to \infty$ as $n \to \infty$ and the number q of nonzero components in the true value is relatively small. In this setting, several kinds of estimation methods have been proposed for various regression-type models. Particularly, the penalized methods such as Lasso [(Tibshirani, 1997; Huang et al., 2013; Bradic et al., 2011) among others] have been

✉ Kou Fujimori
kfujimori@shinshu-u.ac.jp

[1] Department of Economics, Faculty of Economics and Law, Shinshu University, 3-1-1, Asahi, Matsumoto City, Nagano 390-8621, Japan

well studied. In particular, Huang et al. (2013) derived oracle inequalities of the Lasso estimator for the proportional hazards model, which means the Lasso estimator satisfies the consistency even in a high-dimensional setting. Bradic et al. (2011) considered the general penalized estimators including Lasso, SCAD and others and proved that the estimators satisfy the consistency and the asymptotic normality. On the other hand, the Dantzig selector, which was proposed by Candés and Tao (2007) for the linear regression model, is also applied to the proportional hazards model by Antoniadis et al. (2010), who dealt with the l_2 consistency of the estimator. Fujimori and Nishiyama (2017) extended the consistency results of the Dantzig selector for the model to the l_q consistency for every $q \in [1, \infty]$ by a method similar to that of Huang et al. (2013). However, the asymptotic normalities of the Dantzig selector for high-dimensional regression parameter and the Breslow estimator have not yet been studied up to our knowledge.

We establish the asymptotic normalities of estimators in a high-dimensional and sparse setting based on the consistency results from Fujimori and Nishiyama (2017). To discuss this problem, we need to consider the dimension reduction of the regression parameter, which is nearly equivalent to consider the variable selection for a high-dimensional and sparse regression parameter of the proportional hazards model. The variable selection methods for the proportional hazards model in high-dimensional and sparse settings are also discussed by some researchers. For example, Honda and Härdle (2013) studied the group SCAD-type and adaptive group Lasso estimators for time-varying coefficients in the proportional hazards model and proved that these estimators achieve the variable selection. The variable selection consistency, in particular, the sign consistency of the Dantzig selector for the regression models, is proved under some technical conditions called the *irrepresentable condition*, which is derived from KKT condition of the optimization problem, see, e.g., Fan et al. (2016). Since the proportional hazards model is a nonlinear model, KKT condition for the Dantzig selector is too complicated, which implies that the irrepresentable condition becomes also complicated. In this paper, we prove the variable selection consistency by constructing the estimator for the support index set, i.e., the index set of the nonzero components of the regression parameter without conditions such as the irrepresentable condition. Next, we construct a new maximum partial likelihood estimator by using the variable selection consistency result and show that this estimator has the asymptotic normality. In addition, we will construct the asymptotically normal estimator for cumulative baseline hazard function by Breslow-type estimator. Moreover, we observe whether our selection criterion works well for simple models numerically and compare the estimators to the classical maximum partial likelihood estimator.

The novelties of this paper are as follows. First, the consistency of the Dantzig selector for the proportional hazards model is proved under the condition that the number of nonzero components of regression parameter allows to diverge, which is the extension of the results from Antoniadis et al. (2010), Fujimori and Nishiyama (2017). Second, the variable selection consistency of the Dantzig selector for Cox's proportional hazards model is proved without some conditions such as the irrepresentable condition. Third, the asymptotically normal estimator for regression

parameter is constructed by the dimension reduction via the Dantzig selector. Finally, we provide an intuitive method to choose tuning parameter by an iterated algorithm.

The rest of the paper is organized as follows. The model setup, some regularity conditions and matrix conditions to deal with a high-dimensional and sparse setting are introduced in Sect. 2. In Sect. 3, we prove the asymptotic properties of the estimators for the regression parameter, that is, the variable selection consistency of the Dantzig selector. The asymptotic normality of the maximum partial likelihood estimator and the Breslow estimator after dimension reduction is established in Sect. 4. In Sect. 5, we introduce the intuitive method to choose the tuning parameter which is used to construct the Dantzig selector. We present an algorithm to compute the Dantzig selector for the proportional hazards model, which is essentially introduced by Antoniadis et al. (2010) and some simulation studies for simple models to verify the variable selection consistency of the Dantzig selector in Sect. 6. This section also includes the real data application to the gene expression data. Some comments about the differences between Lasso and the Dantzig selector for the proportional hazards model are given in Sect. 7. The proofs for theorems are given in Sect. 8.

Throughout this paper, we denote by $\|\cdot\|_q$ the l_q norm of vector for every $q \in [1, \infty]$, i.e., for $v = (v_1, v_2, \ldots, v_p)^\top \in \mathbb{R}^p$, we define:

$$\|v\|_q = \left(\sum_{j=1}^{p} |v_j|^q \right)^{\frac{1}{q}}, \quad q < \infty;$$
$$\|v\|_\infty = \sup_{1 \le j \le p} |v_j|.$$

In addition, for a $m \times n$ matrix A, where $m,\ n \in \mathbb{N}$, we define $\|A\|_\infty$ by

$$\|A\|_\infty := \sup_{1 \le i \le m} \sup_{1 \le j \le n} |A_{ij}|,$$

where A_{ij} denotes the (i, j)-component of the matrix A. For a vector $v \in \mathbb{R}^p$, and an index set $T \subset \{1, 2, \ldots, p\}$, we denote the $|T|$-dimensional sub-vector of v restricted by the index set T by v_T, where $|T|$ is the number of elements of the set T. Similarly, for a $p \times p$ matrix A and index sets $T, T' \subset \{1, 2, \ldots, p\}$, we define the $|T| \times |T'|$ submatrix $A_{T,T'}$ by

$$A_{T,T'} := (A_{ij})_{i \in T, j \in T'}.$$

2 Preliminaries

Let T_i be a survival time and C_i a censoring time of i-th individual for every $i = 1, 2, \ldots, n$, which are positive real-valued random variables on a probability space $(\Omega, \mathcal{F}, P)$. Assume that each i-th individual has an $\mathbb{R}^p$-valued covariate process $\{Z_i(t)\}_{t \in [0,1]}$, and that the survival time T_i is conditionally independent of the censoring time C_i given $Z_i(t)$. Moreover, we assume that T_i's never occur

simultaneously. For every $n \in \mathbb{N}$ and $t \in [0,1]$, we observe $\{(X_i, D_i, Z_i(t))\}_{i=1}^n$, where $X_i := T_i \wedge C_i$ and $D_i := 1_{\{T_i \le C_i\}}$. We define the counting process $\{N_i(t)\}_{t\in[0,1]}$ and $\{Y_i(t)\}_{t\in[0,1]}$ for every $i = 1, 2, \dots, n$ as follows:

$$N_i(t) := 1_{\{t \ge X_i, D_i = 1\}}, \quad Y_i(t) := 1_{\{X_i \ge t\}}, \quad t \in [0,1].$$

Let $\{\mathcal{F}_t\}_{t\in[0,1]}$ be the filtration defined as follows:

$$\mathcal{F}_t := \sigma\{N_i(u),\ Y_i(u),\ Z_i(u); 0 \le u \le t,\ i = 1, 2, \dots, n\}.$$

