The Dantzig selector for statistical models of stochastic processes in high-dimensional and sparse settings

高次元・スパースな設定における確率過程の統計モデルに対するDantzig selector

February, 2019

Kou FUJIMORI

The Dantzig selector for statistical models of stochastic processes in high-dimensional and sparse settings

高次元・スパースな設定における確率過程の統計モデルに対するDantzig selector

February, 2019

Waseda University
Graduate School of Fundamental Science and Engineering
Department of Pure and Applied Mathematics,
Research on Stochastic Processes and Statistical Inference

Kou FUJIMORI

Acknowledgements

First of all, I would like to express the deepest appreciation to my supervisor, Professor Yoichi Nishiyama of Waseda University, for his wholehearted research guidance. His lecture, when I was a bachelor course student, gave me the great encounter with the theory of statistical inference for stochastic processes. Thanks to his wonderful lecture, I was able to make a decision to go toward this research field. Moreover, I cannot thank him enough for the long hours discussions since I was a master course student. He has given me a lot of constructive, new and unique ideas, and shown what a researcher should be. I have learned many things from him including researcher's spirits.

I am also deeply grateful to Professor Masanobu Taniguchi and Professor Yasutaka Shimizu of Waseda University for their comments on the earlier version of this thesis. Prof. Taniguchi has invited me many seminars and international symposiums and given me many nice opinions and kind comments. Prof. Shimizu has given me the chance to study a lot of new research fields through the seminars with him and his students. These opportunities made me grow up not only as a researcher, but also as a human.

Professor Takeru Suzuki supervised me when I was a bachelor course student. He led me to the world of mathematical statistics. There is no doubt that his leading is one of my triggers to start my research.

I would like to express my thanks to my senior apprentice, Dr. K. Tsukuda of the University of Tokyo, who took part in our seminar as an advisor. He always listens to my academic problems and any other sufferings. Without his advices, I might not have continued my research life until now.

I want to express my gratitude to Dr. F. Akashi of Waseda University and Dr. Y. Liu of Kyoto University. They taught me a lot of academic topics and gave me many comments on my researches. Their advices are also indispensable for my research activities.

I thank all colleagues including Mr. H. Nagahata, Ms. Y. Xue and Mr. Y. Tanida and all students in the laboratories of Prof. Taniguchi and Prof. Shimizu. I have studied many things from the seminars with them, and they are so precious for my mental support.

Last but not least, I thank my family and friends. Without their help, I could not have completed this thesis or anything.

Abstract

The Dantzig selector, which was proposed by Candés and Tao in 2007, is an estimation procedure for regression models in high-dimensional and sparse settings. The Dantzig selectors for some statistical models of stochastic processes are studied in this thesis. We apply this procedure to Cox's proportional hazards model and some specific models of diffusion processes and prove the l_q consistencies for every $q \in [1, \infty]$ and the variable selection consistencies of the estimators. Based on partial likelihood and quasi-likelihood methods which were studied intensively in low-dimensional settings, we study these statistical models of stochastic processes in high-dimensional and sparse settings, which need some mathematically challenging tasks to prove the asymptotic properties of the estimators. The consistencies in the sense of the l_q norm for every $q \in [1, \infty]$ of the estimators are derived from the stochastic maximal inequalities to deal the curse of dimension and some matrix factors and conditions on Hessian matrices of likelihood functions to deal with the sparsities. We use Bernstein's inequalities for martingales and the maximal inequalities using Orlicz norm for the former problem and matrix conditions using restricted eigenvalue, compatibility factor and weak cone invertibility factor for the latter problem, which are known to be weaker conditions than others. We prove that l_q consistency of the estimator implies the variable selection consistency which enables us to reduce the dimension. Using the dimension reduction, asymptotically normal estimators can be constructed.

Contents

Chapter 1

Introduction

Recently, data with much more measurements than number of observations are paid more and more attentions in statistical applications. We call this kind of data "high-dimensional" data. We need much more parameters than sample size in order to construct models for these high-dimensional data. To estimate high-dimensional parameters, classical estimation procedures may not work well because the parameter spaces are too large and rich. Therefore, many researchers have been proposed new estimation procedures. Most of them are constructed by adding some penalty functions of parameters, which induce the sparsity of parameters, *i.e.*, most of components in parameters are zero, to score functions.

For instance, let us consider the following linear regression model:

$$Y = Z\beta + \epsilon,$$

where $Y \in \mathbb{R}^n$ is a response vector, Z is a $n \times p$ design matrix, $\beta \in \mathbb{R}^p$ is an unknown parameter, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ is an error vector such that $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, are independent and identically distributed. We consider the estimation problem of β in a high-dimensional and sparse setting, *i.e.*, $p \gg n$ and the number S of nonzero components of the true value β_0 is smaller than n. One of the most famous estimation procedures for this problem is l_1 penalized method called Lasso (Least absolute shrinkage and selection operator) which was proposed by Tibshirani (1996):

$$\hat{\beta}_L := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - Z\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

where λ is a tuning parameter. Lasso estimator $\hat{\beta}_L$ is a least square estimator penalized by l_1 norm of parameter. It is well known that Lasso estimator has good asymptotic properties such as consistency under some regularity conditions.

On the other hand, Candés and Tao (2007) proposed a relatively new method called the Dantzig selector which is defined as follows:

$$\hat{\beta}_D := \arg\min_{\beta \in \mathcal{C}} \|\beta\|_1, \quad \mathcal{C} := \left\{ \beta \in \mathbb{R}^p : \sup_{1 \le j \le p} |Z^{j\top}(Y - Z\beta)| \le \lambda \right\},$$

where $\lambda \ge 0$ is a tuning parameter. When $\lambda = 0$, the Dantzig selector returns to the LSE (least square estimator) in general settings or MLE (maximum likelihood estimator) in Gaussian settings. For $\lambda > 0$, the Dantzig selector searches for the sparsest β within the given distance of the classical estimators such as LSE or MLE. Notice that this method has a good potential to be applied to other models. We can see that the Dantzig selectors have also good asymptotic properties. In particular, Bickel et al. (2009) proved that Lasso estimator and the Dantzig selector exhibit similar behaviors for the linear regression model and the nonparametric regression model. Based on their results, we prove the consistency of the Dantzig selector for a linear regression model in Chapter 2 of this thesis. The Dantzig selector's advantages are not only consistency but also the variable selection consistency. We can construct an consistent estimator for the support index set of the true value which enables us to reduce the dimension and define an asymptotically normal estimator after the selection as in Chapter 2 of this thesis. In addition, the Dantzig selector has a computational advantage because it can be solved by a linear programming. We can easily verify the finite sample performance of the Dantzig selector for linear model numerically as presented in Chapter 6.

Lasso and the Dantzig selector have been studied for various models including models of stochastic processes. Especially, there are many existing literatures dealing with Cox's proportional hazards model. 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. In a high-dimensional and sparse setting, Huang et al. (2013) proved l_q consistency for every $q \in [1, \infty)$ of Lasso estimator, which is proposed by Tibshirani (1997), under some appropriate conditions. On the other hand, Antoniadis et al. (2010) proposed the Survival Dantzig selector which is an application of the Dantzig selector for the proportional hazards model and proved the l_2 consistency of the estimator. In Chapter 3 of this thesis, we will prove the l_q consistency of the Dantzig selector for the proportional hazards model for all $q \in [1, \infty]$ under some conditions which are similar to those of Huang et al. (2013). Moreover, based on the l_q consistency result, we prove the variable selection

consistency of the Dantzig selector. The selection result enables us to reduce the dimension and construct asymptotically normal estimators for regression parameter and cumulative baseline hazards function.

We can also consider the estimation problems in a high-dimensional and sparse setting for some models of diffusion processes. In Chapter 4, we will consider the one-dimensional stochastic process which is a solution to the stochastic differential equation given by

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \exp(\theta^\top Z_s)dW_s, \tag{1.1}$$

where $\{W_t\}_{t\geq 0}$ is a standard Brownian motion, $b(\cdot)$ is a drift function, which is treated as a nuisance parameter, $\{Z_t\}_{t\geq 0} = \{(Z_t^1, Z_t^2, \ldots, Z_t^p)\}_{t\geq 0}$ is a uniformly bounded p-dimensional continuous process, which is regarded as a covariate vector, and θ is an unknown parameter of interest. We observe the process $\{X_t\}_{t\geq 0}$ at $n+1$ equidistant time points $0 =: t_0^n < t_1^n < \ldots < t_n^n$, where $t_k^n = kt_n^n/n$ for $k = 0, 1, \ldots, n$. Assume that $p = p_n \gg n$ and the number of non-zero components S in the true value θ_0 is relatively small. In this high-dimensional and sparse setting, we will consider the estimation problem of θ_0. The covariate processes $\{Z_t^i\}_{t\geq 0}$, $i = 1, 2, \ldots, p_n$, are, for example, some functionals $\{\phi_i(X_t^i)\}_{t\geq 0}$ of solutions to other stochastic differential equations $\{X_t^i\}_{t\geq 0}$, where ϕ_i's are uniformly bounded smooth functions or random variables which do not depend on t. Using these discretely observed data, we will apply the Dantzig selector in order to estimate θ and prove the l_q consistency of the estimator for all $q \in [1, \infty]$ in Chapter 4 of this thesis. Our estimation procedure is based on the quasi-likelihood method for discretely observed data which has been studied intensively in low-dimensional cases. The study on this subject started at early 90s by the works of, for instance, Yoshida (1992), Genon-Catalot and Jacod (1993) and Kessler (1997) among others. For recent developments on this subject, see Yoshida (2011) and Uchida and Yoshida (2012). Estimation problems for models of stochastic processes in high or fixed dimensional and sparse settings have been studied by many authors. De Gregorio and Iacus (2012), Masuda and Shimizu (2017) dealt with some penalized estimators in discretely observed multi-dimensional models of diffusion processes under fixed dimensional settings. In high-dimensional setting, there are some researches on the estimation problems for diffusion coefficients or volatility matrices of models of diffusion processes. For example, Wang and Zou (2010) proposed the estimator for the sparse volatility matrix in high-dimensional settings by the thresholding, and derived the rate of convergence of the estimator in the sense of L_β-norm for $\beta \geq 2$.

The statistical inference for high-dimensional linear diffusion processes was especially discussed by some researchers. Periera and Ibrahimi (2014) studied vari-

ous models of multi-dimensional diffusion processes observed continuously in high-dimensional settings including the following p-dimensional linear model:

$$X_t = X_0 + \int_0^t \Theta^\top X_s ds + W_t, \quad t \in [0, T]. \tag{1.2}$$

This model may be useful for various fields such as statistical physics, chemical reactions, finances and network systems. They proposed a Lasso type estimator for the true value Θ^0 of Θ and discussed the support recovery of the estimator when the dimension of the process p and the time interval T tends to ∞ independently. Similarly, Gaiffas and Matulewicz (2017) studied the drift estimation based on the Lasso-type estimators for the high-dimensional Ornstein-Uhlenbeck processes described by the SDE like (1.2). They derived the oracle properties of the estimator for the sparse drift matrix and showed some applications for financial data. However, there are few previous researches dealing with the estimation problems for these linear models based on discrete observations. In Chapter 5, we will consider the process $\{X_t\}_{t\geq 0}$ which is a solution to the following linear stochastic differential equation which is more general than (1.2):

$$X_t = X_0 + \int_0^t \Theta X_s ds + \sigma W_t, \tag{1.3}$$

where $\{X_t\}_{t\geq 0} = \{(X_t^1, \ldots, X_t^p)\}_{t\geq 0}$ is a p-dimensional process, Θ is an unknown $p \times p$ drift matrix, and $\sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ is an unknown $p \times p$ diagonal matrix. We will consider the estimation problems for Θ and σ based on the discrete time observation in a high-dimensional and sparse setting for Θ and prove the consistency and the variable selection consistency of the estimators similar to those in other chapters.

The consistency of Lasso estimator and the Dantzig selector depends on some matrix conditions for the Hessian matrices of the log likelihood functions. Candés and Tao (2007) and Antoniadis et al. (2010) proved the consistency of the Dantzig selector under the condition called UUP condition (Uniform Uncertainty Principle condition). On the other hand, Bickel et al. (2009) showed the consistency of Lasso estimator and the Dantzig selectors for the linear regression model and nonparametric regression model using the factor called restricted eigenvalue and the conditions for this factor. Huang et al. (2013) proved the consistency of Lasso estimator for the proportional hazards model using restricted eigenvalue and related factors called compatibility factor and weak cone invertibility factors. In this paper, we will use restricted eigenvalue type conditions to prove the consistency of the Dantzig selectors for statistical models of stochastic processes because it is known that these type of conditions are weaker than UUP condition. In Chapter 2 of this paper, we will discuss these matrix conditions using linear regression models.

It is well-known that the variable selection consistency is generally almost equivalent to the, so called, Irrepresentable condition which is implied from the condition for existing the unique solution to the optimization problem determining the estimator such as Lasso and the Dantzig selector (See for example, Fan et al. (2016)). For Lasso type problem, Irrepresentable condition is relatively simple. However, for the Dantzig selector, this condition becomes very complicated and hard to present explicitly even for simple linear regression models. Since it is hard to consider this condition for models of stochastic processes, we propose another method to prove the variable selection consistency by using thresholding methods.

This thesis is organized as follows. First, we introduce an overview of the asymptotic theory for the Dantzig selector by using linear regression models in Chapter 2. The techniques to prove the consistency and the variable selection consistency in this chapter are very similar to those in other chapters. In Chapter 3, the Dantzig selector for Cox's proportional hazards model is presented. This part is based on the paper Fujimori and Nishiyama (2017a) and Fujimori (2017). The proportional hazards model includes regression parameters and baseline hazards function as unknown parameters. We propose asymptotically normal estimators for both parameters in a high-dimensional and sparse setting by using the variable selection consistency of the Dantzig selector for the proportional hazards model. Chapters 4 and 5 concern the Dantzig selector for models of diffusion processes. In Chapter 4, we consider the model of diffusion process including a high-dimensional and sparse parameter in diffusion coefficient which is the regression coefficient for the high-dimensional covariate process. We propose the Dantzig selector for the parameter based on the quasi-likelihood method and prove the l_q consistency and the variable selection consistency of the estimator. In addition, we propose an asymptotically normal estimator by dimension reduction under an ergodic assumption on the covariate process. The l_q consistency of the Dantzig selector for the model can be seen in Fujimori and Nishiyama (2017b). Chapter 5, which is based on the paper Fujimori (2018), deals with a linear model of diffusion processes which has an unknown high-dimensional and sparse matrix in the drift coefficient and unknown high-dimensional diagonal matrix in the diffusion coefficient. We estimate the diffusion matrix by maximum quasi likelihood estimator and the drift matrix by the Dantzig selector which are proved to satisfy the consistency. Similar to other chapters, we also discuss the variable selection consistency and the construction of the asymptotically normal estimator for the drift matrix in this chapter. Finally, we provide numerical studies for a linear regression model and Cox's proportional hazards model in Chapter 6. We discuss the way to choose the tuning parameter and verify the l_1 consistency and the variable selection consistency numerically for finite sample.

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

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

where A_i^j 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 $|T| \times |T'|$ sub-matrix $A_{T,T'}$ by

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

For an $\mathbb{R}$-valued random variable X on a probability space $(\Omega, \mathcal{F}, P)$, we define the L_q norm of X by

$$\|X\|_{L_q} := (E[|X|^q])^{\frac{1}{q}},$$

where $E[\cdot]$ denotes the expectation with respect to the probability measure P.

For a nondecreasing, convex $\mathbb{R}$-valued function Φ with $\Phi(0) = 0$ and X a random variable, we introduce Orlicz norm of X with respect to Φ by

$$\|X\|_\Phi := \inf \left\{ C > 0 : E\left[\Phi\left(\frac{|X|}{C} \right) \right] \le 1 \right\}.$$

We will use this norm with respect to the function $\Phi_q(X) = e^{x^q} - 1$ for $q \ge 1$.

Chapter 2

An overview of the Dantzig selector for linear regression models

In this chapter, we will introduce the asymptotic properties of the Dantzig selector for the following p-dimensional linear regression model on a probability space $(\Omega, \mathcal{F}, P)$:

$$Y_i = Z_i^\top \beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where Y_i's are $\mathbb{R}$-valued response variables, $\{Z_i\}$'s are $\mathbb{R}^p$-valued independent random variables whose components are bounded P-almost surely, $\beta \in \mathbb{R}^p$ is an unknown parameter of interest and $\{\epsilon_i\}_{i=1}^n$ is an independent zero mean random sequence. In high-dimensional and sparse setting, that is, $p = p_n \to \infty$ as $n \to \infty$ and the number S of nonzero components in the true value β_0 of the unknown parameter β is smaller than n, we will show the consistency and the variable selection consistency of the Dantzig selector in Sections 2.3 and 2.4. Moreover, we will show that the variable selection consistency enables us to construct an asymptotically normal estimator for β_0 by dimension reduction in Section 2.5. The main concept of asymptotic theory in this chapter can be seen in other chapters which deal with models of stochastic processes.

2.1 Model setups and preliminaries

2.1.1 Model setups

We use the matrix form of the model (2.1) as follows:

$$Y = Z\beta + \epsilon,$$

where $Y = (Y_1, \ldots, Y_n)^\top$ is an $\mathbb{R}^n$-valued response vector, $Z = (Z_1^\top, \ldots, Z_n^\top)^\top$ is a $n \times p$ random design matrix and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is an $\mathbb{R}^n$-valued zero-mean random variable whose components are mutually independent. We write T_0 for the support index set of the true value β_0 and S for the number of indices in T_0 which means a sparsity of the parameter, *i.e.*,

$$T_0 := \{j : \beta_0^j \neq 0\}, \quad S = |T_0|.$$

We assume the following conditions for the sample size n, the dimension $p = p_n$, sparsity S.

Assumption 2.1. *The following (i) and (ii) are satisfied.*

(*i*) *The dimension $p = p_n$ allows to tends to ∞ as $n \to \infty$. Moreover, it holds that*

$$\frac{\log p_n}{n} \to 0, \quad n \to \infty.$$

(*ii*) *S is a fixed constant independent of n.*

Moreover, we assume that the random variable Z and ϵ satisfy the following conditions.

Assumption 2.2. (*i*) *The sequence $\{Z_i\}_{i\in\mathbb{N}}$ is an i.i.d. sequence of random vectors. For every $i = 1, 2, \ldots, n$, Z_i^j, $j = 1, \ldots, p_n$ are mutually independent. Moreover, there exists a positive constant M such that*

$$\sup_{i,j\in\mathbb{N}} |Z_i^j| \leq M, \quad a.s.$$

(*ii*) *The sequence $\{\epsilon_i\}_{i\in\mathbb{N}}$ is an independent sub-Gaussian random sequence with zero mean, i.e., there exist positive constants C and ν such that*

$$P\left(|\epsilon_i| > t\right) \leq C \exp\left(-\nu t^2\right), \quad t > 0, \quad i \in \mathbb{N}.$$

Moreover, $\{\epsilon_i\}_{i\in\mathbb{N}}$ is independent of $\{Z_i\}_{i\in\mathbb{N}}$.

Note that it is well-known that bounded and centered random variables are sub-Gaussian. Under Assumption 2.2, we can deal with the curse of dimension by using maximal inequalities described in the next subsection.

2.1.2 Maximal inequalities for sub-Gaussian random variables

To prove the consistency of high-dimensional estimators, we will use Orlicz norm introduced as follows.

Definition 2.3. *Let Φ be a nondecreasing, convex function with $\Phi(0) = 0$ and X a random variable. Orlicz norm of X with respect to Φ is defined by*

$$\|X\|_\Phi := \inf\left\{C > 0 : E\left[\Phi\left(\frac{|X|}{C}\right)\right] \leq 1\right\}.$$

We will use this norm with respect to the function $\Phi_p(X) = e^{x^p} - 1$ for $p \geq 1$. For example, we can easily show that Φ_2-Orlicz norm is bounded if and only if X is sub-Gaussian. Especially, when X is a standard normal random variable, it holds that

$$\|X\|_{\Phi_2} = \sqrt{\frac{8}{3}},$$

which will be used in the proof of Theorem 5.4. It is well-known that Φ_p-Orlicz norm$\|\cdot\|_{\Phi_p}$ and L_p-norm $\|\cdot\|_p$ satisfies the following inequalities.

$$\|X\|_{\Phi_p} \leq \|X\|_{\Phi_q}(\log 2)^{\frac{1}{q}-\frac{1}{p}}, \quad p \leq q.$$

$$\|X\|_p \leq p!\|X\|_{\Phi_1}.$$

Moreover, we have the following maximal inequality for Orlicz norm.

Lemma 2.4. *Let Φ be a nondecreasing, convex, nonzero function with $\Phi(0) = 0$ and*

$$\limsup_{x,y\to\infty} \frac{\Phi(x)\Phi(y)}{\Phi(cxy)} < \infty$$

for some constant $c > 0$. It holds for any random variables $X_1, \ldots, X_m$ that

$$\left\|\max_{1\leq i\leq m} X_i\right\|_\Phi \leq K\Phi^{-1}(m) \max_{1\leq i\leq m} \|X_i\|_\Phi,$$

where K is a positive constant depending only on the function Φ.

Note that if we put $\Phi = \Phi_p$, then we have that

$$\Phi_p^{-1}(m) = (\log(1+m))^{\frac{1}{p}}.$$

In addition, combining the Bernstein's inequality and Orlicz norm, we have the another type of maximal inequality as follows.

Lemma 2.5. *If random variables $X_1, \ldots, X_m$ satisfy the following tail bound*

$$P\left(|X_i| > x\right) \leq 2\exp\left(-\frac{x^2}{2(b+ax)}\right), \quad i = 1, \ldots, m$$

for all x and fixed $a, b > 0$, then it holds that

$$\left\| \max_{1\leq i\leq m} X_i \right\|_{\Phi_1} \leq K\left\{ a\log(1+m) + \sqrt{b}\sqrt{\log(1+m)} \right\},$$

where K is a universal constant.

See e.g. van der Vaart and Wellner (1996) for the details of Orlicz norm and maximal inequalities.

2.2 The Dantzig selector for linear regression models

Now, we define the estimator for β_0 by the Dantzig selector proposed by Candés and Tao (2007) defined as follows:

$$\hat{\beta}_n := \arg\min_{\beta\in\mathcal{C}_n} \|\beta\|_1, \quad \mathcal{C}_n := \left\{\beta \in \mathbb{R}^p : \|\psi_n(\beta)\|_\infty \leq \lambda_n\right\}, \tag{2.2}$$

where

$$\psi_n(\beta) := \frac{1}{n} Z^\top (Y - Z\beta)$$

and $\lambda_n \geq 0$ is a tuning parameter which satisfies the following condition.

Assumption 2.6. *The tuning parameter λ_n tends to 0 as $n \to \infty$. Moreover, it holds that*

$$\frac{\lambda_n}{\sqrt{n^{-1}\log p_n}} \to \infty, \quad n \to \infty.$$

Some remarks on the choice of the tuning parameter are presented in Chapter 6. In this chapter, we prove the consistency of this estimator in the sense of l_q norm for every $q \in [1, \infty]$ and the variable selection consistency. The proof of l_q consistency of the Dantzig selector for the linear regression model is similar to that in Bickel et al. (2009). It is well-known that the variable selection consistency of the Lasso type and Dantzig selector type estimator is equivalent to "Irrepresentable condition" which is obtained from KKT conditions for the optimization problems (see e.g. Fan et al. (2016)). However, Irrepresentable condition for the Dantzig selector is very complicated compared with that for Lasso type problems even in a linear regression models. We therefore provide another proof of the variable selection consistency by using the thresholding method.

2.3 The l_q-consistency of the Dantzig selector

First, we prove that the true value β_0 belongs to the constrain set $\mathcal{C}_n$ with probability tending to 1.

Lemma 2.7. *Under Assumptions 2.1, 2.2 and 2.6, it holds that*

$$\lim_{n\to\infty} P\left(\|\psi_n(\beta_0)\| > \lambda_n\right) = 0.$$

Proof. Noting that

$$\|\psi_n(\beta_0)\|_\infty = \left\|\frac{1}{n}Z^\top\epsilon\right\|_\infty,$$

we first evaluate the tail probability;

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n Z_i^j\epsilon_i\right| > x\right)$$

for every $x \in \mathbb{R}$ and $j \in \{1,\ldots,p_n\}$. Since Z_i^j's are $[-M,M]$-valued P.-$a.s.$ and ϵ_i's are independent sub-Gaussian random variables, it holds for every $s > 0$ and some positive constant σ that

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n Z_i^j\epsilon_i > x\right) &\leq P\left(\exp\left(\frac{s}{n}\sum_{i=1}^n Z_i^j\epsilon_i\right) > e^{sx}\right) \\
&\leq e^{-sx}\prod_{i=1}^n E\left[\exp\left(\frac{Ms}{n}\epsilon_i\right)\right] \\
&\leq e^{-sx}\prod_{i=1}^n \exp\left(\frac{\sigma^2(Ms)^2}{2n^2}\right) \\
&\leq \exp\left(-\frac{nx^2}{2M\sigma^2}\right).
\end{aligned}$$

We therefore obtain that

$$P\left(\frac{1}{n}\left|\sum_{i=1}^n Z_i^j\epsilon_i\right| > x\right) \leq 2\exp\left(-\frac{nx^2}{2M\sigma^2}\right).$$

From Lemma 2.5, it holds for a universal constant $K > 0$ that

$$\left\|\sup_{1\leq j\leq p_n}\left|\frac{1}{n}\sum_{i=1}^n Z_i^j\epsilon_i\right|\right\|_{\Phi_1} \leq K\sqrt{\frac{M\sigma^2\log(1+p_n)}{n}}.$$

It follows from Markov's inequality that

$$
\begin{aligned}
P(\|\psi_n(\beta_0)\|_\infty \geq \lambda_n) &= P(\sup_{1\leq j\leq p_n} |\psi_n^j(\beta_0)| \geq \lambda_n) \\
&\leq P\left(\Phi_1\left(\frac{\sup_{1\leq j\leq p_n}|\psi_n^j(\beta_0)|}{\|\sup_{1\leq j\leq p_n}|\psi_n^j(\beta_0)|\|_{\Phi_1}}\right) \geq \Phi_1\left(\frac{\lambda_n}{\|\sup_{1\leq j\leq p_n}|\psi_n^j(\beta_0)|\|_{\Phi_1}}\right)\right) \\
&\leq \Phi_1\left(\frac{\lambda_n}{\|\sup_{1\leq j\leq p_n}|\psi_n^j(\beta_0)|\|_{\Phi_1}}\right)^{-1} \\
&\leq \Phi_1\left(\frac{\lambda_n}{\left(K\sqrt{\frac{M\sigma^2\log(1+p_n)}{n}}\right)}\right)^{-1}.
\end{aligned}
$$

The right-hand side of this inequality converges to 0 if the tuning parameter λ_n satisfies that

$$\frac{\lambda_n}{\sqrt{n^{-1}\log p_n}} \to \infty,$$

which is verified under Assumption 2.6. We thus obtain the conclusion. □

Next, we prove that the $p_n \times p_n$ Hessian matrices $J_n := 1/nZ^\top Z$ can be approximated by the deterministic matrix $\mathcal{I}_n := E[Z^\top Z]$.

Lemma 2.8. *Under Assumption 2.2, the random sequence e_n defined by*

$$e_n := \|J_n - \mathcal{I}_n\|_\infty$$

converges to 0 in probability as $n \to \infty$.

Proof. Since $\{Z_i\}_{i=1}^n$ is *i.i.d.* random sequence, (k,l) component of the matrix $J_n - \mathcal{I}_n$ for each $k, l \in \{1, 2, \ldots, p_n\}$ is

$$(J_n - \mathcal{I}_n)_{k.l} = \frac{1}{n}\sum_{i=1}^n Z_i^k Z_i^l - E[Z_1^k Z_1^l].$$

It follows from the weak law of large numbers that

$$\frac{1}{n}\sum_{i=1}^n Z_i^k Z_i^l - E[Z_1^k Z_1^l] \to^p 0, \quad n \to \infty$$

for every $k, l \in \{1, 2, \ldots, p_n\}$. Moreover, as mentioned in earlier, bounded and centered random variables $\{(J_n - \mathcal{I}_n)\}_{k,l}$ are sub-Gaussian. We therefore can show that

$$\|J_n - \mathcal{I}_n\|_\infty \to^p 0$$

by the similar way to the proof of Lemma 2.7. □

The asymptotic properties of classical estimators such as LSE and MLE are follows from non-singularities of Hessian matrices or Fisher information matrices. However, in high-dimensional settings, such conditions cannot be generally verified. To deal with such phenomena, we introduce the following factors for the high-dimensional matrix $\mathcal{I}_n$ which can be seen in, for example, Bickel et al. (2009) and van de Geer and Bühlmann (2009).

Definition 2.9. *For every index set $T \subset \{1, 2, \cdots, p_n\}$ and $h \in \mathbb{R}^{p_n}$, h_T is a $\mathbb{R}^{|T|}$ dimensional sub-vector of h constructed by extracting the components of h corresponding to the indices in T. Define the set C_T by*

$$C_T := \{h \in \mathbb{R}^{p_n} : \|h_{T^c}\|_1 \leq \|h_T\|_1\}.$$

We introduce the following three factors.