Suppose that $\{Z_i(t)\}_{t\in[0,1]}$, $i = 1, 2, \dots, n$ are predictable and bounded processes. In Cox's proportional hazards model, it is assumed that each $\{N_i(t)\}_{t\in[0,1]}$ for every $i = 1, 2, \dots, n$ has the following intensity:

$$\lambda_i(t) := Y_i(t)\lambda_0(t)\exp(\beta_0^\top Z_i(t)), \quad t \in [0,1],$$

where $\lambda_0 \in L^1[0,1]$ is the unknown deterministic baseline hazard function and $\beta_0 \in \mathbb{R}^p$ is the unknown regression parameter. We have that the following process $\{M_i(t)\}_{t\in[0,1]}$ for every $i = 1, 2, \dots, n$ is a square integrable martingale:

$$M_i(t) := N_i(t) - \int_0^t \lambda_i(s)\mathrm{d}s, \quad t \in [0,1].$$

Note that predictable variation process of $\{M_i(t)\}_{t\in[0,1]}$ is given by:

$$\langle M_i, M_i\rangle(t) = \int_0^t \lambda_i(s)\mathrm{d}s, \quad t \in [0,1]$$

and

$$\langle M_i, M_j\rangle(t) = 0, \quad i \ne j,\ t \in [0,1].$$

Hereafter, we write Λ_0 for the cumulative baseline hazard function, i.e.,

$$\Lambda_0(t) := \int_0^t \lambda_0(s)\mathrm{d}s, \quad t \in [0,1].$$

The aim of this paper is to estimate the regression parameter β_0 and the cumulative baseline hazard Λ_0 in a high-dimensional and sparse setting for β_0, i.e., $p = p_n \to \infty$ as $n \to \infty$ and $|T_0^n| =: q_n$ which is allowed to tend to infinity as $n \to \infty$, where $T_0^n := \{j; \beta_0^j \ne 0\}$ is the support index set of the true value. To estimate β_0, we use Cox's log-partial likelihood which is given by;

$$C_n(\beta) := \sum_{i=1}^n \int_0^1 \{\beta^\top Z_i(t) - \log S_n^{(0)}(\beta, t)\}\mathrm{d}N_i(t),$$

where

$$S_n^{(0)}(\beta, t) := \sum_{i=1}^{n} Y_i(t) \exp(\beta^\top Z_i(t)).$$

Put $l_n(\beta) = C_n(\beta)/n$. We write $U_n(\beta)$ for the gradient of $l_n(\beta)$ and $J_n(\beta)$ for the Hessian of $-l_n(\beta)$, i.e.,

$$U_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left\{ Z_i(t) - \frac{S_n^{(1)}}{S_n^{(0)}}(\beta, t) \right\} \mathrm{d}N_i(t)$$

and

$$J_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left\{ \frac{S_n^{(2)}}{S_n^{(0)}}(\beta, t) - \left(\frac{S_n^{(1)}}{S_n^{(0)}} \right)^{\otimes 2} (\beta, t) \right\} \mathrm{d}N_i(t),$$

where

$$S_n^{(1)}(\beta, t) := \sum_{i=1}^{n} Z_i(t) Y_i(t) \exp(\beta^\top Z_i(t))$$

and

$$S_n^{(2)}(\beta, t) := \sum_{i=1}^{n} Z_i(t)^{\otimes 2} Y_i(t) \exp(\beta^\top Z_i(t)).$$

Note that $U_n(\beta_0)$ is a terminal value of the following square integrable martingale:

$$U_n(\beta_0, t) := \frac{1}{n} \sum_{i=1}^{n} \int_0^t \left\{ Z_i(s) - \frac{S_n^{(1)}}{S_n^{(0)}}(\beta, s) \right\} \mathrm{d}M_i(s).$$

We assume the following conditions.

Assumption 1 (i) The covariate processes $\{Z_i(t)\}_{t\in[0,1]}$, $i = 1, 2, \ldots, n$, are uniformly bounded, i.e., there exists global constant $K_1 > 0$ such that

$$\sup_{t\in[0,1]} \sup_i \|Z_i(t)\|_\infty < K_1 \quad a.s.$$

(ii) The baseline hazard function λ_0 is integrable, i.e.,

$$\int_0^1 \lambda_0(t) \mathrm{d}t < \infty.$$

(iii) For every $n \in \mathbb{N}$, there exist deterministic $\mathbb{R}$-valued function $s_n^{(0)}(\beta, t)$, $\mathbb{R}^{p_n}$-valued function $s_n^{(1)}(\beta, t)$ and $\mathbb{R}^{p_n \times p_n}$-valued function $s_n^{(2)}(\beta, t)$ which satisfy the following conditions:

$$\sup_{\beta}\sup_{t\in[0,1]}\left\|\frac{1}{n}S_n^{(l)}(\beta,t)-s_n^{(l)}(\beta,t)\right\|_{\infty}\to^p 0,\quad l=0,1,2$$

as $n\to\infty$.

(iv) The functions $s_n^{(l)}(\beta,t)$, $l=0,1,2$, satisfy the following conditions:

$$\limsup_{n\to\infty}\sup_{\beta}\sup_{t\in[0,1]}\|s_n^{(l)}(\beta,t)\|_{\infty}<\infty,\quad l=0,1,2,$$
$$\liminf_{n\to\infty}\inf_{\beta}\inf_{t\in[0,1]}s_n^{(0)}(\beta,t)>0.$$

(v) For every β, the following $p_n\times p_n$ matrix $I_n(\beta)$ is nonnegative definite:

$$I_n(\beta):=\int_0^1\left[\frac{s_n^{(2)}}{s_n^{(0)}}(\beta,t)-\left(\frac{s_n^{(1)}}{s_n^{(0)}}\right)^{\otimes 2}(\beta,t)\right]s_n^{(0)}(\beta_0,t)\lambda_0(t)\mathrm{d}t.$$

Recalling that $T_0^n=\{j;\beta_0^j\neq 0\}$ is the support index set of the true value β_0, we introduce the following factor for the matrix $I_n(\beta_0)$.

Definition 2 Define the following factors for the matrix $J_n(\beta_0)$.

Compatibility factor

$$\kappa(T_0^n;J_n(\beta_0))=\inf_{0\neq h\in C_{T_0^n}}\frac{q_n^{\frac{1}{2}}(h^\top J_n(\beta_0)h)^{\frac{1}{2}}}{\|h_{T_0^n}\|_1}.$$

l_∞-cone invertibility factor:

$$F_\infty(T_0^n;J_n(\beta_0)):=\inf_{0\neq h\in C_{T_0^n}}\frac{h^\top J_n(\beta_0)h}{\|h_{T_0^n}\|_1\|h\|_\infty},$$

where the set $C_{T_0^n}\subset\mathbb{R}^{p_n}$ is defined as follows:

$$C_{T_0^n}:=\{h\in\mathbb{R}^{p_n};\|h_{(T_0^n)^c}\|_1\le\|h_{T_0^n}\|_1\}.$$

Note that $h_{(T_0^n)^c}$ and $h_{T_0^n}$ are the p_n-q_n- and q_n-dimensional sub-vector of $h\in\mathbb{R}^{p_n}$ constructed by extracting h corresponding to the indices in the index set $(T_0^n)^c$ and T_0^n, respectively, as we mentioned in Introduction. This factor is a modification of l_q cone invertibility factor:

$$F_q(T_0^n;M):=\inf_{0\neq h\in C_{T_0^n}}\frac{q_n^{1/q}h^\top Mh}{\|h_{T_0^n}\|_1\|h\|_q},\quad q\ge 1,$$

for a matrix M, which is given by Huang et al. (2013). The matrix factors can be seen in many papers which deal with high-dimensional and sparse setting. See, e.g., Bickel et al. (2009), van de Geer and Bühlmann (2009), Huang et al. (2013) among others for the details.