(A) **Compatibility factor**

$$\kappa(T_0; \mathcal{I}_n) := \inf_{0 \neq h \in C_{T_0}} \frac{S^{\frac{1}{2}}(h^\top \mathcal{I}_n h)^{\frac{1}{2}}}{\|h_{T_0}\|_1}.$$

(B) **Weak cone invertibility factor**

$$\begin{aligned} F_q(T_0; I_n(\beta_0)) &:= \inf_{0 \neq h \in C_{T_0}} \frac{S^{\frac{1}{q}} h^\top \mathcal{I}_n h}{\|h_{T_0}\|_1 \|h\|_q}, \quad q \in [1, \infty), \\ F_\infty(T_0; \mathcal{I}_n) &:= \inf_{0 \neq h \in C_{T_0}} \frac{(h^\top \mathcal{I}_n h)^{\frac{1}{2}}}{\|h\|_\infty}. \end{aligned}$$

(C) **Restricted eigenvalue**

$$RE(T_0; \mathcal{I}_n) := \inf_{0 \neq h \in C_{T_0}} \frac{(h^\top \mathcal{I}_n h)^{\frac{1}{2}}}{\|h\|_2}.$$

These factors are similar to the minimal eigenvalue of matrix $\mathcal{I}_n$, which cannot expect to be positive. However, since we can show that the restriction set C_{T_0} satisfy that $\hat{\beta}_n - \beta_0 \in C_{T_0}$, it is sufficient to prove the consistency of the estimator that these three factors are positive. We thus assume that the compatibility factor is asymptotically positive instead of the non-singularity of $\mathcal{I}_n$.

Assumption 2.10. *It holds that*

$$\liminf_{n\to\infty} \kappa(T_0; \mathcal{I}_n) > 0.$$

Noting also that $\|h_{T_0}\|_1^q \geq \|h_{T_0}\|_q^q$ for all $q \in [1, \infty)$, and $\|h\|_\infty \leq \|h\|_1$, we can see that $\kappa(T_0; \mathcal{I}_n) \leq 2\sqrt{S}RE(T_0; \mathcal{I}_n)$, $\kappa(T_0; \mathcal{I}_n) \leq F_q(T_0; \mathcal{I}_n)$ and $\kappa(T_0; \mathcal{I}_n) \leq 2\sqrt{S}F_\infty(T_0; \mathcal{I}_n)$. We therefore have that (B) and (C) are also asymptotically positive under Assumption 2.10. There exist other matrix conditions for Dantzig selector such as the Uniform Uncertainty Principle (UUP) condition, which is used in Candés and Tao (2007) and Antoniadis et al. (2010). To discuss the relationship between the UUP condition and our conditions, let us introduce some objects.

Note that there exists a matrix A such that $A^\top A = \mathcal{I}_n$, because $\mathcal{I}_n$ is a non-negative definite matrix. Given a index set $T \subset \{1, 2, \ldots, p_n\}$, we write A_T for the $p_n \times |T|$ matrix constructed by extracting the columns of A corresponding to the indices in T. The restricted isometry constant $\delta_N(\mathcal{I}_n)$ is the smallest quantity such that

$$(1 - \delta_N(\mathcal{I}_n))\|h\|_2^2 \leq \|A_T h\|_2^2 \leq (1 + \delta_N(\mathcal{I}_n))\|h\|_2^2,$$

for all $T \subset \{1, 2, \ldots, p_n\}$ with $|T| \leq N$, where $N \leq p_n$ is an integer, and all $h \in \mathbb{R}^{|T|}$. The restricted orthogonality constant $\theta_{S,S'}(\mathcal{I}_n)$ is the smallest quantity such that

$$|(A_T h)^\top A_{T'} h'| \leq \theta_{S,S'}(\mathcal{I}_n)\|h\|_2\|h'\|_2$$

for all disjoint sets $T,\ T' \subset \{1, 2, \ldots, p_n\}$ with $|T| \leq S,\ |T'| \leq S'$, where $S + S' \leq p_n$ and all vectors $h \in \mathbb{R}^{|T|}$ and $h' \in \mathbb{R}^{|T'|}$. For $\delta_{2S}(\mathcal{I}_n)$ and $\theta_{S,2S}(\mathcal{I}_n)$, the UUP condition is described that $0 < 1 - \delta_{2S}(\mathcal{I}_n) - \theta_{S,2S}(\mathcal{I}_n)$.

In addition, we introduce another factor $\phi_{2S}(T_0; \mathcal{I}_n)$ by

$$\phi_{2S}(T_0; \mathcal{I}_n) := \inf_{T \supset T_0,\ |T| \leq 2S,\ h \in D_{T_0,T}} \frac{(h^\top \mathcal{I}_n h)^{\frac{1}{2}}}{\|h_T\|_2},$$

where

$$D_{T_0,T} := \left\{ h \in C_{T_0} : \|h_{T^c}\|_\infty \leq \min_{j \in T \setminus T_0} |h_j| \right\}, \quad T \supset T_0.$$

Define that $\min_{j \in T \setminus T_0} |h_j| = \infty$ when $T = T_0$. The next lemma provides the asymptotic relationship between the UUP condition and a condition for $\phi_{2S}(T_0; \mathcal{I}_n)$. The proof is an adaptation of that in van de Geer and Bühlmann (2009), so it is omitted.

Lemma 2.11. *The UUP condition that*

$$\liminf_{n\to\infty} \{1 - \delta_{2S}(\mathcal{I}_n) - \theta_{S,2S}(\mathcal{I}_n)\} > 0$$

implies the following condition;

$$\liminf_{n\to\infty} \phi_{2S}(T_0; \mathcal{I}_n) > 0.$$

Noting that $\|h_{T_0}\|_1^2 \leq S\|h_{T_0}\|_2^2$, we have that

$$\kappa(T_0; \mathcal{I}_n) \geq \phi_{2S}(T_0; \mathcal{I}_n),$$

which implies that

$$\liminf_{n\to\infty} \kappa(T_0; \mathcal{I}_n) \geq \liminf_{n\to\infty} \phi_{2S}(T_0; \mathcal{I}_n) > 0.$$

Therefore, Assumption 2.10 is weaker than other conditions described above. Moreover, we can easily observe that Assumption 2.10 is satisfied when, for example, $S \times S$ sub-matrix $\mathcal{I}_{nT_0,T_0}$ is positive definite.

Now, we are ready to prove the l_q consistency of the Dantzig selector.

Theorem 2.12. *Under Assumptions 2.1, 2.2, 2.6 and 2.10, the following (i)-(iv) hold true.*

(*i*) *It holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_2^2 \geq \frac{4\|\beta_0\|_1^2 e_n + 4\|\beta_0\|_1 \lambda_n}{RE^2(T_0; \mathcal{I}_n)}\right) = 0, \tag{2.3}$$

where $e_n = \|J_n - \mathcal{I}_n\|_\infty = o_p(1)$ as stated in Lemma 2.8. In particular, $\|\hat{\beta}_n - \beta_0\|_2 \to^p 0$.

(*ii*) *It holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_\infty^2 \geq \frac{4\|\beta_0\|_1^2 e_n + 4\|\beta_0\|_1 \lambda_n}{F_\infty^2(T_0; \mathcal{I}_n)}\right) = 0. \tag{2.4}$$

In particular, $\|\hat{\beta}_n - \beta_0\|_\infty \to^p 0$.

(*iii*) *It holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_1 \geq \frac{2S\lambda_n}{\kappa^2(T_0; \mathcal{I}_n) - 2Se_n}\right) = 0. \tag{2.5}$$

In particular, $\|\hat{\beta}_n - \beta_0\|_1 \to^p 0$.

(*iv*) *It holds for any* $q \in (1, \infty)$ *that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_q \geq \xi_{n,q}\right) = 0, \tag{2.6}$$

where

$$\xi_{n,q} = \frac{2S^{\frac{1}{q}} e_n}{F_q(T_0; \mathcal{I}_n)} \cdot \frac{2S\lambda_n}{\kappa^2(T_0; \mathcal{I}_n) - 2Se_n} + \frac{2S^{\frac{1}{q}}\lambda_n}{F_q(T_0; \mathcal{I}_n)}.$$

In particular, $\|\hat{\beta}_n - \beta_0\|_q \to^p 0$.

Proof. Put $h = \hat{\beta}_n - \beta_0$. It is sufficient to prove that

$$\|\psi_n(\beta_0)\|_\infty \leq \lambda_n$$

implies the inequality (2.3)-(2.6).

(i) and (ii) It is obvious that

$$|h^\top J_n h| = |h^\top(\psi_n(\hat{\beta}_n) - \psi_n(\beta_0))|.$$

Noting that $\hat{\beta}_n \in \mathcal{C}_n$ and $\|\hat{\beta}_n\|_1 \leq \|\beta_0\|_1$ by the definition of the Dantzig selector, we have that

$$\begin{aligned} |h^\top J_n h| &\leq \|h\|_1 \|\psi_n(\hat{\beta}_n) - \psi_n(\beta_0)\|_\infty \\ &\leq 2\|\beta_0\|_1 2\lambda_n \\ &= 4\|\beta_0\|_1 \lambda_n. \end{aligned}$$

We therefore obtain that

$$\begin{aligned} |h^\top \mathcal{I}_n h| &\leq |h^\top(J_n - \mathcal{I}_n)h| + |h^\top J_n h| \\ &\leq \|h\|_1^2 e_n + 4\|\beta_0\|_1\lambda_n \\ &\leq 4\|\beta_0\|_1^2 e_n + 4\|\beta_0\|_1\lambda_n. \end{aligned}$$

Moreover, we can prove that $h = \hat{\beta}_n - \beta_0 \in C_{T_0}$ as follows:

$$\begin{aligned} 0 \geq \|\beta_0 + h\|_1 - \|\beta_0\|_1 &= \sum_{j\in T_0^c} |h_{T_{0j}^c}| + \sum_{j\in T_0} (|\beta_{0j} + h_{T_{0j}}| - |\beta_{0j}|) \\ &\geq \sum_{j\in T_0^c} |h_{T_0^c j}| - \sum_{j\in T_0} |h_{T_{0j}}| \\ &= \|h_{T_0^c}\|_1 - \|h_{T_0}\|_1. \end{aligned}$$

It follows from the definition of $RE(T_0; \mathcal{I}_n)$ that

$$\begin{aligned} RE^2(T_0; \mathcal{I}_n) &\leq \frac{h^\top \mathcal{I}_n h}{\|h\|_2^2} \\ &\leq \frac{4\|\beta_0\|_1^2 e_n + 4\|\beta_0\|_1 \lambda_n}{\|h\|_2^2}, \end{aligned}$$

which implies the conclusion in (i). Using $F_\infty(T_0; \mathcal{I}_n)$ instead of $RE(T_0; \mathcal{I}_n)$ in the above inequality, we obtain the conclusion in (ii).

(iii) and (iv) Similarly to the proof of (i), we have that

$$|h^\top \mathcal{I}_n h| \leq \|h\|_1^2 e_n + 2\|h\|_1 \lambda_n.$$

From the definition of $\kappa(T_0; \mathcal{I}_n)$, we have that

$$\kappa^2(T_0; \mathcal{I}_n) \leq \frac{S\{\|h\|_1^2 e_n + 2\|h\|_1 \lambda_n\}}{\|h_{T_0}\|_1^2}.$$

Noting that

$$\|h\|_1 \leq 2\|h_{T_0}\|_1,$$

We obtain that

$$\|h\|_1 \leq \frac{2S\lambda_n}{\kappa^2(T_0; \mathcal{I}_n) - 2Se_n},$$

which implies the conclusion in (iii). Moreover, it follows from the definition of $F_q(T_0; \mathcal{I}_n)$ for every $q \in (1, \infty)$ that

$$F_q(T_0; \mathcal{I}_n) \leq \frac{S^{\frac{1}{q}}\{\|h\|_1^2 e_n + 2\|h\|_1 \lambda_n\}}{\|h_{T_0}\|_1 \|h\|_q},$$

which implies the conclusion in (iv). □

2.4 The variable selection consistency

In this subsection, we will show that $\hat{\beta}_n$ satisfies the variable selection consistency. To do this, we construct an estimator for the support index set T_0 of the true value β_0 as follows:

$$\hat{T}_n := \{j : |\hat{\beta}_n^j| > \lambda_n\}. \tag{2.7}$$

The next theorem states that $\hat{T}_n = T_0$ with probability tending to 1, which means the variable selection consistency of the Dantzig selector.

Theorem 2.13. *Under Assumptions 2.1, 2.2, 2.6 and 2.10, it holds that*

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

Proof. Note that $\|\hat{\beta}_n - \beta_0\|_\infty \leq \|\hat{\beta}_n - \beta_0\|_1$ and that the sparsity S is assumed to be fixed. We have that

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

by the l_1 bound from Theorem 2.6 (i). Therefore, it is sufficient to show that the next inequality

$$\|\hat{\beta}_n - \beta_0\|_\infty \leq \lambda_n$$

implies that

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

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

$$|\beta_0^j| - |\hat{\beta}_n^j| \leq |\hat{\beta}_n^j - \beta_0^j| \leq \lambda_n.$$

We have that

$$|\hat{\beta}_n^j| \geq |\beta_0^j| - \gamma_{n,p_n} > \lambda_n$$

for sufficiently large n, which implies that $T_0 \subset \hat{T}_n$. On the other hand, for every $j \in T_0^c$, we have that

$$|\hat{\beta}_n^j - \beta_0^j| = |\hat{\beta}_n^j| \leq \lambda_n$$

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

2.5 An asymptotically normal estimator post variable selection

Using the estimator $\hat{T}_n$ for T_0, we can reduce the dimension which enables us to construct an asymptotically normal estimator. We construct an asymptotically normal estimator $\hat{\beta}_n^{(2)}$ by the solution to the next equation:

$$\psi_n(\beta)_{\hat{T}_n} = 0, \quad \beta_{\hat{T}_n^c} = 0. \tag{2.8}$$

In this section, we prove the asymptotic normality under the next condition.

Assumption 2.14. *The $S \times S$ sub-matrix $\mathcal{I}_{nT_0,T_0}$ is positive definite.*

Using the central limit theorem and Slutsky's lemma, we can prove the following theorem.

Theorem 2.15. *Under Assumptions 2.1, 2.2, 2.6, 2.10 and 2.14, it holds that*

$$\sqrt{n}\left(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0}\right) 1_{\{\hat{T}_n = T_0\}} \to^d N\left(0, \mathcal{I}^{-1}_{nT_0,T_0}\right), \quad n \to \infty.$$

Proof. Since we have already proved that

$$\|J_n - \mathcal{I}_n\|_\infty \to^p 0$$

as $n \to \infty$, we have the corresponding result for $S \times S$ sub-matrices. We therefore obtain that

$$\sqrt{n}\left(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0}\right) 1_{\{\hat{T}_n = T_0\}} = \mathcal{I}_n^{-1} \frac{1}{\sqrt{n}} Z^\top_{\hat{T}_n} \epsilon 1_{\{\hat{T}_n = T_0\}} + o_p(1).$$

Noting that it follows from Lemma 2.13 that

$$1_{\{\hat{T}_n = T_0\}} \to^p 1, \quad n \to \infty,$$

we obtain the conclusion by using the central limit theorem for independent random sequences and Slutsky's lemma. □

2.6 Concluding remarks

2.6.1 Remarks on the tuning parameter

Our results strongly depends on the tuning parameter λ_n. To ensure our l_q consistency results, it is sufficient that λ_n satisfies Assumption 2.6. We therefore can choose λ_n, for example, by

$$\lambda_n = c_0 \tilde{\lambda}_n,$$

where $c_0 > 0$ is a positive constant and

$$\tilde{\lambda}_n = \left(\frac{\log p_n}{n}\right)^\alpha, \quad \alpha \in \left(0, \frac{1}{2}\right).$$

The exponent α can be chosen close to $1/2$ and constant c_0 can be chosen arbitrary for asymptotic results. However, to ensure the finite sample performance for the variable selection, how to choose c_0 is an important point. We will discuss this problem in Chapter 6, Numerical studies.