To verify the l_∞ consistency of the Dantzig selector defined in the next section, we assume the following condition for $J_n(\beta_0)$ as well as Huang et al. (2013) among other studies.

Assumption 3 For every $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$P\big(\kappa(T_0^n, J_n(\beta_0)) > \delta\big) \geq 1 - \epsilon$$

and

$$P\big(F_\infty(T_0^n, J_n(\beta_0)) > \delta\big) \geq 1 - \epsilon.$$

3 The estimator for the regression parameter

3.1 The Dantzig selector for the proportional hazards model

Now, we define the estimator for the regression parameter β_0 by the Dantzig selector for the proportional hazards model given by:

$$\hat{\beta}_n := \arg\min_{\beta \in \mathcal{B}_n} \|\beta\|_1, \quad \mathcal{B}_n := \{\beta \in \mathbb{R}^{p_n}; \|U_n(\beta)\|_\infty \leq \gamma\}, \tag{1}$$

where γ is a tuning parameter. This type of estimator was proposed by Antoniadis et al. (2010) and was further discussed by Fujimori and Nishiyama (2017).

Hereafter, we assume that the dimension p_n and the sparsity q_n of the parameter satisfy the following conditions.

Assumption 4 It holds that

$$\log p_n = O(n^a), \quad a \in \left(0, \frac{1}{2}\right). \tag{2}$$

Moreover, the tuning parameter γ_n and the sparsity q_n satisfy that

$$\gamma_n = c\frac{\log p_n}{n^\alpha}, \quad \alpha \in \left(a, \frac{1}{2}\right), \tag{3}$$

where $c > 0$ is a positive constant and that

$$q_n\gamma_n = o(1), \quad n \to \infty.$$

Antoniadis et al. (2010), Fujimori and Nishiyama (2017) assume that the sparsity q_n is independent of n and finite. In contrast, we consider the case where q_n allows to diverge and prove the consistency and the variable selection consistency of the

Dantzig selector. The next lemma implies the consistency of the Dantzig selector. Since it can be proved as well as Lemmas 4.1 in Fujimori and Nishiyama (2017) or corresponding results from Bradic et al. (2011) by using the concentration inequality established by van de Geer (1995) and the maximal inequality provided in van der Vaart and Wellner (1996), the proof is omitted.

Lemma 5 *Under Assumptions 1 and 4, it holds that*

$$\lim_{n\to\infty} P\big(\|U_n(\beta_0)\|_\infty \geq \gamma_n\big) = 0.$$

The following theorem states the l_1 and l_∞ consistency of the estimator $\hat{\beta}_n$. It can be proved as well as the corresponding result in Huang et al. (2013).

Theorem 6 *Under Assumptions 1, 3 and 4, it holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_1 \geq \frac{K_3 q_n \gamma_n}{\kappa^2(T_0^n; J_n(\beta_0))} \right) = 0 \tag{4}$$

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_\infty \geq \frac{K_4 \gamma_n}{F_\infty(T_0^n; J_n(\beta_0))} \right) = 0, \tag{5}$$

where K_3 and K_4 are some constants.

3.2 The variable selection consistency of the Dantzig selector

The aim of this subsection is to show that $\hat{\beta}_n$ selects nonzero components of β_0 correctly. To do this, we define the following estimator for the support index set T_0^n of the true value β_0:

$$\hat{T}_n := \{j; |\hat{\beta}_n^j| > \gamma_n\}. \tag{6}$$

The estimator similar to $\hat{T}_n$ can be seen in Fujimori (2019) which considers a linear model of diffusion processes in a high-dimensional and sparse setting. The following theorem states that $\hat{\beta}_n$ has a variable selection consistency.

Theorem 7 *Suppose that*

$$\liminf_{n\to\infty} \inf_{j\in T_0^n} |\beta_0^j| n^\zeta > 0, \tag{7}$$

where $0 < \zeta < \alpha - a$ with α and a appeared in (2) and (3). Under Assumptions1, 3 and 4, it holds that

$$\lim_{n\to\infty} P(\hat{T}_n = T_0^n) = 1.$$

Remark 8 If there exists a constant $c > 0$ such that $\beta_0^j > c$, $j \in T_0^n$ or the sparsity q_n is finite, the condition (7) is valid.

4 After the variable selection

4.1 The maximum likelihood estimator after the selection

Hereafter, we assume that the sparsity satisfies that

$$q_n = S$$

for some positive constant S which is independent of n. In this case, we write the index set $T_0^n = T_0$, since it does not depend on n. Using the set $\hat{T}_n$, we construct a new estimator $\hat{\beta}_n^{(2)}$ by the solution to the next equation:

$$U_n(\beta)_{\hat{T}_n} = 0, \quad \beta_{\hat{T}_n^c} = 0. \tag{8}$$

We prove the asymptotic normality of $\hat{\beta}_n^{(2)}$. In this subsection, we impose the following assumption.

Assumption 9

(i) For every $\epsilon > 0$, it holds that

$$\sum_{i=1}^{n} \int_0^1 \left\| \xi_{nT_0,i} \right\|_2^2 1_{\{\|\xi_{nT_0,i}\|^2 > \epsilon\}} Y_i(t) \exp(\beta_0^\top Z_i(t)) \lambda_0(t) \mathrm{d}t \to^p 0,$$

where

$$\xi_{nT_0,i} := \frac{1}{\sqrt{n}} \left\{ Z_{iT_0}(t) - \frac{S_{nT_0}^{(1)}}{S_n^{(0)}}(\beta_{0T_0}, t) \right\}.$$

(ii) The following $S \times S$ matrix $\mathcal{I}$ is positive definite:

$$\mathcal{I} := \int_0^1 \left[\frac{s^{(2)}}{s^{(0)}}(\beta_{0T_0}, s) - \left(\frac{s^{(1)}}{s^{(0)}} \right)^{\otimes 2} (\beta_{0T_0}, s) \right] \lambda_0(s) s^{(0)}(\beta_{0T_0}, s) \mathrm{d}s,$$

where

$$s^{(0)}(\beta_{0T_0}, t) := s_n^{(0)}(\beta_{0T_0}, t),$$
$$s^{(1)}(\beta_{0T_0}, t) := s_{nT_0}^{(1)}(\beta_{0T_0}, t)$$

and

$$s^{(2)}(\beta_{0T_0}, t) := s_{nT_0,T_0}^{(2)}(\beta_{0T_0}, t).$$

The condition (i) is the Lindeberg condition imposed to derive the asymptotic normality, and the condition (ii) is agree with the condition in Fleming and Harrington (1991) when p is fixed. Note that $\mathcal{I}$ is a sub-matrix of $I_n(\beta_0)$. We can prove that $I_n(\beta_0)$ approximates the Hessian matrix in the following sense.