2.6.2 Summary

The Dantzig selector can be defined by the following form for general parametric models with unknown parameter θ:

$$\hat{\theta}_n := \arg\min_{\theta \in \mathcal{C}_n} \|\theta\|_1, \quad \mathcal{C}_n = \{\theta \in \mathbb{R}^{p_n} : \|\Psi_n(\theta)\|_\infty \leq \gamma_n\},$$

where $\gamma_n \geq 0$ is a tuning parameter and $\Psi_n(\cdot)$ is the normalized score function for the model. To verify the l_q consistency of this estimator,we need the following two properties:

Proposition 2.16. *It holds that*

$$\lim_{n\to\infty} P\left(\|\Psi_n(\theta_0)\|_\infty \geq \tilde{\gamma}_n\right) = 0,$$

where θ_0 is a true value of θ and $\tilde{\gamma}_n \to 0$ as $n \to \infty$.

Assumption 2.17. *Let $-M_n(\theta)$ be Hessian matrix of the model. For every $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$*

$$P(\kappa(T_0; M_n(\theta_0)) > \delta) \geq 1 - \epsilon.$$

where $T_0 = \{j : \theta_0^j \neq 0\}$ is the support index set of θ_0 and $\kappa(T_0; M_n(\theta_0))$ is defined by the same way as in Definition 2.9.

In this chapter, we prove the first property for linear regression model (2.1) by using the maximal inequality for Orlicz norm under sub-Gaussian settings to deal with high-dimensional settings. For other models of stochastic processes in this paper, we can prove the corresponding property by using stochastic inequalities for martingales instead of sub-Gaussian properties and maximal inequality for Orlicz norm. The second property for Hessian matrices is assumed in various models, since this is not so strong assumption as we mentioned in this chapter if the true value has sparse structure.

Moreover, if the tuning parameter has the same asymptotic rate as the rate of convergence of $\|\Psi_n(\theta_0)\|_\infty$, we can prove that the rate of convergence of the l_q error for every $q \in [1, \infty]$ can be written by the tuning parameter γ_n. Using this fact, we can construct an consistent estimator $\hat{T}_n$ for T_0 by the thresholding method using the tuning parameter. We therefore can construct an asymptotically normal estimator $\hat{\beta}_n^{(2)}$ by using $\hat{T}_n$.

Chapter 3

Cox's proportional hazards model

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 S 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. Especially, the penalized methods such as Lasso (Tibshirani (1997), Huang et al. (2013), Bradic et al. (2011) and others) have been 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. 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 will extend the consistency results by Antoniadis et al. (2010) to the l_q consistency for every $q \in [1, \infty]$ by a method similar to that of Huang et al. (2013). Moreover, we will establish the asymptotic normalities of estimators in a high-dimensional

and sparse setting based on the consistency results as in the previous chapter. 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. On the other hand, we will show that the Dantzig selector has a variable selection consistency, which enables us to reduce the dimension. Next, we will 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 prove that a Breslow type estimator, which is obtained by using the maximum partial likelihood estimator after dimension reduction, satisfies the asymptotic normality. In addition, we will observe whether our selection criterion works well for simple models numerically and compare the estimators to the classical maximum partial likelihood estimator.

This chapter 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 Section 3.1. In Section 3.2, we prove the l_q consistency of the estimators for the regression parameter. The variable selection consistency of the Dantzig selector is proved in Section 3.3 and the asymptotic normality of the maximum partial likelihood estimator after dimension reduction in Section 3.4. The asymptotic property of the Breslow estimator is established in Section 3.5. Summary and an example which satisfies our matrix condition are presented in Section 3.6.

3.1 Model setups

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 $i = 1, 2, \ldots, n$ 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:

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

Suppose that $\{Z_i(t)\}_{t\in[0,1]},\ i = 1, 2, \ldots, n$ 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 $i = 1, 2, \ldots, n$ has the following intensity:

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

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

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

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

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

and

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

Hereafter, we write $\Lambda(\cdot)$ for the cumulative baseline hazard function, *i.e.*,

$$\Lambda(t) := \int_0^t \lambda(s)ds, \quad t \in [0, 1].$$

The aim of this chapter is to estimate the true value β_0 of the regression parameter β and the true function $\Lambda_0(\cdot)$ of the cumulative baseline hazard function $\Lambda(\cdot)$ with respect to the true baseline function $\lambda_0(\cdot)$ in a high-dimensional and sparse setting for β_0, *i.e.*, $p = p_n \to \infty$ as $n \to \infty$ and $S := |T_0|$ is a fixed constant which is independent of n, where $T_0 := \{j; \beta_0^j \neq 0\}$ is the support index set of the true value. To estimate β_0, we use Cox's log-partial likelihood which is given by;

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

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.*,

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

and

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

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:

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

3.1.1 Regularity conditions and matrix conditions

We assume the following conditions.

Assumption 3.1. (*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.,*

$$\int_0^1 \lambda_0(t) dt < \infty.$$

(*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:*

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

$$\liminf_{n\to\infty} \inf_{\beta} \inf_{t\in[0,1]} s_n^{(l)}(\beta, t) > 0.$$

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

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

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

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

where

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

We define the estimator $\hat{\beta}_n$ of β_0 as

$$\hat{\beta}_n := \arg\min_{\beta\in\mathcal{B}_n} \|\beta\|_1, \tag{3.1}$$

where $\mathcal{B}_n := \{\beta \in \mathbb{R}^{p_n} : \|U_n(\beta)\|_\infty \le \gamma\}$, and $\gamma \ge 0$ is a suitable constant. We call the estimator $\hat{\beta}_n$ the Dantzig Selector for Proportional Hazards model (DSfPH). Note that this estimator is called the Survival Dantzig Selector (SDS) by Antoniadis et al. (2010).

Now, we introduce some matrix conditions to derive the theoretical results for DSfPH $\hat{\beta}_n$. Hereafter, we write T_0 for the support of β_0, *i.e.*,

$$T_0 := \{j : \beta_0^j \neq 0\}.$$

To begin with, we introduce the following three factors (A), (B) and (C), all of which are used by Huang et al. (2013) for Lasso in Cox's proportional hazards model.

Definition 3.2. *For every index set* $T \subset \{1,\ 2,\ \cdots,\ p_n\}$ *and* $h \in \mathbb{R}^{p_n}$, h_T *is a* $\mathbb{R}^{|T|}$ *dimensional sub-vector of* h *constructed by extracting the components of* h *corresponding to the indices in* T. *Define the set* C_T *by*

$$C_T := \{h \in \mathbb{R}^{p_n} : \|h_{T^c}\|_1 \leq \|h_T\|_1\}.$$

We introduce the following three factors.

(A) **Compatibility factor**

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

(B) **Weak cone invertibility factor**

$$\begin{aligned} F_q(T_0; I_n(\beta_0)) &:= \inf_{0 \neq h \in C_{T_0}} \frac{S^{\frac{1}{q}} h^\top I_n(\beta_0)h}{\|h_{T_0}\|_1 \|h\|_q}, \quad q \in [1, \infty), \\ F_\infty(T_0; I_n(\beta_0)) &:= \inf_{0 \neq h \in C_{T_0}} \frac{(h^\top I_n(\beta_0)h)^{\frac{1}{2}}}{\|h\|_\infty}. \end{aligned}$$

(C) **Restricted eigenvalue**

$$RE(T_0; I_n(\beta_0)) := \inf_{0 \neq h \in C_{T_0}} \frac{(h^\top I_n(\beta_0)h)^{\frac{1}{2}}}{\|h\|_2}.$$

Huang et al. (2013) defined these factors for the random matrix $J_n(\beta_0)$, and derived some conditions to treat them as deterministic constants. On the other hand, we define them not for $J_n(\beta_0)$, but for the deterministic matrix $I_n(\beta_0)$, since we will prove that $\|I_n(\beta_0) - J_n(\beta_0)\|_\infty = o_p(1)$ later. Noting that $\|h_{T_0}\|_1^q \geq \|h_{T_0}\|_q^q$ for all $q \in [1, \infty)$, and $\|h\|_\infty \leq \|h\|_1$, we can see that $\kappa(T_0; I_n(\beta_0)) \leq 2\sqrt{S} RE(T_0; I_n(\beta_0))$, $\kappa(T_0; I_n(\beta_0)) \leq F_q(T_0; I_n(\beta_0))$ and $\kappa(T_0; I_n(\beta_0)) \leq 2\sqrt{S} F_\infty(T_0; I_n(\beta_0))$. We therefore assume in our main theorems that c factors are "asymptotically positive".

Assumption 3.3. *It holds that*

$$\liminf_{n \to \infty} \kappa(T_0; I_n(\beta_0)) > 0.$$

An example for this matrix condition is provided in Section 3.6.

3.2 Consistency

In this section, we will prove the consistency of DSfPH $\hat{\beta}_n$. To do this, we will prepare three lemmas. The next lemma states that the true parameter β_0 is an element of $\mathcal{B}_n$ appearing in (3.1) with large probability when the sample size n is large.

Lemma 3.4. *Put $\gamma = \gamma_{n,p_n} = K_2 \log(1+p_n)/n^\alpha$, where $0 < \alpha \leq 1/2$ and $K_2 > 0$ are constants. If $\log p_n = O(n^\zeta)$ for some $0 < \zeta < \alpha$, then it holds that*

$$\lim_{n\to\infty} P(\|U_n(\beta_0)\|_\infty \geq \gamma_{n,p_n}) = 0$$

and that $\gamma_{n,p_n} \to 0$ as $n \to \infty$.

Proof. Recall that

$$U_n^j(\beta_0, u) = \frac{1}{n}\sum_{i=1}^n \int_0^u \left[\sum_{k=1}^n \{Z_i^j(u) - Z_k^j(u)\} w_k(\beta_0, u)\right] dM_i(u),$$

where $u \in [0,1]$ and

$$w_k(\beta_0, u) = \frac{\exp(Z_k^\top(u)\beta_0) Y_k(u)}{\sum_{l=1}^n \exp(Z_l^\top(u)\beta_0) Y_l(u)}.$$

We use Lemma 2.1 from van de Geer (1995). To do this, we shall evaluate $\Delta U_n^j(\beta_0, u),\ u \in [0,1]$, and $\langle U_n^j(\beta_0,\cdot), U_n^j(\beta_0,\cdot)\rangle_1$. Since the jumps of M_i do not occur at the same time and are all of magnitude 1, it holds that

$$\begin{aligned}
|\Delta U_n^j(\beta_0, u)| &= \left|\frac{1}{n}\sum_{i=1}^n \int_{u-}^u \left[\sum_{k=1}^n \{Z_i^j(s) - Z_k^j(s)\} w_k(\beta_0, s)\right] dM_i(s)\right| \\
&\leq \frac{1}{n} \sup_{i,j,k,s} |Z_i^j(s) - Z_i^j(s)| \sum_{k=1}^n w_k(\beta_0, u) \\
&\leq \frac{2K_1}{n}.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\langle U_n^j(\beta_0,\cdot), U_n^j(\beta_0,\cdot)\rangle_1 &= \frac{1}{n^2}\sum_{i=1}^n \int_0^1 \left[\sum_{k=1}^n \{Z_i^j(u) - Z_k^j(u)\} w_k(\beta_0,u)\right]^2 d\langle M_i\rangle_u \\
&= \frac{1}{n^2}\sum_{i=1}^n \int_0^\tau \left[\sum_{k=1}^n \{Z_i^j(u) - Z_k^j(u)\} w_k(\beta_0,\ u)\right]^2 \exp(Z_i^\top(u)\beta_0) Y_i(u)\lambda_0(u)du \\
&\leq \frac{1}{n^2}\sup_{i,j,k,u} |Z_i^j(u) - Z_k^j(u)|^2 \sum_{i=1}^n \int_0^1 \exp(Z_i^\top(u)\beta_0) Y_i(u)\lambda_0(u)du \\
&\leq \frac{4K_1^2}{n^2} n \exp(S \sup_{i,j,u} |Z_i^j(u)| \|\beta_0\|_\infty|) \int_0^1 \lambda_0(u)du \\
&\leq \frac{K_3}{n},
\end{aligned}$$

where K_3 is a positive constant. We now use the Lemma 2.1 from van de Geer (1995):

$$\begin{aligned}
P(|U_n^j(\beta_0)| \geq \gamma_{n,p_n}) &= P\left(|U_n^j(\beta_0)| \geq \gamma_{n,p_n},\ \langle U_n^j(\beta_0,\cdot), U_n^j(\beta_0,\cdot)\rangle_1 \leq \frac{K_3}{n}\right) \\
&\leq 2\exp\left(-\frac{\gamma_{n,p_n}^2}{2\left(\frac{2K_1}{n}\gamma_{n,p_n} + \frac{K_3}{n}\right)}\right).
\end{aligned}$$

Write $\|\cdot\|_{\Phi_1}$ for the Orlicz norm with respect to $\Phi_1(x) = e^x - 1$. We apply Lemma 2.5 to deduce that there exists a constant $L > 0$ depending only on Φ_1 such that

$$\left\| \max_{1\leq j\leq p_n} |U_n^j(\beta_0)| \right\|_{\Phi_1} \leq L\left(\frac{2K_1}{n}\log(1+p_n) + \sqrt{\frac{K_3}{n}\log(1+p_n)}\right).$$

Using Markov's inequality, we have that

$$\begin{aligned}
P(\|U_n(\beta)\|_\infty \geq \gamma_{n,p_n}) &= P(\max_{1\leq j\leq p_n} |U_n^j(\beta_0)| \geq \gamma_{n,p_n}) \\
&\leq P\left(\Phi_1\left(\frac{\max\limits_{\leq j\leq p_n} |U_n^j(\beta_0)|}{\|\max\limits_{1\leq j\leq p_n} |U_n^j(\beta_0)|\|_{\Phi_1}}\right) \geq \Phi_1\left(\frac{\gamma_{n,p_n}}{\|\max\limits_{1\leq j\leq p_n} |U_n^j(\beta_0)|\|_{\Phi_1}}\right)\right) \\
&\leq \psi\left(\frac{\gamma_{n,p_n}}{\|\max\limits_{1\leq j\leq p_n} |U_n^j(\beta_0)|\|_{\Phi_1}}\right)^{-1} \\
&\leq \Phi_1\left(\frac{\gamma_{n,p_n}}{L\left(\frac{2K_1}{n}\log(1+p_n) + \sqrt{\frac{K_3}{n}\log(1+p_n)}\right)}\right)^{-1}
\end{aligned}$$

In our settings, the right-hand side of this inequality converges to 0. □

Next we will show that $J_n(\beta_0)$ is approximated by $I_n(\beta_0)$.