Lemma 10 *Under Assumption 1, it holds that*

$$\|J_n(\beta_n^*) - I_n(\beta_0)\|_\infty = o_p(1)$$

for every random sequence $\{\beta_n^*\}_{n\in\mathbb{N}}$ *which satisfies that*

$$\|\beta_n^* - \beta_0\|_1 \to^p 0$$

as $n \to \infty$.

Since we can prove Lemma 10 as well as Lemma 4.2 in Fujimori and Nishiyama (2017), the proof of this lemma is omitted in this paper.

Noticing the result in Lemma 10, we have that the new estimator $\hat{\beta}_n^{(2)}$ is well defined.

Lemma 11 *Under Assumptions 1 and 9, the solution to Eq. (8) exists with probability tending to* 1 *as* $n \to \infty$, *i.e., the estimator* $\hat{\beta}_n^{(2)}$ *is well defined.*

The following theorem states that this estimator $\hat{\beta}_n^{(2)}$ satisfies l_1 consistency.

Theorem 12 *Under Assumptions 1 and 3, it holds that*

$$\|\hat{\beta}_n^{(2)} - \beta_0\|_1 \to^p 0$$

as $n \to \infty$.

Now, we can prove the asymptotic normality in the following sense by a similar way to that in Andersen and Gill (1982).

Theorem 13 *Under Assumptions 1 and 3, it holds that*

$$\sqrt{n}(\hat{\beta}_{n\hat{T}_n}^{(2)} - \beta_{0T_0})1_{\{\hat{T}_n = T_0\}} \to^d N(0, \mathcal{I}^{-1}).$$

Remark 14 We may prove the consistency and asymptotic normality of $\hat{\beta}_n^{(2)}$ for the case when the sparsity $q_n \to \infty$ as $n \to \infty$ by the similar way to those in Bradic et al. (2011) under some suitable conditions. In such cases, the asymptotic normality is written in the following form:

$$\sqrt{n} b_n^\top \mathcal{I}^{\frac{1}{2}} (\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0}) 1_{\{\hat{T}_n = T_0\}} \to^d N(0,1), \quad n \to \infty,$$

for every l_2 unit vector $b_n \in \mathbb{R}^{q_n}$.

4.2 The estimator for the cumulative baseline hazard function

We define the estimator for $\Lambda_0(t)$ by the following Breslow-type estimator:

$$\hat{\Lambda}(t) := \int_0^t \frac{\mathrm{d}\bar{N}(s)}{\sum_{i=1}^n Y_i(s) \exp(\hat{\beta}_n^{(2)T} Z_i(s))}, \quad t \in [0,1], \tag{9}$$

where $\hat{\beta}_n^{(2)}$ is defined by Eq. (8). We discuss the asymptotic property of $\hat{\Lambda}$ in this section. For every $t \in [0,1]$, we have that

$$\sqrt{n}\{\hat{\Lambda}(t) - \Lambda_0(t)\} = (I) + (II) + (III),$$

where

$$(I) = \sqrt{n} \int_0^t \left\{ \frac{1}{S_n^{(0)}(\hat{\beta}_n^{(2)}, s)} - \frac{1}{S_n^{(0)}(\beta_0, s)} \right\} \mathrm{d}\bar{N}(s),$$

$$(II) = \sqrt{n} \left\{ \int_0^t \frac{\mathrm{d}\bar{N}(s)}{S_n^{(0)}(\beta_0, s)} - \int_0^t \lambda_0(s) 1_{\{\sum_{i=1}^n Y_i(s) > 0\}} \right\}$$

and

$$(III) = \sqrt{n} \left\{ \int_0^t \lambda_0(s) 1_{\{\sum_{i=1}^n Y_i(s) > 0\}} - \Lambda_0(t) \right\}.$$

The third term (III) is asymptotically negligible because it follows from Assumption 1 that

$$\lim_{n \to \infty} P\left(\left\{ \int_0^t \lambda_0(s) 1_{\{\sum_{i=1}^n Y_i(s) > 0\}} \mathrm{d}s - \Lambda_0(t) \right\} = 0 \right) = 1.$$

Moreover, we have that (II) is equal to the following process $\{W_n(t)\}_{t \in [0,1]}$:

$$W_n(t) = \sqrt{n} \int_0^t \frac{d\bar{M}(s)}{S_n^{(0)}(\beta_0, s)},$$

which is a square integrable martingale. Using the Taylor expansion, we have that

$$(I) = H_n(\beta_n^*, t)^\top (\hat{\beta}_n^{(2)} - \beta_0),$$

where

$$H_n(\beta_n^*, t) := -\int_0^t \frac{S_n^{(1)}}{\{S_n^{(0)}\}^2}(\beta_n^*, s) d\bar{N}(s)$$

and β_n^* lies between $\hat{\beta}_n^{(2)}$ and β_0. Since it holds that $\|\beta_n^* - \beta_0\|_1 = o_p(1)$ by Theorem 12, we can see that

$$\sup_{t\in[0,1]} \left\| H_n(\beta_n^*, t) + \int_0^t \frac{s_n^{(1)}}{s_n^{(0)}}(\beta_0, s)\lambda_0(s)\mathrm{d}s \right\|_\infty = o_p(1) \tag{10}$$

by a similar way to the proof of Lemma 10. Therefore, we obtain the following theorem, which is proved by using Slutsky's theorem and a similar way to that in Andersen and Gill (1982).

Theorem 15 *Under Assumptions 1 and 3, it holds that* $\sqrt{n}(\hat{\beta}_{n\hat{T}_n}^{(2)} - \beta_{0T_0})1_{\{\hat{T}_n = T_0\}}$ *and the process equal in the point t to*

$$\left[\sqrt{n}\{\hat{\Lambda}(t) - \Lambda_0(t)\} + \sqrt{n}\int_0^t (\hat{\beta}_{n\hat{T}_n}^{(2)} - \beta_{0T_0})^\top \frac{s^{(1)}}{s^{(0)}}(\beta_{0T_0}, s)\lambda_0(s)\mathrm{d}s\right] 1_{\{\hat{T}_n = T_0\}}$$

is asymptotically independent. The latter process is asymptotically distributed as a Gaussian martingale with the variance function

$$\int_0^t \frac{\lambda_0(s)}{s^{(0)}(\beta_{0T_0}, s)}\mathrm{d}s.$$

5 Some discussion on the tuning parameter

In this section, we present some comments of the tuning parameter because our theoretical results strongly depend on the choice of the tuning parameter. Recall that the Dantzig selector $\hat{\beta}_n$ is defined by the following form

$$\hat{\beta}_n := \arg\min_{\beta\in\mathcal{B}_n} \|\beta\|_1, \quad \mathcal{B}_n = \{\beta \in \mathbb{R}^{p_n} : \|U_n(\beta)\|_\infty \le \gamma_n\},$$

where $\gamma_n \ge 0$ is a tuning parameter and $U_n(\cdot)$ is the score function. In this paper, we assume that the tuning parameter γ_n satisfies the following condition:

$$\gamma_n = c\tilde{\gamma}_n,$$

where $c \geq 0$ is a constant not depending on n and $\tilde{\gamma}_n = \log p_n / n^{\alpha},\ \alpha \in (a, 1/2)$. To ensure the l_∞ consistency and the variable selection consistency for finite sample, c has to satisfy that

$$\frac{\|U_n(\beta_0)\|_\infty}{\tilde{\gamma}_n} \leq c \leq \frac{\inf_{j \in T_0} |\beta_0^j|}{\tilde{\gamma}_n}, \tag{11}$$

where T_0 is the support index set of the true value θ_0. The problem is how to choose c which satisfies (11). Note that

$$\frac{\|U_n(\beta_0)\|_\infty}{\tilde{\gamma}_n} = O_p(1)$$

and that

$$\frac{\inf_{j \in T_0^i} |\beta_0^j|}{\tilde{\gamma}_n} \to \infty$$

as $n \to \infty$ under our setting. When we choose the small tuning parameter satisfying the inequality (11), we may have that at least $T_0 \subset \hat{T}_n^i$, which means a conservative variable selection. We thus propose an intuitive method to choose c^i by the following recursive algorithm as well as Fujimori (2019):

Step 1. Let $c^{[1]} > 0$ be a fixed constant.
Step 2. Calculate the Dantzig selector for $j \geq 1$ by using $c^{[j]}$;

$$\hat{\theta}^{[j]} := \arg\min_{\beta \in \mathcal{B}_n^{[j]}} \|\beta\|_1,\ \mathcal{B}_n^{[j]} := \{\beta \in \mathbb{R}^{p_n} : \|U_n(\beta)\|_\infty \leq c^{[j]} \tilde{\gamma}_n\}.$$

Step 3. Put

$$c^{[j+1]} = \frac{\|U_n(\hat{\beta}^{[j]})\|_\infty}{\tilde{\gamma}_n}, \quad j \geq 1.$$

Step 4. Repeat Steps 2 and 3 until we have that

$$|c^{[j+1]} - c^{[j]}| \leq \epsilon,$$

where $\epsilon > 0$ is an arbitrary small constant.
The prefixed constant $c^{[1]}$ has to be chosen large enough to ensure that

$$\frac{\|U_n(\beta_0)\|_\infty}{\tilde{\gamma}_n} \leq c^{[1]},$$

For each $j \geq 1$, we may observe that $c^{[j]}$ is close to a random variable C, where

$$C := \frac{\|U_n(\beta_0)\|_\infty}{\tilde{\gamma}_n},$$

with probability tending to 1 as $n \to \infty$ since it holds that $\|\hat{\beta}^{[j]} - \beta_0\|_1 \to^p 0$. In addition, for each sufficiently large $n \in \mathbb{N}$, we can also verify that the sequence $\{c^{[j]}\}_{j\in\mathbb{N}}$ is non-increasing for j and bounded below by 0. Therefore, there exists a limit $c_0 \geq 0$ of $\{c^{[j]}\}_{j\in\mathbb{N}}$ which is close to C with probability tending to 1. Note that if the random variable C is close to 0, then it may hold that $c^{[j]} \to 0$ as $j \to \infty$, which means that the Dantzig selector is nearly or exactly equals to the classical Z-estimator, which is a solution to the following estimating equation:

$$U_n(\beta) = 0.$$

Remark 16 This method may work well when the sample size n is sufficiently large so that the rate $\tilde{\gamma}_n$ is sufficiently small. However, for a finite sample scenario, the rate $\tilde{\gamma}_n$ may be still large. To deal with such a case, we first put some initial value $\beta^{[1]} = (\beta_1^{[1]}, \ldots, \beta_p^{[1]}) \in \mathbb{R}^p$. For example, we put $\beta_j^{[1]} \sim U([a, b])$, $j = 1, \ldots, p$ are mutually independent for some $a, b \in \mathbb{R}$. Then, we can define the initial value of γ_n by

$$\gamma_n^{[1]} = \|U_n(\beta^{[1]})\|_\infty.$$

By using $\gamma_n^{[1]}$, we can calculate the Dantzig selector $\hat{\beta}_n^{[1]}$ and we can update the tuning parameter by

$$\gamma_n^{[2]} = \|U_n(\hat{\beta}_n^{[1]})\|_\infty.$$

Repeating these steps until it converges, we can determine the tuning parameter.

For the Dantzig selector for the linear regression models,

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, I_n),$$

where I_n is the $n \times n$ identity matrix, Candés and Tao (2007) suggested that the tuning parameter can be chosen via the Monte Carlo simulations, i.e., it is determined by the sample mean of $\|X^\top Z\|_\infty$ over the several realizations of $Z \sim N(0, I_n)$. This is similar to our method because for the true value β_0, the residual $Y - X\beta_0 = \epsilon \sim N(0, I_n)$ and $X^\top(Y - X\beta)$ is corresponding to the score function for the linear model. This choice by the Monte Carlo simulation may be also applicable to the proportional hazards model after the linear approximation as introduced in Sect. 6.1, when the approximation works well. Note that we can also apply the cross-validation to determine the tuning parameter. See, e.g., Antoniadis et al. (2010) for the detail.

6 Numerical studies

6.1 The algorithm for the Dantzig selector

In this subsection, we briefly introduce the algorithm to compute the Dantzig selector for Cox's proportional hazards model, which is given by Antoniadis et al. (2010).

For every $k = 0, 1, \ldots$, we calculate the gradient vector $U_n(\hat{\beta}^{[k]})$ and the Hessian matrix $J_n(\hat{\beta}^{[k]})$, where k means the k-th iteration. Then, calculate the square root $X_{(k)}$ of $J_n(\hat{\beta}^{[k]})$ as follows:

$$J_n(\hat{\beta}^{[k]}) = X_{(k)} X_{(k)}.$$

Using the generalized inverse $X_{(k)}^-$ of $X_{(k)}$, we define

$$Y = X_{(k)}^-(J_n(\hat{\beta}^{[k]})\hat{\beta}^{[k]} - U_n(\hat{\beta}^{[k]})).$$

We compute $X_{(k)}^*$ by normalizing $X_{(k)}$ such that each column has l_2 norm one and define the modified version Y^* of Y. Then, we apply the algorithm to compute the Dantzig selector for linear regression model provided by Candés and Tao (2007) for the response vector Y^*, the design matrix $X_{(k)}^*$ and the tuning parameter γ. Note that the linear approximation and the normalizing step requires the rescaling for the tuning parameter γ given by Sect. 5. Note moreover that although this algorithm works well numerically, there is no theoretical proof that the obtained estimator converges to the solution $\hat{\hat{\beta}}_n$ defined by (1).