Lemma 3.5. *The random sequence ϵ_n defined by*

$$\epsilon_n := \|J_n(\beta_0) - I_n(\beta_0)\|_\infty$$

converges in probability to 0.

Proof. Define the $p_n \times p_n$ matrices $h_n(\beta_0, t)$ and $H_n(\beta_0, t)$ for $t \in [0, \tau]$ by

$$\begin{aligned}
h_n(\beta_0, t) &:= \frac{s_n^{(2)}}{s_n^{(0)}}(\beta_0, t) - \left(\frac{s_n^{(1)}}{s_n^{(0)}}\right)^{\otimes 2}(\beta_0, t), \\
H_n(\beta_0, t) &:= \frac{S_n^{(2)}}{S_n^{(0)}}(\beta_0, t) - \left(\frac{S_n^{(1)}}{S_n^{(0)}}\right)^{\otimes 2}(\beta_0, t).
\end{aligned}$$

Note that the matrices $I_n(\beta_0)$ and $J_n(\beta_0)$ can be written in this form:

$$\begin{aligned}
I_n(\beta_0) &= \int_0^\tau h_n(\beta_0, u) s_n^{(0)}(\beta_0, u) \lambda_0(u) du, \\
J_n(\beta_0) &= \int_0^\tau H_n(\beta_0, u) \frac{d\bar{N}(u)}{n}.
\end{aligned}$$

Put $\bar{M}(u) = \sum_{i=1}^{n} M_i(u)$. Then, it holds that $\|J_n(\beta_0) - I_n(\beta_0)\|_\infty \leq (I)+(II)+(III)$, where

$$
\begin{aligned}
(I) &= \int_0^\tau \|H_n(\beta_0, u) - h_n(\beta_0, u)\|_\infty \frac{d\bar{N}(u)}{n}, \\
(II) &= \int_0^\tau \left\| h_n(\beta_0, u) \left\{ \frac{S_n^{(0)}(\beta_0, u)}{n} - s_n^{(0)}(\beta_0, u) \right\} \right\|_\infty \lambda_0(u) du, \\
(III) &= \left\| \frac{1}{n} \int_0^\tau h_n(\beta_0, u) d\bar{M}(u) \right\|_\infty .
\end{aligned}
$$

Since the process $t \rightsquigarrow \bar{N}(t)/n$ has bounded variation uniformly in n, Assumption 2.1 implies that $(I) = o_p(1)$ and $(II) = o_p(1)$. Moreover, it follows from Assumption 2.1 that $h_n(\beta_0, u)$ is uniformly bounded. So we obtain that $(III) = o_p(1)$ by the same way as the proof of Lemma 3.4. □

The next lemma is used to control $U_n(\hat{\beta}_n) - U_n(\beta_0)$ and $J_n(\beta_0)$. See Huang et al. (2013) and Hjort and Pollard (1993) for the proofs.

Lemma 3.6. *Define that* $\eta_h = \max_{i,j,s} |h^\top Z_i(s) - h^\top Z_j(s)|$, *for* $h \in \mathbb{R}^{p_n}$. *Then for all* $\beta \in \mathbb{R}^{p_n}$, *it holds that*

$$
e^{-\eta_h} h^\top J_n(\beta) h \leq h^\top [U_n(\beta + h) - U_n(\beta)] \leq e^{\eta_h} h^\top J_n(\beta) h.
$$

Now, we are ready to prove the main result of this paper. The theorem below provides the l_2 consistency and l_∞ consistency of DSfPH.

Theorem 3.7. *Under Assumptions 3.1 and 3.3, the following (i)-(iv) hold true.*

(*i*) *It holds that*

$$
\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_2^2 \geq \frac{K_4 \gamma_{n,p_n} + K_5 \epsilon_n}{RE^2(T_0; I_n(\beta_0))} \right) = 0,
$$

where K_4, K_5 *is a positive constant and* $\epsilon_n = \|I_n(\beta_0) - J_n(\beta_0)\|_\infty = o_p(1)$. *In particular,* $\|\hat{\beta}_n - \beta_0\|_2 \to^p 0$.

(*ii*) *It holds that*

$$
\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_\infty^2 \geq \frac{K_4 \gamma_{n,p_n} + K_5 \epsilon_n}{F_\infty^2(T_0; I_n(\beta_0))} \right) = 0.
$$

In particular, $\|\hat{\beta}_n - \beta_0\|_\infty \to^p 0$.

(*iii*) *It holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_1 \geq \frac{4K_6 S\gamma_{n,p_n}}{\kappa^2(T_0; I_n(\beta_0)) - 4S\epsilon_n}\right) = 0,$$

where K_6 *is a positive constant. In particular,* $\|\hat{\beta}_n - \beta_0\|_1 \to^p 0$.

(*iv*) *It holds for any* $q \in (1,\infty)$ *that*

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_q \geq \xi_{n,q}\right) = 0,$$

where

$$\xi_{n,q} = \frac{2S^{\frac{1}{q}}\epsilon_n}{F_q(T_0; I_n(\beta_0))} \cdot \frac{2K_6 S\gamma_{n,p_n}}{\kappa^2(T_0; I_n(\beta_0)) - 2S\epsilon_n} + \frac{2K_6 S^{\frac{1}{q}}\gamma_{n,p_n}}{F_q(T_0; I_n(\beta_0))}.$$

In particular, $\|\hat{\beta}_n - \beta_0\|_q \to^p 0$.

Proof. (i) and (ii) It is sufficient to prove that $\|U_n(\beta_0)\|_\infty \leq \gamma_{n,p_n}$ implies

$$\|\hat{\beta}_n - \beta_0\|_2^2 \leq \frac{K_4\gamma_{n,p_n} + K_5\epsilon_n}{RE^2(T_0; I_n(\beta_0))}.$$

By the construction of the estimator, we have $\|U(\hat{\beta}_n)\|_\infty \leq \gamma_{n,p_n}$, which implies that

$$\|U_n(\hat{\beta}_n) - U_n(\beta_0)\|_\infty \leq \|U_n(\hat{\beta}_n)\|_\infty + \|U_n(\beta_0)\|_\infty \leq 2\gamma_{n,p_n}.$$

Note that $h := \hat{\beta} - \beta_0 \in C_{T_0}$, since it holds that

$$\begin{aligned} 0 \geq \|\beta_0 + h\|_1 - \|\beta_0\|_1 &= \sum_{j\in T_0^c} |h_{T_{0j}^c}| + \sum_{j\in T_0} (|\beta_{0j} + h_{T_{0j}}| - |\beta_{0j}|) \\ &\geq \sum_{j\in T_0^c} |h_{T_0^c j}| - \sum_{j\in T_0} |h_{T_{0j}}| \\ &= \|h_{T_0^c}\|_1 - \|h_{T_0}\|_1. \end{aligned}$$

Notice moreover that $\|h\|_1 \leq \|\hat{\beta}_n\|_1 + \|\beta_0\|_1 \leq 2\|\beta_0\|_1$ by the definition of $\hat{\beta}_n$. Now, we use Lemma 3.6 for h to deduce that

$$\begin{aligned} h^\top J_n(\beta_0) h &\leq e^{\eta_h} h^\top [U_n(\hat{\beta}_n) - U_n(\beta_0)] \\ &\leq \exp(\max_{i,j,u} |h^\top Z_i(u) - h^\top Z_j(u)|) \cdot 2\gamma_{n,p_n}\|h\|_1 \\ &\leq \exp(4K_1\|\beta_0\|_1) \cdot 4\gamma_{n,p_n}\|\beta_0\|_1 \\ &=: K_4\gamma_{n,p_n}. \end{aligned}$$

Thus it holds that

$$\begin{aligned} h^\top I_n(\beta_0)h &\le |h^\top (I_n(\beta_0) - J_n(\beta_0))h| + h^\top J_n(\beta_0)h \\ &\le \epsilon_n \|h\|_1^2 + K_4\gamma_{n,p_n} \\ &\le 4\epsilon_n \|\beta_0\|_1^2 + K_4\gamma_{n,p_n} \\ &\le K_4\gamma_{n,p_n} + K_5\epsilon_n. \end{aligned}$$

By the definition of the restricted eigenvalue, we have that

$$\begin{aligned} RE^2(T_0; I_n(\beta_0)) &\le \frac{h^T I_n(\beta_0)h}{\|\hat{\beta}_n - \beta_0\|_2^2} \\ &\le \frac{K_4\gamma_{n,p_n} + K_5\epsilon_n}{\|\hat{\beta}_n - \beta_0\|_2^2}. \end{aligned}$$

Noting that $RE^2(T_0; I_n(\beta_0)) > 0$, we obtain that

$$\|\hat{\beta}_n - \beta_0\|_2^2 \le \frac{K_4\gamma_{n,p_n} + K_5\epsilon_n}{RE^2(T_0; I_n(\beta_0))},$$

which is the conclusion in (i).

By the definition of $F_\infty(T_0; I_n(\beta_0))$, we have also that

$$F_\infty^2(T_0; I_n(\beta_0)) \le \frac{K_4\gamma_{n,p_n} + K_5\epsilon_n}{\|\hat{\beta}_n - \beta_0\|_\infty^2},$$

which yields the conclusion in (ii).

(iii) and (iv) It follows from the proof of (i) that

$$h^T J_n(\beta_0)h \le K_6\gamma_{n,p_n}\|\hat{\beta}_n - \beta_0\|_1.$$

We have also that

$$h^T I_n(\beta_0)h \le \epsilon_n\|\hat{\beta}_n - \beta_0\|_1^2 + K_6\gamma_{n,p_n}\|\hat{\beta}_n - \beta_0\|_1.$$

The definition of $\kappa(T_0; I_n(\beta_0))$ implies that

$$\begin{aligned} \kappa^2(T_0; I_n(\beta_0)) &\le \frac{Sh^T I_n(\beta_0)h}{\|h_{T_0}\|_1^2} \\ &\le \frac{S\epsilon_n\|h\|_1^2 + K_6 S\gamma_{n,p_n}\|h\|_1}{\|h_{T_0}\|_1^2}. \end{aligned}$$

Since $\|h\|_1 \leq 2\|h_{T_0}\|_1$, this yields the conclusion in (iii).

On the other hand, using the weak cone invertibility factor for every $q \geq 1$, we have that

$$F_q(T_0; I_n(\beta_0)) \leq \frac{S^{\frac{1}{q}}\epsilon_n \|h\|_1^2 + S^{\frac{1}{q}} K_6 \gamma_{n,p_n} \|h\|_1}{\|h_{T_0}\|_1 \|h\|_q},$$

which implies that

$$\|\hat{\beta}_n - \beta_0\|_q \leq \frac{2S^{\frac{1}{q}}\epsilon_n \|\hat{\beta} - \beta_0\|_1 + 2S^{\frac{1}{q}} K_6 \gamma_{n,p_n}}{F_q(T_0; I_n(\beta_0))}.$$

Using the l_1 bound derived above, we obtain the conclusion in (iv). □

3.3 The variable selection consistency of the Dantzig selector

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

$$\hat{T}_n := \{j; |\hat{\beta}_n^j| > \gamma_{n,p_n}\}. \tag{3.2}$$

The following theorem states that $\hat{\beta}_n$ has a variable selection consistency.

Theorem 3.8. *Under Assumptions 3.1 and 3.3, it holds that*

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

Proof. Note that $\|\hat{\beta}_n - \beta_0\|_\infty \leq \|\hat{\beta}_n - \beta_0\|_1$ and that the sparsity S is assumed to be fixed. We have that

$$\lim_{n\to\infty} P\left(\|\hat{\beta}_n - \beta_0\|_\infty > \gamma_{n,p_n}\right) = 0$$

by the l_1 bound from Theorem 3.7 (*iii*). Therefore, it is sufficient to show that the next inequality

$$\|\hat{\beta}_n - \beta_0\|_\infty \leq \gamma_{n,p_n}$$

implies that

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

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

$$|\beta_0^j| - |\hat{\beta}_n^j| \leq |\hat{\beta}_n^j - \beta_0^j| \leq \gamma_{n,p_n}.$$

We have that

$$|\hat{\beta}_n^j| \geq |\beta_0^j| - \gamma_{n,p_n} > \gamma_{n,p_n}$$

for sufficiently large n, which implies that $T_0 \subset \hat{T}_n$. On the other hand, for every $j \in T_0^c$, we have that

$$|\hat{\beta}_n^j - \beta_0^j| = |\hat{\beta}_n^j| \leq \gamma_{n,p_n}$$

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

3.4 The maximum partial likelihood estimator for the regression parameter after dimension reduction

Using the set $\hat{T}_n$, we construct a new estimator $\hat{\beta}_n^{(2)}$ by the solution to the next equation:

$$U_n(\beta_{\hat{T}_n}) = 0, \quad \beta_{\hat{T}_n^c} = 0. \tag{3.3}$$

We prove the asymptotic normality of $\hat{\beta}_n^{(2)}$.

Assumption 3.9. *In this subsection, we assume that the following $S \times S$ matrix $\mathcal{I}$ is positive definite:*

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

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

Theorem 3.10. *Under Assumptions 3.1, 3.3 and 3.9, it holds that*

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

as $n \to \infty$.

Proof. 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. □

To show the asymptotic normality of $\hat{\beta}_n^{(2)}$, we need to prove the next lemma.

Lemma 3.11. *Under Assumptions 3.1, 3.3 and 3.9, 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 $n \to \infty$.

Proof. Define

$$V_n(\beta, t) := \frac{S_n^{(2)}}{S_n^{(0)}}(\beta, t) - \left(\frac{S_n^{(1)}}{S_n^{(0)}}\right)^{\otimes 2}(\beta, t)$$

and

$$v_n(\beta, t) := \frac{s_n^{(2)}}{s_n^{(0)}}(\beta, t) - \left(\frac{s_n^{(1)}}{s_n^{(0)}}\right)^{\otimes 2}(\beta, t).$$

We have that

$$J_n(\beta_n^*) = \int_0^1 V_n(\beta_n^*, t)dt$$

and that

$$I_n(\beta_0) = \int_0^1 v_n(\beta_0, t)ds.$$

Therefore it holds that $\|J_n(\beta_n^*) - I_n(\beta_0)\|_\infty \leq (I) + (II) + (III) + (IV)$, where

$$(I) = \left\| \int_0^1 \{V_n(\beta_n^*, t) - V_n(\beta_0, t)\} \frac{d\bar{N}(t)}{n} \right\|_\infty ,$$

$$(II) = \left\| \int_0^1 \{V_n(\beta_0, t) - v_n(\beta_0, t)\} \frac{d\bar{N}(t)}{n} \right\|_\infty ,$$

$$(III) = \left\| \int_0^1 v_n(\beta_0, t) \left\{ \frac{1}{n} S_n^{(0)}(\beta_0, t) - s_n^{(0)}(\beta_0, t) \right\} \lambda_0(t) dt \} \right\|_\infty$$

and

$$(IV) = \left\| \int_0^1 v_n(\beta_0, t) \frac{d\bar{M}(t)}{n} \right\|_\infty .$$

Since the process $t \rightsquigarrow \bar{N}(t)/n$ has bounded variation, Assumption 3.1 implies that (II) and (III) are $o_p(1)$. In addition, we have for every $l = 0, 1, 2$ and $t \in [0, 1]$ that

$$\begin{aligned} & \frac{1}{n} \|S_n^{(l)}(\beta_n^*, t) - S_n^{(l)}(\beta_0, t)\|_\infty \\ \leq \; & \frac{1}{n} \left\| \sum_{i=1}^n Y_i(t) Z_i(t)^{\otimes l}(t) \exp(\beta_0^\top Z_i(t)) \left\{ \exp[\|Z_i(t)\|_\infty \|\beta_n^* - \beta_0\|_1] - 1 \right\} \right\|_\infty \\ \leq \; & K_1 \exp(K_1 \|\beta_0\|_1) |\exp(K_1 \|\beta_n^* - \beta_0\|_1) - 1|. \end{aligned}$$

We have that the right-hand side of this inequality converges to 0 in probability when $\|\beta_n^* - \beta_0\|_1 \to^p 0$ as $n \to \infty$. This fact and the continuous mapping theorem imply that $\|V_n(\beta_n^*, t) - V_n(\beta_0, t)\|_\infty \to^p 0$ as $n \to \infty$. Therefore, we obtain that $(I) = o_p(1)$. Finally, we can see that $(IV) = o_p(1)$ by using Bernstein's inequality for martingales (see e.g. van de Geer (1995)). □

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

Theorem 3.12. *Under Assumptions 3.1, 3.3 and 3.9, it holds that*

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

Proof. It follows from the Taylor expansion that

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

where β_n^* is the point between $\hat{\beta}_n^{(2)}$ and β_0. We therefore have that

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

We can see that $\sqrt{n}U_{nT_0}(\beta_{0T_0})$ is the terminal value of the S-dimensional martingale $\{\tilde{M}_n(t)\}_{t\in[0,1]}$ defined by

$$\tilde{M}_n(t) := \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \int_0^t \left\{ Z_{iT_0}(s) - \frac{S^{(1)}_{nT_0}}{S^{(0)}_n}(\beta_{0T_0}, s) \right\} dM_i(s).$$

It holds that

$$\begin{aligned} \langle \tilde{M}_n, \tilde{M}_n \rangle(1) &= \frac{1}{n} \int_0^1 V_{nT_0,T_0}(\beta_{0T_0,t}) S^{(0)}_n(\beta_{0T_0}, t) \lambda_0(t) dt \\ &\to^p \mathcal{I} \end{aligned}$$

as $n \to \infty$, where $V_n(\beta_0, t)$ is defined in proof of Lemma 3.11. By this fact and Lindeberg's condition assumed in Assumption 3.1, we can apply the martingale central limit theorem to $\tilde{M}_n(t)$ to deduce that

$$\sqrt{n}U_n(\beta_{0T_0}) \to^d N(0, \mathcal{I})$$

as $n \to \infty$. It follows from Theorem 3.10 and Lemma 3.11 that $\|J_{nT_0,T_0}(\beta^*_{nT_0}) - \mathcal{I}\|_\infty = o_p(1)$. We therefore have that

$$\sqrt{n}(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0}) 1_{\{\hat{T}_n = T_0\}} = \mathcal{I}^{-1} \sqrt{n} U_{nT_0}(\beta_{0T_0}) 1_{\{\hat{T}_n = T_0\}} + o_p(1).$$

Since Theorem 3.8 implies that $1_{\{\hat{T}_n = T_0\}} \to^p 1$ as $n \to \infty$, we obtain the conclusion by using Slutsky's theorem. □

3.5 The estimator for the cumulative baseline hazard function

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

$$\hat{\Lambda}(t) := \int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) \exp(\hat{\beta}^{(2)T}_n Z_i(s))}, \quad t \in [0,1], \tag{3.4}$$

where $\hat{\beta}^{(2)}_n$ is defined by the equation (3.3). 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

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

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

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 3.1 that

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

Moreover, we have that (II) equals 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 3.10, we can see that

$$\sup_{t \in [0,1]} \left\| H_n(\beta_n^*, t) + \int_0^t \frac{s_n^{(1)}}{s_n^{(0)}}(\beta_0, s) \lambda_0(s) ds \right\|_\infty = o_p(1) \tag{3.5}$$

by a similar way to the proof of Lemma 3.11. 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 3.13. *Under Assumptions 3.1, 3.3 and 3.9, 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*

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

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

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

Proof. We have that

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

We can use the fact (3.5) to deduce that

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

We apply the martingale central limit theorem to the process $\{\sqrt{n}W_n(t)\}_{t\in[0,1]}$. It holds that

$$\begin{aligned}
\langle \sqrt{n}W_n(\cdot), \sqrt{n}W_n(\cdot)\rangle(t) &= \int_0^t \frac{\lambda_0(s)}{n^{-1}S_n(\beta_0, s)^{(0)}} ds \\
&\to^p \int_0^t \frac{\lambda_0(s)}{s^{(0)}(\beta_{0T_0}, s)} ds
\end{aligned}$$

as $n \to \infty$. Moreover, we can see Lindeberg's condition, *i.e.*, it holds for every $\epsilon > 0$ that,

$$\begin{aligned}
\int_0^t \frac{1}{n^{-2}\{S_n^{(0)}(\beta_0, s)\}^2} 1_{\{n^{-1}S_n^{(0)}(\beta_0,s) > \epsilon\}} S_n^{(0)}(\beta_0, s)\lambda_0(s) ds &< \frac{\epsilon}{n}\int_0^t \lambda_0(s) ds \\
&\to 0.
\end{aligned}$$

Therefore, we have that $\{\sqrt{n}W_n(t)\}_{t\in[0,1]}$ is asymptotically distributed as a Gaussian martingale with variance function

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

by using the martingale central limit theorem. Next, we check the asymptotic orthogonality. It follows from a direct calculation that

$$\langle U^j_{nT_0}(\beta_{0T_0}, \cdot), W_n(\cdot)\rangle(t) = 0, \quad t \in [0,1], \quad j \in T_0$$

where

$$U_{nT_0}(\beta_{0T_0}, t) := \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{ Z_{iT_0}(s) - \frac{S^{(1)}_{nT_0}}{S^{(0)}_{nT_0}}(\beta_{0T_0}, s) dM_i(s) \right\}.$$

Since $\hat{\beta}_{nT_0} - \beta_{0T_0}$ is the linear combination of $U_{nT_0}(\beta_{0T_0})$, we have that $\sqrt{n}(\hat{\beta}^{(2)}_{n\hat{T}_n} - \beta_{0T_0})1_{\{\hat{T}_n = T_0\}}$ and

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

are asymptotically orthogonal for every $t \in [0,1]$. This fact implies that they are asymptotically independent because both are asymptotically normal distributed. Noting that $1_{\{\hat{T}_n = T_0\}} \to^p 1$ as $n \to \infty$, we obtain the conclusion by using Slutsky's theorem. □

3.6 Concluding remarks

3.6.1 An example for matrix conditions

In this subsection, we provide an example which satisfies the high level matrix condition Assumption 3.3 in a high-dimensional setting.

Let covariates $\{Z_i\}_{i=1}^n$ be $\mathbb{R}^p$-valued *i.i.d.* random vectors which are independent of the time $t \in [0,1]$. For each $i \in \{1, \ldots, n\}$, we assume that $\{Z_i^j\}_{j=1}^p$ are independent and bounded random variables. Moreover, we suppose that the censoring time C_i is independent of T_i for every $i = 1, \ldots, n$, the baseline hazard function $\lambda(\cdot)$ is integrable and the true value $\beta_0 = (\beta_0^1, \beta_0^2, \ldots, \beta_0^S, 0, \ldots, 0) \in \mathbb{R}^p$. Under these assumptions, T_i and Y_i are independent of Z_i^l for every $l = S+1, \ldots, p$. In this case, we have that

$$S_n^0(\beta_0, t) = \sum_{i=1}^n Y_i(t)\exp(\beta_0^\top Z_i),$$

$$S_n^1(\beta_0, t) = \sum_{i=1}^n Z_i Y_i(t)\exp(\beta_0^\top Z_i)$$

and

$$S_n^2(\beta_0, t) = \sum_{i=1}^n Z_i Z_i^\top Y_i(t)\exp(\beta_0^\top Z_i).$$

By using the weak law of large numbers for each component of vector or matrix, we obtain that

$$s_n^{(0)}(\beta_0, t) = E[Y_1(t)\exp(\beta_0^\top Z_i)] = \prod_{l=1}^{S} E[Y_1(t)\exp(\beta_0^l Z_1^l)],$$

$$s_{nj}^{(1)}(\beta_0, t) = E[Z_1^j Y_1(t)\exp(\beta_0^\top Z_1)], \quad j \in \{1, 2, \ldots, p\}$$

and

$$s_{njk}^{(2)}(\beta_0, t) = E[Z_1^j Z_1^k Y_1(t)\exp(\beta_0^\top Z_1)], \quad j, k \in \{1, 2, \ldots, p\}.$$

For $s_n^{(0)}(\beta_0, t)$, it holds that

$$\begin{aligned} s_n^{(0)}(\beta, t) &= E[\exp(\beta_0^\top Z_1) E[1_{\{X_1 \geq t\}} | Z_1]] \\ &= E\left[\prod_{l=1}^{S} \exp(\beta_0^l Z_1^l)\pi_Z(t)\right], \end{aligned}$$

where $\pi_Z(\cdot)$ is the survival function of $X_1 = T_1 \wedge C_1$, which is independent of Z_1^l for every $S < l \leq p$. We can also calculate $s_{nj}^{(1)}(\beta_0, t)$ and $s_{njk}^{(2)}(\beta_0, t)$ for every $j, k \in \{1, 2, \ldots, p\}$ as follows:

$$\begin{aligned} s_{nj}^{(1)}(\beta_0, t) &= E[Z_1^j \exp(\beta_0^\top Z_1)\pi_Z(t)] \\ &= \begin{cases} E\left[Z_1^j \prod_{l=1}^{S} \exp(\beta_0^\top Z_1)\pi_Z(t)\right] & (1 \leq j \leq S) \\ E[Z_1^j] s_n^{(0)}(\beta_0, t) & (S < j \leq p); \end{cases} \end{aligned}$$

$$\begin{aligned} s_{njk}^{(2)}(\beta_0, t) &= E[Z_1^j Z_1^k \exp(\beta_0^\top Z_1)\pi_Z(t)] \\ &= \begin{cases} E\left[Z_1^j Z_1^k \prod_{l=1}^{S} \exp(\beta_0^\top Z_1)\pi_Z(t)\right] & (1 \leq j, k \leq p) \\ E[Z_1^k] s_{nj}^{(1)}(\beta_0, t) & (1 \leq j \leq p,\ S < k \leq p) \\ E[Z_1^j] E[Z_1^k] s_n^{(0)}(\beta_0, t) & (S < j, k \leq p). \end{cases} \end{aligned}$$

We therefore obtain the matrix $I_n(\beta_0)$ as follows:

$$I_n(\beta_0) = \begin{pmatrix} \mathcal{I}(\beta_0) & 0 \\ 0 & \mathrm{diag}(Var[Z_1^j] \int_0^1 s_n^{(0)}(\beta_0, t)\lambda_0(t)dt) \end{pmatrix},$$

where $\mathcal{I}(\beta_0)$ is a $S \times S$ matrix which can be proved to be a positive definite (see e.g. Fleming and Harrington (1991)). Therefore, we have that the model satisfies Assumption 3.3.

3.6.2 Summary

In summary, we have been able to construct the asymptotically normal estimators 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 3.8 which is obtained from l_1 consistency proved in Theorem 3.7. If the sparsity is not fixed, we may not reduce the dimension of the parameter since we cannot prove Theorem 3.8. In such cases, the asymptotically normal estimators cannot be constructed by (3.3) and (3.4).

It is well known that 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) and this thesis 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 are equivalent to the irrepresentable conditions, which are obtained from KKT conditions of the optimization problems. Since the KKT conditions of 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 becomes 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 Lasso. In contrast, we have proved that l_1 consistency implies the variable selection consistency when the sparsity S is fixed in this paper. This type of theoretical results for various regression models may be proved for Lasso type estimators because the l_1 consistency results are nearly equivalent to that for the Dantzig selector.

Chapter 4

Diffusion processes with covariates

The purpose of this chapter is to discuss a parametric estimation problem in a high-dimensional and sparse setting for a special parametric model of diffusion processes. We consider the stochastic process $\{X_t\}_{t\geq 0}$ which is a solution to the stochastic differential equation given by

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \exp(\theta^\top Z_s)dW_s, \tag{4.1}$$

where $\{W_t\}_{t\geq 0}$ is a standard Brownian motion, $b(\cdot)$ is a nuisance drift function, $\{Z_t\}_{t\geq 0} = \{(Z_t^1, Z_t^2, \ldots, Z_t^p)\}_{t\geq 0}$ is a uniformly bounded p dimensional continuous process, which is regarded as a covariate vector, and θ is an unknown parameter of interest. We observe the processes $\{X_t\}_{t\geq 0}$ and $\{Z_t\}_{t\geq 0}$ at $n+1$ equidistant time points $0 =: t_0^n < t_1^n < \ldots < t_n^n$, where $t_k^n = kt_n^n/n$ for $k = 0, 1, \ldots, n$. Assume that $p = p_n \gg n$ and the number of non-zero components S in the true value θ_0 of θ is relatively small. In this high-dimensional and sparse setting, we consider the estimation problem of θ_0. The covariate processes $\{Z_t^i\}_{t\geq 0}$, $i = 1, 2, \ldots, p_n$, are, for example, some functionals $\{\phi_i(X_t^i)\}_{t\geq 0}$ of solutions to other stochastic differential equations $\{X_t^i\}_{t\geq 0}$, where ϕ_i's are uniformly bounded smooth functions or random variables which do not depend on t.

This chapter is organized as follows. The settings for the model, some regularity conditions, and the estimation procedure are given in Section 4.1. In Section 4.2, we state our l_q consistency results for every $q \in [1, \infty]$. Our methods of proofs are similar to Huang et al. (2013) who proved the consistency of Lasso estimator for Cox's proportional hazards model and to Chapter 3 of this thesis which dealt with the Dantzig selector for the proportional hazards model. Based on the l_1 consistency result, we discuss the variable selection consistency of the Dantzig selector in Section 4.3. Moreover, we construct an asymptotically normal estimator for θ_0 in Section 4.4

under the ergodic assumption on the covariate process $\{Z_t\}_{t\geq 0}$. Finally, in Section 4.5, we will present some concluding remarks concerning with the rate of convergence appearing in our main results.