6.2 Simulation studies

In this subsection, we verify the finite sample performance of the Dantzig selector. We omit the asymptotic normalities of the estimators obtained after variable selection since these are the consequences of the variable selection consistency of the Dantzig selector and the asymptotic normalities of the maximum partial likelihood estimator (MPLE) and the Breslow estimator. We consider the following deigns for the simulation studies. For all cases, the covariates $Z_1, \ldots, Z_n$ are *i.i.d.* uniform random vectors on $[-2, 2]$ whose components are mutually independent, survival time T_i's are *i.i.d.* exponentially distributed and censoring time C_i's are also *i.i.d.* exponentially distributed independently of T_i's. The data are generated to have about 5% censoring. The tuning parameter γ_n is determined by the algorithm in Sect. 5 with initial value

$$\gamma_n^{[1]} = \|U_n(\beta^{[1])}\|_\infty,$$

where each component of $\beta^{[1]}$ is independently generated from the uniform distribution on $[-1, 1]$. We put the true value as follows:

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^p.$$

Table 1 Variable selection results for Case 1

	$(n,p)=(30,50)$		$(n,p)=(50,100)$		$(n,p)=(80,300)$	
	DS (%)	Lasso (%)	DS (%)	Lasso (%)	DS (%)	Lasso (%)
$F-$	93.4	44.0	78.4	0.2	60.4	0
$F+$	53.4	76.0	51.6	99.8	45.5	100

Table 2 Variable selection results for Case 2

	$(n,p)=(50,50)$		$(n,p)=(100,100)$		$(n,p)=(150,300)$	
	DS (%)	Lasso (%)	DS (%)	Lasso (%)	DS (%)	Lasso (%)
$F-$	60.2	0	31.6	0	17	0
$F+$	45.2	78.6	22.8	100	8.8	100

Let us consider the following two cases for the sample size n and the dimension p of covariates:

Case 1. $(n,p) = (30,50),\ (50,100),\ (80,300).$

Case 2. $(n,p) = (50,50),\ (100,100),\ (150,300).$

We have that

$$\frac{\log(\log p)}{\log n} \fallingdotseq 0.4 \iff \log p = O(n^{0.4})$$

and

$$\frac{\log(\log p)}{\log n} \fallingdotseq 0.35 \iff \log p = O(n^{0.35})$$

for Cases 1 and 2, respectively. Then, the rate of convergence of the Dantzig selector for Case 2 should be faster than that for Case 1.

We apply the Dantzig selector and the Lasso, which are calculated by the algorithm described in Sect. 6.1 and the R-package "glmnet" (see Friedman et al. 2010 and Simon et al. 2011), respectively, to these data for 500 replications. The tuning parameter for the Lasso is determined by the cross-validation.

We calculate the estimator of the support index set which is proposed in Theorem 7 for the Dantzig selector. On the other hand, we use the estimator for the Lasso given by

$$\hat{T}_n^L = \{j : \hat{\beta}_j^L \neq 0\},$$

where $\hat{\beta}^L$ is the Lasso estimator. To evaluate the performance of the variable selection, we calculate $F-$, which is the proportion of $T_0 \not\subset \hat{T}$ and $F+$, which is the proportion of $\hat{T} \not\subset T_0$ for $\hat{T} = \hat{T}_n$ and $\hat{T} = \hat{T}_n^L$, respectively, via 500 replications. If the estimators select the variables correctly, both $F-$ and $F+$ are close to zero. Tables 1

Table 3 Mean absolute errors for Case 1

	$(n,p) = (30,50)$			$(n,p) = (50,100)$			$(n,p) = (80,300)$		
	DS	Lasso	$\hat{\beta}_n^{(2)}$	DS	Lasso	$\hat{\beta}_n^{(2)}$	DS	Lasso	$\hat{\beta}_n^{(2)}$
MAE	0.173	0.182	0.166	0.079	0.063	0.069	0.024	0.019	0.018

Table 4 Mean absolute errors for Case 2

	$(n,p) = (50,50)$			$(n,p) = (100,100)$			$(n,p) = (150,300)$		
	DS	Lasso	$\hat{\beta}_n^{(2)}$	DS	Lasso	$\hat{\beta}_n^{(2)}$	DS	Lasso	$\hat{\beta}_n^{(2)}$
MAE	0.138	0.145	0.111	0.039	0.032	0.024	0.023	0.016	0.0097

and 2 show $F-$ and $F+$ of the Dantzig selector (DS) and the Lasso for each case. We can observe that both $F-$ and $F+$ for the Dantzig selector tend to be small as n tends to be large and the rate of the decay for Case 2 is faster than Case 1. The Dantzig selector is better than the Lasso in the sense of $F+$. On the other hand, in terms of $F-$, the Lasso is better than the Dantzig selector. Since the estimator $\hat{T}_n$ of the support index set by the Dantzig selector removes the coefficients such that the absolute values are smaller than the threshold level, $\hat{T}_n$ tends to be smaller than $\hat{T}_n^L$, which implies this selection result. In summary, we can observe that the Dantzig selector enables us to construct a sparser estimator.

Tables 3 and 4 show the average of mean absolute errors (MAEs) of the Dantzig selector, the Lasso estimator and the second estimator $\hat{\beta}_n^{(2)}$ defined by Eq. (8), via 500 replications, respectively. In the sense of MAE, the Dantzig selector and the Lasso perform similarly. On the other hand, the second estimator $\hat{\beta}_n^{(2)}$ seems to be better when n is large since the Dantzig selector reduces more variables than the Lasso.

6.3 Real data analysis

In this subsection, we apply the variable selection method via the Dantzig selector for the gene expression data. CuratedOvarianData from Bioconductor in R (Ganzfried et al., 2013) provides gene expression data for curated ovarian cancer. We use one of their studies "GSE13876eset" which provides survival information (T_i, C_i), $i = 1, \ldots, n$ and normalized gene expression data Z_i, $i = 1, \ldots, n$. We use the gene expression data as the covariates and apply the Dantzig selector and the Lasso for the proportional hazards model. The sample size is 157, and the dimension of covariates is 16788. To apply the Dantzig selector and the Lasso, the dimension is too large. Therefore, some screening method is required. To overcome this issue, we apply the univariate screening as follows. For each $j = 1, \ldots, p$, we fit the univariate proportional hazards model by using (T_i, C_i, Z_i^j), $i = 1, \ldots, n$. Then, we consider the test whose null hypothesis is that the regression parameter β^j is zero and evaluate the p value. Taking the variables

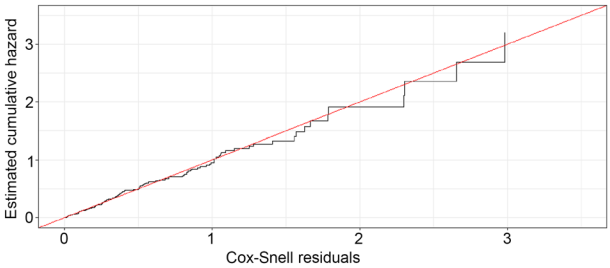

Fig. 1 Plot of Cox–Snell residual

whose p-values are less than 0.05, we obtain 826-dimensional covariates. Applying the Dantzig selector and the Lasso for these 826 covariates and survival information, we can select 3 and 20 variables, respectively. Therefore, we observe that the Dantzig selector and our selection method enable us to construct model with fewer variables for this data. Moreover, for the proportional hazards model constructed by the selected variables via the Dantzig selector, the classical methods such as maximum partial likelihood estimation and the Breslow estimator can be applied, which enables us to ordinal residual analysis such as fit test by the Schoenfeld residual and the Cox–Snell plot. Figure 1 shows the plot of Cox–Snell residual and the estimated cumulative baseline hazard function for the model constructed via the Dantzig selector. We can see that the proportional hazards model is well fitted to the data. Moreover, we apply the goodness of fit test by using Schoenfeld residual whose null hypothesis is that the data follow the proportional hazards model. Then, the p-value is calculated as 0.57, which implies that the null hypothesis cannot be rejected. Therefore, we can conclude that our method enables the constructive model selection for this data.