4.1 Model set up and matrix conditions

Let $\{W_t\}_{t\geq 0}$ be a standard Brownian motion defined on a probability space $(\Omega, \mathcal{F}, P)$, and $\{Z_t\}_{t\geq 0} := \{(Z_t^1, Z_t^2, \ldots, Z_t^p)\}_{t\geq 0}$ be a uniformly bounded p dimensional continuous process. We introduce the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ defined by

$$\mathcal{F}_t := \mathcal{F}_0 \vee \sigma(W_s, Z_s : s \in [0,t]), \quad t \geq 0,$$

where $\mathcal{F}_0$ is a σ-field independent of $\{W_t\}_{t\geq 0}$, and $\{Z_t\}_{t\geq 0}$. Let us consider the 1 dimensional stochastic differential equation (5.1):

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \exp(\theta^\top Z_s)dW_s,$$

where $x \mapsto b(x)$ is a nuisance drift function which satisfies appropriate regularity conditions presented later, and $\theta \in \mathbb{R}^p$ is an unknown parameter of interest. We observe the process $\{X_t\}_{t\geq 0}$ at $n+1$ discrete time points $0 =: t_0^n < t_1^n < t_2^n < \cdots < t_n^n$, where $t_k^n := n^{-1}kt_n^n$. Assume that $p = p_n \gg n$ and the number of non-zero components S in the true value θ_0 is a fixed constant. In this high dimensional and sparse setting, we consider the estimation problem of θ_0 with finite l_1 norm. The quasi-likelihood function $L_n(b;\theta)$ is given by

$$L_n(b;\theta) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi \exp(2\theta^T Z_{t_{k-1}^n})\Delta_n}} \exp\left(-\frac{|X_{t_k^n} - X_{t_{k-1}^n} - b(X_{t_{k-1}^n})\Delta_n|^2}{2\exp(2\theta^T Z_{t_{k-1}^n})\Delta_n}\right),$$

where $\Delta_n := t_k^n - t_{k-1}^n = t_n^n/n$. Put $l_n(b;\theta) := \log L_n(b;\theta)$, and define the $\mathbb{R}^{p_n}$-valued function $\psi_n(b;\theta) = (\psi_n^1(b;\theta), \psi_n^2(b;\theta), \ldots, \psi_n^{p_n}(b;\theta))$ by

$$\begin{aligned}
\psi_n(b;\theta) &:= \frac{1}{n}\dot{l}_n(b;\theta) \\
&= \frac{1}{n\Delta_n}\sum_{k=1}^n Z_{t_{k-1}^n} \exp(-2\theta^T Z_{t_{k-1}^n})|X_{t_k^n} - X_{t_{k-1}^n} - b(X_{t_{k-1}^n})\Delta_n|^2 \\
&\qquad - Z_{t_{k-1}^n}\Delta_n.
\end{aligned}$$

Moreover, we define the $p_n \times p_n$ matrix-valued function $V_n(b;\theta)$ by

$$\begin{aligned} V_n(b;\theta) &:= -\frac{1}{n}\ddot{l}_n(b;\theta) \\ &= \frac{2}{n\Delta_n}\sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^{\top} \exp(-2\theta^T Z_{t_{k-1}^n})|X_{t_k^n} - X_{t_{k-1}^n} - b(X_{t_{k-1}^n})\Delta_n|^2. \end{aligned}$$

Note that $V_n(b;\theta)$ is a nonnegative definite matrix. Hereafter, we assume the following conditions.

Assumption 4.1. *(i) The terminal value t_n^n of the time interval $[0, t_n^n]$ satisfies the either condition;*

Case 1. *It holds that t_n^n is a fixed constant. Without loss of generality, we put $t_n^n = 1$.*

Case 2. *It holds that*

$$t_n^n = n\Delta_n \to \infty, \quad n\Delta_n^2 \to 0$$

as $n \to \infty$.

(ii) There exists a constant $\tilde{L} > 0$ such that for all $x, y \in \mathbb{R}$,

$$|b(x) - b(y)| \le \tilde{L}|x - y|.$$

(iii) There exists a constant $C > 0$ such that

$$\sup_{t\in[0,\infty)} \sup_{1\le i\le\infty} |Z_t^i| \le C \quad a.s.$$

(iv) For every $r \ge 1$, it holds that

$$\sup_{t\in[0,\infty)} E\left[|X_t|^r\right] < \infty.$$

(v) For every $r \in \mathbb{N}$, there exists a constant $\tilde{C}_r$ such that for every $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, p_n\}$ and $k = 1, 2, \ldots, n$,

$$E\left[\sup_{s\in[t_{k-1}^n, t_k^n]} |X_s - X_{t_{k-1}^n}|^r\right] \le \tilde{C}_r \Delta_n^{\frac{r}{2}},$$

$$E\left[\sup_{s\in[t_{k-1}^n, t_k^n]} |Z_s^i - Z_{t_{k-1}^n}^i|^r\right] \le \tilde{C}_r \Delta_n^{\frac{r}{2}}.$$

Assumption (iv) is satisfied if Z_t^i, $i = 1, 2, \ldots, p_n$, are appropriate transformation of stochastic processes which are solutions to other SDEs as mentioned in Introduction. In Section 4.2, we will show that $b(\cdot)$ can be ignored under Assumption 4.1. We thus define the estimator $\hat{\theta}_n$ by the Dantzig selector as

$$\hat{\theta}_n := \arg\min_{\theta \in \mathcal{C}_n} \|\theta\|_1, \quad \mathcal{C}_n := \{\theta \in \mathbb{R}^{p_n} : \|\psi_n(0;\theta)\|_\infty \leq \gamma\},$$

where γ is a tuning parameter by setting $b = 0$.

Define the $p_n \times p_n$ matrix J_n by

$$J_n := \frac{2}{n} \sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top,$$

which will be proved to approximate $V_n(0;\theta_0)$ in Section 4.2. We introduce the following factors (A), (B) and (C) in order to prove the consistency of the estimator $\hat{\theta}_n$.

Definition 4.2. *For every index set $T \subset \{1,\ 2,\ \cdots,\ p_n\}$ and $h \in \mathbb{R}^{p_n}$, h_T is a $\mathbb{R}^{|T|}$ dimensional sub-vector of h constructed by extracting the components of h corresponding to the indices in T. Define the set C_T by*

$$C_T := \{h \in \mathbb{R}^{p_n} : \|h_{T^c}\|_1 \leq \|h_T\|_1\}.$$

We introduce the following factors.

(A) **Compatibility factor**

$$\kappa(T_0; J_n) := \inf_{0 \neq h \in C_{T_0}} \frac{S^{\frac{1}{2}} (h^T J_n h)^{\frac{1}{2}}}{\|h_{T_0}\|_1}.$$

(B) **Weak cone invertibility factor**

$$F_q(T_0; J_n) := \inf_{0 \neq h \in C_{T_0}} \frac{S^{\frac{1}{q}} h^T J_n h}{\|h_{T_0}\|_1 \|h\|_q}, \quad q \in [1, \infty),$$

$$F_\infty(T_0; J_n) := \inf_{0 \neq h \in C_{T_0}} \frac{(h^T J_n h)^{\frac{1}{2}}}{\|h\|_\infty}.$$

(C) **Restricted eigenvalue**

$$RE(T_0; J_n) := \inf_{0 \neq h \in C_{T_0}} \frac{(h^T J_n h)^{\frac{1}{2}}}{\|h\|_2}.$$

We assume the next condition to derive our main results.

Assumption 4.3. *For every $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$*

$$P(\kappa(T_0; J_n) > \delta) \geq 1 - \epsilon.$$

Noting that $\|h_{T_0}\|_1^q \geq \|h_{T_0}\|_q^q$ for all $q \geq 1$, we can see that $\kappa(T_0; J_n) \leq 2\sqrt{S}RE(T_0; J_n)$, and $\kappa(T_0; J_n) \leq F_q(T_0; J_n)$. So under Assumption 4.3, $RE(T_0; J_n)$ and $F_q(T_0; J_n)$ also satisfy the corresponding conditions.

4.2 The $l_q$consistency of the Dantzig selector

In this section, we will prove the l_q consistency of the estimator $\hat{\theta}_n$. For Case 1, *i.e.*, the case where $t_n^n = 1$, the consistency result can be seen in Fujimori and Nishiyama (2017b). In addition, we can prove the consistency in Case 2 by the same way as that in Case 1. To do this, we will evaluate the gradient of quasi-likelihood at the true value and show that $V_n(0; \theta_0)$ is approximated by J_n. The following theorems are our main results. Hereafter, we assume that γ_n and p_n satisfy that

$$\gamma_n = K_0 \Delta_n^{\frac{1}{2}-\alpha}, \tag{4.2}$$

$$\log(1 + p_n) = O(n^{\zeta}), \tag{4.3}$$

where $K_0 > 0$, $0 < \alpha < 1/2$, $0 < \zeta < 2\alpha$ are some constants.

First of all, we will show that under Assumption 4.1,

$$\lim_{n\to\infty} P(\|\psi_n(0; \theta_0)\|_\infty \geq \gamma_n) = 0.$$

Let us decompose $\psi_n^i(0; \theta_0) = A_n^i + B_n^i + C_n^i$, where

$$A_n^i := \frac{1}{n\Delta_n} \sum_{k=1}^{n} Z_{t_{k-1}^n}^i \exp(-2\theta_0^T Z_{t_{k-1}^n}) \left| \int_{t_{k-1}^n}^{t_k^n} b(X_s) ds \right|^2,$$

$$\begin{aligned} B_n^i := \frac{2}{n\Delta_n} \sum_{k=1}^{n} Z_{t_{k-1}^n}^i \exp(-2\theta_0^T Z_{t_{k-1}^n}) \left(\int_{t_{k-1}^n}^{t_k^n} b(X_s) ds \right) \\ \times \left(\int_{t_{k-1}^n}^{t_k^n} \exp(\theta_0^T Z_s) dW_s \right) \end{aligned}$$

and

$$C_n^i := \frac{1}{n\Delta_n}\sum_{k=1}^{n} Z^i_{t^n_{k-1}} \exp(-2\theta_0^T Z_{t^n_{k-1}}) \left| \int_{t^n_{k-1}}^{t^n_k} \exp(\theta_0^T Z_s) dW_s \right|^2 - Z^i_{t^n_{k-1}} \Delta_n.$$

We further decompose $C_n^i = D_n^i + E_n^i$, where

$$\begin{aligned} D_n^i \;:=\; & \frac{1}{n\Delta_n}\sum_{k=1}^{n} Z^i_{t^n_{k-1}} \exp(-2\theta_0^T Z_{t^n_{k-1}}) \left| \int_{t^n_{k-1}}^{t^n_k} \exp(\theta_0^T Z_s) dW_s \right|^2 \\ & - Z^i_{t^n_{k-1}} (W_{t^n_k} - W_{t^n_{k-1}})^2 \end{aligned}$$

and

$$E_n^i := \frac{1}{n\Delta_n}\sum_{k=1}^{n} Z^i_{t^n_{k-1}} \{(W_{t^n_k} - W_{t^n_{k-1}})^2 - \Delta_n\}.$$

Lemma 4.4. *Suppose that γ_n satisfies (4.2). Under Assumption 4.1, it holds in both of the Cases 1 and 2 that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} |A_n^i| \ge \gamma_n \right) = 0.$$

Proof. It follows from Markov's inequality and Schwartz's inequality and Assumption 4.1 that

$$\begin{aligned} P\left(\sup_{1\le i\le p_n} |A_n^i| \ge \gamma_n \right) &\le \frac{C\exp(2C\|\theta_0\|_1)}{n\Delta_n\gamma_n} \sum_{k=1}^{n} E\left[\left| \int_{t^n_{k-1}}^{t^n_k} b(X_s) ds \right|^2 \right] \\ &\le \frac{C\exp(2C\|\theta_0\|_1)}{n\Delta_n\gamma_n} \sum_{k=1}^{n} E\left[\Delta_n \int_{t^n_{k-1}}^{t^n_k} |b(X_s)|^2 ds \right] \\ &\le \frac{C\exp(2C\|\theta_0\|_1)}{n\gamma_n} \sum_{k=1}^{n} \int_{t^n_{k-1}}^{t^n_k} E[|b(X_s)|^2] ds \\ &\le \frac{C\exp(2C\|\theta_0\|_1)}{\gamma_n} \tilde{C}\Delta_n. \end{aligned}$$

Noting that $\Delta_n \to 0$ and $\gamma_n = K_0\Delta_n^{\frac{1}{2}-\alpha}$, we obtain the conclusion. □

Lemma 4.5. *Under Assumption 4.1, it holds for both of the Cases 1 and 2 that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} |B_n^i| \ge \gamma_n\right) = 0.$$

Proof. Using Markov's inequality and Schwartz's inequality, we have that

$$\begin{aligned}
&P\left(\sup_{1\le i\le p_n} |B_n^i| \ge \gamma_n\right) \\
&\le \frac{2C\exp(2C\|\theta_0\|_1)}{n\Delta_n\gamma_n}\sum_{k=1}^{n}\left(E\left[\left|\int_{t_{k-1}^n}^{t_k^n} b(X_s)ds\right|^2\right]\right)^{\frac{1}{2}} \\
&\qquad\qquad \times\left(E\left[\left|\int_{t_{k-1}^n}^{t_k^n}\exp(\theta_0^T Z_s)dW_s\right|^2\right]\right)^{\frac{1}{2}} \\
&\le \frac{2C\exp(2C\|\theta_0\|_1)}{n\Delta_n\gamma_n}\sum_{k=1}^{n}\left(E\left[\Delta_n\int_{t_{k-1}^n}^{t_k^n}|b(X_s)|^2ds\right]\right)^{\frac{1}{2}} \\
&\qquad\qquad \times\left(E\left[\int_{t_{k-1}^n}^{t_k^n}\exp(2\theta_0^T Z_s)ds\right]\right)^{\frac{1}{2}} \\
&\le \frac{2C\exp(2C\|\theta_0\|_1)}{n\Delta_n\gamma_n} n\left(\tilde{C}\Delta_n^2\right)^{\frac{1}{2}}(\exp(2C\|\theta_0\|_1)\Delta_n)^{\frac{1}{2}} \\
&\le \frac{C\tilde{C}^{\frac{1}{2}}\Delta_n^{\frac{1}{2}}\exp(3C\|\theta_0\|_1)}{\gamma_n}.
\end{aligned}$$

The right-hand side of this inequality tends to 0 as $n\to\infty$. □

Lemma 4.4, and Lemma 4.5 imply that we can ignore the effect of $b(\cdot)$. So we may take $b(x)=0$ when we define the estimator $\hat{\theta}_n$. The following lemmas give some inequalities about D_n^i and E_n^i.