7 Concluding remarks

In summary, we have been able to construct the asymptotically normal estimator for the proportional hazards model in high-dimensional settings if the sparsity of the regression parameter is fixed. This results are based on the selection result Theorem 7 which is obtained from the l_∞ consistency.

It is well known that the Lasso and the Dantzig selector exhibit similar behaviors for linear regression models. We can see the same phenomena in the proportional hazards model in the sense of l_q consistency for every $q \in [1, \infty]$ since the error bounds for the Dantzig selector in Antoniadis et al. (2010), Fujimori and Nishiyama (2017) are similar to those for Lasso in Huang et al. (2013). On the other hand, the differences between two procedures may occur in the sense of the variable selection consistency. According to Fan et al. (2016), the variable selection consistency, in particular, sign consistencies for estimators, is equivalent to the irrepresentable conditions, which are obtained from KKT conditions of the

optimization problems. Since the KKT conditions of the Lasso-type optimization problems are relatively simple, we can prove the sign consistency of the Lasso estimator for the proportional hazards model by using the irrepresentable condition (see, e.g., Yu 2010). However, the KKT conditions of the Dantzig selector become quite complicated. Although it is possible to derive the sign consistency of the Dantzig selector from the irrepresentable condition for a linear model, it may be difficult to construct the selection results of the Dantzig selector for nonlinear models such as the proportional hazards model by the similar way to that for the Lasso. In contrast, we have proved that l_∞ consistency implies the variable selection consistency in this paper. This type of theoretical results for various regression models may be proved for the Lasso-type estimators because the consistency results are nearly equivalent to that for the Dantzig selector.

8 Proofs of main theorems

Proof of Theorem 6 To prove (4), it suffices to show that

$$\|\hat{\beta}_n - \beta_0\|_1 \leq \frac{K_3 q_n \gamma_n}{\kappa(T_0^n; J_n(\beta_0))}$$

for some positive constant K_3 under the condition that

$$\|U_n(\beta_0)\|_\infty \leq \gamma_n,$$

which is satisfied with probability tending to one by Lemma 5. Put $\tilde{h} = \hat{\beta}_n - \beta_0$ and $h = \tilde{h}/\|\tilde{h}\|_1$, which is proved that $h \in C_{T_0^n}$. Noticing that $-C_n(\beta_0 + xh)$ is a convex function with respect to $x \geq 0$ and it follows from the definition of the estimator that $\|U_n(\hat{\beta}_n)\|_\infty \leq \gamma_n$, we have that

$$\begin{aligned} h^\top [U_n(\beta_0) - U_n(\beta_0 + xh)] \leq & h^\top [U_n(\beta_0) - U_n(\hat{\beta}_n)] \\ \leq & 2\gamma_n \|h\|_1 = 2\gamma_n, \quad x \in [0, \|\tilde{h}\|_1]. \end{aligned} \tag{12}$$

Using the inequality from Huang et al. (2013), we have that

$$xh^\top [U_n(\beta_0) - U_n(\beta_0 + xh)] \geq x^2 \exp(-K^* x) h^\top J_n(\beta_0) h \tag{13}$$

for every x satisfying the inequality (12) and K^* defined by

$$K^* := \sup_{s \in [0,1]} \sup_{1 \leq i, i' \leq n} \sup_{1 \leq j \leq p_n} |Z_i^j(s) - Z_i'^j(s)|,$$

which is bounded almost surely under Assumption 1. This inequality, the fact that $h \in C_{T_0^n}$ and the definition of the factor $\kappa(T_0^n; J_n(\beta_0))$ imply that

$$\begin{aligned} xe^{-K^*x}\kappa^2(T_0^n;J_n(\beta_0))\|h_{T_0^n}\|_1^2 &\le xe^{-K^*x}h^\top J_n(\beta_0)h \\ &\le 2\gamma_n. \end{aligned}$$

Under Assumption 3, it holds for sufficiently large n, every x satisfying (12) that

$$\begin{aligned} xe^{-K^*x} &\le \frac{2q_n\gamma_n}{\kappa^2(T_0^n;J_n(\beta_0))\|h_{T_0^n}\|_1^2} \\ &\le \frac{4q_n\gamma_n}{\kappa^2(T_0^n;J_n(\beta_0))} =: \tau_n \end{aligned}$$

where the last inequality is implied from the fact that

$$\|h_{T_0^n}\|_1 \le \|h\|_1 \le 2\|h_{T_0^n}\|_1, \quad h \in C_{T_0^n}$$

and that $\|h\|_1 = 1$. Note that the set of all x satisfying (12) is closed interval $[0, \tilde{x}]$, where $\tilde{x}$ is some positive constant. We thus have that

$$K^*\tilde{x} \le K^*\tau_n e^{K^*x} = \eta_n,$$

where η_n is a solution to the smaller equation $\eta_n e^{-\eta_n} = K^*\tau_n$. We observe that when τ_n is smaller than $1/e$, the smaller solution of the equation $\eta_n e^{-\eta_n} = \tau_n$ is less than 1. Noticing that $\tau_n \to^p 0$ as $n \to \infty$, we see that $\eta_n < 1$ with probability tending to one as $n \to \infty$. Therefore, we obtain that

$$\|\tilde{h}\|_1 \le \tilde{x} \le \frac{K_3 q_n \gamma_n}{\kappa^2(T_0^n;J_n(\beta_0))}$$

for sufficiently large n and some positive constant K_3. To prove (5), it follows from the definition of $F_\infty(T_0^n;J_n(\beta_0))$ that

$$\begin{aligned} xe^{-K^*x} &\le \frac{2\gamma_n}{F_\infty(T_0^n;J_n(\beta_0))\|h_{T_0^n}\|_1\|h\|_\infty} \\ &\le \frac{4\gamma_n}{F_\infty(T_0^n;J_n(\beta_0))\|h\|_1\|h\|_\infty} \\ &\le \frac{4\gamma_n}{F_\infty(T_0;J_n(\beta_0))\|h\|_\infty}. \end{aligned}$$

Therefore, we obtain that for $x = \|\tilde{h}\|_1$

$$\begin{aligned} \|\tilde{h}\|_\infty = x\|h\|_\infty &\le \frac{4e^{\eta_n}\gamma_n}{F_\infty(T_0;J_n(\beta_0))\|h\|_\infty} \\ &\le \frac{K_4\gamma_n}{F_\infty(T_0;J_n(\beta_0))\|h\|_\infty}. \end{aligned}$$

for sufficiently large n and some positive constant K_4, which ends the proof. □

Proof of Theorem 7 We have that

$$\lim_{n\to\infty} P\big(\|\hat{\beta}_n - \beta_0\|_\infty > \gamma_n\big) = 0$$

by the l_∞ bound from Theorem 6. Therefore, it is sufficient to show that the next inequality

$$\|\hat{\beta}_n - \beta_0\|_\infty \le \gamma_n$$

implies that

$$\hat{T}_n = T_0^n.$$

For every $j \in T_0^n$, it follows from the triangle inequality that

$$|\beta_0^j| - |\hat{\beta}_n^j| \le |\hat{\beta}_n^j - \beta_0^j| \le \|\hat{\beta}_n - \beta_0\|_\infty \le \gamma_n.$$

Noticing that $\inf_{j\in T_0^n} |\beta_0^j| n^\zeta$ is bounded from below by some positive constant c in our assumption, we find that

$$|\hat{\beta}_n^j| \ge |\beta_0^j| - \gamma_n > \frac{c}{n^\zeta} - \gamma_n,$$

for sufficiently large n where $c > 0$ is a constant. Noticing that the right-hand side of this inequality is larger than γ_n for sufficiently large n, we obtain that $T_0^n \subset \hat{T}_n$. On the other hand, for every $j \in (T_0^n)^c$, we have that

$$|\hat{\beta}_n^j - \beta_0^j| = |\hat{\beta}_n^j| \le \gamma_n$$

since it holds that $\beta_0^j = 0$. From this fact, we can see that $j \in \hat{T}_n^c$ which implies that $\hat{T}_n \subset T_0^n$. We thus obtain the conclusion. □

Proof of Theorem 12 We have that

$$\|\hat{\beta}_n^{(2)} - \beta_0\|_1 = \|\hat{\beta}_{nT_0}^{(2)} - \beta_{0T_0}\|_1 + \|\hat{\beta}_{nT_0^c}^{(2)}\|_1.$$

It follows from Lemma 3.1 of Andersen and Gill (1982) that the first term of right-hand side is $o_p(1)$ since the sparsity S is assumed to be fixed. Moreover, we have that

$$\|\hat{\beta}_{nT_0^c}^{(2)}\|_1 1_{\{\hat{T}_n = T_0\}} = 0$$

by the definition of $\hat{\beta}_n^{(2)}$. Noting that $1_{\{\hat{T}_n=T_0\}} \to^p 1$, we obtain the conclusion by using Slutsky's theorem. □

Since Theorems 13 and 15 can be proved by similar ways to the corresponding results of Andersen and Gill (1982), the proofs of them are described in Appendix.

Appendix

Proof of Lemma 11 Under the condition that $\hat{\beta}^{(2)}_{n\hat{T}_n^c} = 0$, we use Taylor expansion to deduce that

$$J_{n\hat{T}_n,\hat{T}_n}(\beta_0)\Big(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0\hat{T}_n}\Big)1_{\{\hat{T}_n=T_0\}} = U_n(\beta_0)_{\hat{T}_n}1_{\{\hat{T}_n=T_0\}} + o_p(\|\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0\hat{T}_n}\|_2).$$

Therefore, under Assumption 9, it follows from Lemma 10 that

$$\mathcal{I}\Big(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0\hat{T}_n}\Big)1_{\{\hat{T}_n=T_0\}} = U_n(\beta_0)_{\hat{T}_n}1_{\{\hat{T}_n=T_0\}} + o_p(\|\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0\hat{T}_n}\|_2) + o_p(1).$$

Since $\mathcal{I}$ is assumed to be non-singular and $P(\hat{T}_n = T_0) \to 1$ as $n \to \infty$ by Theorem 7, we obtain the conclusion. □

Proof of Theorem 13 It follows from the Taylor expansion that

$$\Big\{U_{n\hat{T}_n}(\hat{\beta}^{(2)}_{n\hat{T}_n}) - U_{nT_0}(\beta_{0T_0})\Big\}1_{\{\hat{T}_n=T_0\}} = -J_{nT_0,T_0}(\beta^*_{nT_0})(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0})1_{\{\hat{T}_n=T_0\}},$$

where β^*_n is the point between $\hat{\beta}^{(2)}_n$ and β_0. Then, the assertion is obtained by using Slutsky's theorem and the corresponding result from Andersen and Gill (1982). □

Proof of Theorem 15 We have that

$$\begin{aligned}&\sqrt{n}\{\hat{\Lambda}(t) - \Lambda_0(t)\}1_{\{\hat{T}_n=T_0\}}\\ &\quad= \Big[H_{nT_0}(\beta^*_{nT_0}, t)^\top\sqrt{n}(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0}) + \sqrt{n}W_n(t)\Big]1_{\{\hat{T}_n=T_0\}} + o_p(1).\end{aligned}$$

We can use the fact (10) to deduce that

$$\begin{aligned}&\sqrt{n}\{\hat{\Lambda}(t) - \Lambda_0(t)\}1_{\{\hat{T}_n=T_0\}}\\ &\qquad+ \sqrt{n}\int_0^t(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0})^\top\frac{s^{(1)}}{s^{(0)}}(\beta_{0T_0}, s)\lambda_0(s)\mathrm{d}s1_{\{\hat{T}_n=T_0\}}\\ &\quad= \sqrt{n}W_n(t)1_{\{\hat{T}_n=T_0\}} + o_p(1),\end{aligned}$$

where

$$W_n(t) = \sqrt{n}\int_0^t\frac{d\bar{M}(s)}{S^{(0)}_n(\beta_0, s)}, \quad t \in [0, 1].$$

Then, the conclusion is obtained by using Slutsky's theorem and the corresponding result from Andersen and Gill (1982). □

Acknowledgements The author is grateful to the associate editor and two reviewers for their instructive comments to improve this paper. The author thanks Prof. Y. Nishiyama of Waseda University and Dr. K. Tsukuda of Kyushu University for helpful discussion.

References

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statististics, 10*(4), 1100–1120.

Antoniadis, A., Fryzlewicz, P., & Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics, 37*(4), 531–552.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Annals of Statististics, 37*(4), 1705–1732.

Bradic, J., Fan, J., & Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statististics, 39*(6), 3092–3120.

Candés, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statististics, 35*(6), 2313–2351.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B, 34,* 187–220.

Fan, Y., Gai, Y., & Zhu, L. (2016). Asymtotics of Dantzig selector for a general single-index model. *Journal of Systems Science and Complexity, 29*(4), 1123–1144.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1–22.

Fujimori, K. (2019). The Dantzig selector for a linear model of diffusion processes. *Statistical Inference for Stochastic Processes, 22*(3), 475–498.

Fujimori, K., & Nishiyama, Y. (2017). The l_q consistency of the Dantzig selector for Cox's proportional hazards model. *Journal of Statistical Planning and Inference, 181,* 62–70.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., & Huttenhower, C. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database, 2013,* bat013.

Honda, T., & Härdle, W. K. (2013). Variable selection in Cox regression model with varying coefficients. *Journal of Statistical Planning and Inference, 148,* 67–81.

Huang, J., Sun, T., Ying, Z., Yu, Y., & Zhang, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Annals of Statistics, 41*(3), 1142–1165.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software, 39*(5), 1–13.

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

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Annals of Statistics, 23*(5), 1779–1801.

van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics, 3,* 1360–1392.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer.

Yu, Y. (2010). High-dimensional variable selection in Cox model with generalized Lasso-type convex penalty. *Preprint*.