Lemma 4.6. *Under Assumption 4.1, it holds for both of the Cases 1 and 2 that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} |D_n^i| \ge \gamma_n\right) = 0.$$

Proof. It follows from Markov's inequality and Schwartz's inequality that

$$\begin{aligned} P\left(\sup_{1\le i\le p_n}|D_n^i|\ge\gamma_n\right) &\le \frac{C}{n\Delta_n\gamma_n}\sum_{k=1}^{n}E\left[|D_1|\cdot|D_2|\right] \\ &\le \frac{C}{n\Delta_n\gamma_n}\sum_{k=1}^{n}(E[|D_1|^2])^{\frac{1}{2}}(E[|D_2|^2])^{\frac{1}{2}}, \end{aligned}$$

where D_1 and D_2 are defined as follows

$$D_1 := \int_{t_{k-1}^n}^{t_k^n}\{\exp(\theta_0^T[Z_s - Z_{t_{k-1}^n}]) + 1\}dW_s,$$

$$D_2 := \int_{t_{k-1}^n}^{t_k^n}\{\exp(\theta_0^T[Z_s - Z_{t_{k-1}^n}]) - 1\}dW_s.$$

We can see that

$$\begin{aligned} (E[|D_1|^2])^{\frac{1}{2}} &= \left(E\left[\int_{t_{k-1}^n}^{t_k^n}\{\exp(\theta_0^T[Z_s - Z_{t_{k-1}^n}]) + 1\}^2 ds\right]\right)^{\frac{1}{2}} \\ &\le (\exp(2C\|\theta_0\|_1) + 1)\Delta_n^{\frac{1}{2}}. \end{aligned}$$

Noting that there exists a positive constant C_1 such that

$$\begin{aligned} |\exp(\theta_0^T[Z_s - Z_{t_{k-1}^n}]) - 1| &\le C_1|\theta_0^T[Z_s - Z_{t_{k-1}^n}]| \\ &\le C_1\|\theta_0\|_1 \max_{i\in T_0}|Z_s^i - Z_{t_{k-1}^n}^i|, \end{aligned}$$

where $T_0 := \{i : \theta_0^i \neq 0\}$, we have that

$$\begin{aligned} (E[|D_2|^2])^{\frac{1}{2}} &= \left(E\left[\int_{t_{k-1}^n}^{t_k^n}\{\exp(\theta_0^T[Z_s - Z_{t_{k-1}^n}]) - 1\}^2 ds\right]\right)^{\frac{1}{2}} \\ &\le \left(E\left[\int_{t_{k-1}^n}^{t_k^n} C_1^2\|\theta_0\|_1^2 \max_{i\in T_0}|Z_s^i - Z_{t_{k-1}^n}^i|^2 ds\right]\right)^{\frac{1}{2}} \\ &\le C_1\tilde{C}_2\|\theta_0\|_1\Delta_n. \end{aligned}$$

Consequently, it holds that

$$P\left(\sup_{1\le i\le p_n}|D_n^i|\ge\gamma_n\right) \le \frac{CC_1\tilde{C}_2\|\theta_0\|_1(\exp(2C\|\theta_0\|_1) + 1)\Delta_n^{\frac{1}{2}}}{\gamma_n} \to 0.$$

We thus obtain the conclusion. □

Lemma 4.7. *Suppose that γ_n and p_n satisfy (4.2) and (4.3) respectively. Under Assumption 4.1, it holds for both of the Cases 1 and 2 that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} |E_n^i| \ge 3\gamma_n\right) = 0.$$

Proof. Put $U_{t_k^n} := |W_{t_k^n} - W_{t_{k-1}^n}|^2 - \Delta_n$ and $\eta := \Delta_n^{1/2+\alpha-\beta}$, where $0 < \beta < 2\alpha - \zeta$ is a constant. Then, we have that

$$\begin{aligned} E_n^i &= \frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}} U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} + \frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}} U_{t_k^n} 1_{\{|U_{t_k^n}|>\eta\}} \\ &=: F_n^i + G_n^i. \end{aligned}$$

It is sufficient to prove that $P(\sup_i |F_n^i| \ge 2\gamma_n) \to 0$ and $P(\sup_i |G_n^i| \ge \gamma_n) \to 0$. Note that

$$\begin{aligned} F_n^i &= \frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}} \{U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} - E[U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} | \mathcal{F}_{t^n_{k-1}}]\} \\ &\qquad + Z^i_{t^n_{k-1}} E[U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} | \mathcal{F}_{t^n_{k-1}}] \\ &=: H_n^i + I_n^i. \end{aligned}$$

We can see that for all k and i,

$$|Z^i_{t^n_{k-1}} \{U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} - E[U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}}]\}| \le 2C\eta$$

$$E[|Z^i_{t^n_{k-1}}|^2 \{U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}} - E[U_{t_k^n} 1_{\{|U_{t_k^n}|\le\eta\}}] | \mathcal{F}_{t^n_{k-1}}]\}^2 | \mathcal{F}_{t^n_{k-1}}] \le C^2\Delta_n^2.$$

Now, it follows from Bernstein's inequality for martingales (See Theorem 1.6 from Freedman (1975).) that

$$P\left(|H_n^i| \ge \gamma_n\right) \le 2\exp\left(-\frac{\gamma_n^2}{2(2C\eta\gamma_n + C^2\Delta_n^2)}\right).$$

Write $\|\cdot\|_{\Phi_1}$ for Orlicz norm with respect to $\Phi_1(x) := e^x - 1$. Lemma 2.5 implies that there exists a constant $L > 0$ depending only on Φ_1 such that

$$\left\| \sup_{1\le i\le p_n} |H_n^i| \right\|_{\Phi_1} \le L\left\{2C\eta\log(1+p_n) + \sqrt{C^2\Delta_n^2\log(1+p_n)}\right\}.$$

Using Markov's inequality, we have that

$$\begin{aligned}
P\left(\sup_{1\le i\le p_n}|H_n^i|\ge\gamma_n\right) &= P\left(\Phi_1\left(\frac{\sup_i|H_n^i|}{\|\sup_i|H_n^i|\|_{\Phi_1}}\right)\ge\Phi_1\left(\frac{\gamma_n}{\|\sup_i|H_n^i|\|_{\Phi_1}}\right)\right)\\
&\le \Phi_1\left(\frac{\gamma_n}{\|\sup_i|H_n^i|\|_{\Phi_1}}\right)^{-1}\\
&\le \Phi_1\left(\frac{\gamma_n}{L\left\{2C\eta\log(1+p_n)+\sqrt{C^2\Delta_n^2\log(1+p_n)}\right\}}\right)^{-1}\\
&\to 0.
\end{aligned}$$

On the other hand, it holds that

$$\begin{aligned}
I_n^i &= \frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}}\left\{E[U_{t^n_k}-U_{t^n_k}1_{\{|U_{t^n_k}|>\eta\}}|\mathcal{F}_{t^n_{k-1}}]\right\}\\
&= \frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}}E[-U_{t^n_k}1_{\{|U_{t^n_k}|>\eta\}}|\mathcal{F}_{t^n_{k-1}}].
\end{aligned}$$

So we thus obtain that

$$\begin{aligned}
P\left(\sup_{1\le i\le p_n}|I_n^i|\ge\gamma_n\right) &\le \frac{1}{\gamma_n}E\left[\sup_{1\le i\le p_n}\left|\frac{1}{n\Delta_n}\sum_{k=1}^n Z^i_{t^n_{k-1}}E[U_{t^n_k}1_{\{|U_{t^n_k}|>\eta\}}|\mathcal{F}_{t^n_{k-1}}]\right|\right]\\
&\le \frac{C}{n\Delta_n\gamma_n}\sum_{k=1}^n E\left[E\left[\frac{|U_{t^n_k}|^2}{\eta}|\mathcal{F}_{t^n_{k-1}}\right]\right]\\
&= \frac{2C\Delta_n}{\gamma_n\eta}\\
&\to 0.
\end{aligned}$$

A similar calculation leads us that

$$P\left(\sup_{1\le i\le p_n}|G_n^i|\ge\gamma_n\right)\to 0.$$

This yields the conclusion. □

After all, we obtain the next lemma.

Lemma 4.8. *Suppose that γ_n and p_n satisfy (4.2) and (4.3) respectively. Under Assumption 4.1, it holds for both of the Cases 1 and 2 that*

$$\lim_{n\to\infty}P\left(\|\psi_n(0;\theta_0)\|_\infty\ge 6\gamma_n\right)=0.$$

This lemma states that the true value θ_0 belongs to the constraint set $\mathcal{C}_n$ with large probability when the sample size n is large. Then, we prepare two lemmas for $V_n(0;\theta_0)$. The next lemma states that $V_n(0;\theta_0)$ is approximated by J_n.

Lemma 4.9. *The random sequence ϵ_n defined by*

$$\epsilon_n := \|V_n(0;\theta_0) - J_n\|_\infty$$

converges in probability to 0.

Proof. It holds that

$$\begin{aligned} V_n(0;\theta_0) &= \frac{2}{n\Delta_n}\sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top \exp(-2\theta_0^T Z_{t_{k-1}^n}) \left| \int_{t_{k-1}^n}^{t_k^n} \exp(\theta_0^T Z_s) dW_s \right|^2 \\ &= (I) + (II) + (III), \end{aligned}$$

where

$$\begin{aligned} (I) \;=\; & \frac{2}{n\Delta_n}\sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top \left| \int_{t_{k-1}^n}^{t_k^n} \exp(\theta_0^T [Z_s - Z_{t_{k-1}^n}]) dW_s \right|^2 \\ & - Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top |W_{t_k^n} - W_{t_{k-1}^n}|^2, \end{aligned}$$

$$(II) = \frac{2}{n\Delta_n}\sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top \left\{ |W_{t_k^n} - W_{t_{k-1}^n}|^2 - \Delta_n \right\},$$

and

$$(III) = \frac{2}{n\Delta_n}\sum_{k=1}^{n} Z_{t_{k-1}^n} Z_{t_{k-1}^n}^\top \Delta_n = J_n.$$

Using triangle inequality, we have that

$$\|V_n(0;\theta_0) - J_n\|_\infty \leq \|(I)\|_\infty + \|(II)\|_\infty.$$

As well as the proof of Lemma 4.6 and Lemma 4.7, we can prove that $\|(I)\|_\infty$ and $\|(II)\|_\infty$ are $o_p(1)$. □

The relationship between $\psi_n(0;\hat{\theta}_n) - \psi_n(0;\theta_0)$ and $V_n(0;\theta_0)$ are provided by the lemma below.

Lemma 4.10. *Define that* $I := [-2C\|\theta_0\|_1, 2C\|\theta_0\|_1]$,

$$g(x) := \begin{cases} \frac{e^{2x}-1}{x} & (x \neq 0) \\ 2 & (x = 0) \end{cases}$$

and $\nu := \min_{x \in I} g(x)$. *Then, it holds for* $h := \theta_0 - \hat{\theta}_n$ *that*

$$\frac{\nu}{2} h^T V_n(0; \theta_0) h \leq h^T [\psi_n(0; \hat{\theta}_n) - \psi_n(0; \theta_0)].$$

Proof. We have that

$$\begin{aligned} h^T [\psi_n(0; \hat{\theta}_n) - \psi_n(0; \theta_0)] &= \frac{1}{n\Delta_n} \sum_{k=1}^{n} h^T Z_{t^n_{k-1}} \exp(-2\theta_0^T Z_{t^n_{k-1}}) |X_{t^n_k} - X_{t^n_{k-1}}|^2 \\ &\qquad \times \{\exp(2h^T Z_{t^n_{k-1}}) - 1\} \end{aligned}$$

Note that $h^T Z_{t^n_{k-1}} \in I$ for all $k = 1, 2, \ldots, n$. Noting moreover that $x(e^{2x} - 1) \geq \nu x^2$, we can see that

$$\begin{aligned} h^T [\psi_n(0; \hat{\theta}_n) - \psi_n(0; \theta_0)] &\geq \frac{1}{n\Delta_n} \sum_{k=1}^{n} \exp(-2\theta_0^T Z_{t^n_{k-1}}) |X_{t^n_k} - X_{t^n_{k-1}}|^2 (\nu h^T Z_{t^n_{k-1}})^2 \\ &= \frac{\nu}{2} h^T V_n(0; \theta_0) h. \end{aligned}$$

We thus obtain the conclusion. □

Now, we are ready to prove our main results. The next theorem states the l_q consistency of the estimator $\hat{\theta}_n$.

Theorem 4.11. *Suppose that* γ_n *and* p_n *satisfy (4.2) and (4.3) respectively. Under Assumptions 4.1 and 4.3, the following (i)-(iv) hold true for some positive constants* K_2 *and* K_3 *in both of Cases 1 and 2.*

(*i*) *It holds that*

$$\lim_{n \to \infty} P\left(\|\hat{\theta}_n - \theta_0\|_2^2 \geq \frac{K_2 \gamma_n + K_3 \epsilon_n}{R E^2(T_0; J_n)} \right) = 0.$$

In particular, it holds that $\|\hat{\theta}_n - \theta_0\|_2 \to^p 0$.

(*ii*) *It holds that*

$$\lim_{n \to \infty} P\left(\|\hat{\theta}_n - \theta_0\|_\infty^2 \geq \frac{K_2 \gamma_n + K_3 \epsilon_n}{F_\infty^2(T_0; J_n)} \right) = 0.$$

In particular, it holds that $\|\hat{\theta}_n - \theta_0\|_\infty \to^p 0$.

(*iii*) *It holds that*

$$\lim_{n\to\infty} P\left(\|\hat{\theta}_n - \theta_0\|_1 \geq \frac{4K_4 S\gamma_n}{\kappa^2(T_0; J_n) - 4S\epsilon_n}\right) = 0.$$

In particular, it holds that $\|\hat{\theta}_n - \theta_0\|_1 \to^p 0$.

(*iv*) *It holds for every* $q \in (1, \infty)$ *that*

$$\lim_{n\to\infty} P\left(\|\hat{\theta}_n - \theta_0\|_q \geq \xi_{n,q}\right) = 0,$$

where

$$\xi_{n,q} := \frac{2S^{\frac{1}{q}}\epsilon_n}{F_q(T_0; J_n)} \cdot \frac{2K_4 S\gamma_n}{\kappa^2(T_0; J_n) - 2S\epsilon_n} + \frac{2K_4 S^{\frac{1}{q}}\gamma_n}{F_q(T_0; J_n)}.$$

In particular, it holds for all $q \in (1, \infty)$ *that* $\|\hat{\theta}_n - \theta_0\|_q \to^p 0$.

Proof. It is sufficient to prove that $\|\psi_n(0; \theta_0)\|_\infty \leq \gamma_n$ implies that

$$\|\hat{\theta}_n - \theta_0\|_2^2 \leq \frac{K_2\gamma_n + K_3\epsilon_n}{RE^2(T_0; J_n)}.$$

By the construction of the estimator $\hat{\theta}_n$, we have $\|\psi_n(0; \hat{\theta}_n)\|_\infty \leq \gamma_n$, which implies that

$$\|\psi_n(0; \hat{\theta}_n) - \psi_n(0; \theta_0)\|_\infty \leq \|\psi_n(0; \hat{\theta}_n)\|_\infty + \|\psi_n(0; \theta_0)\|_\infty \leq 2\gamma_n.$$

Put $h := \theta_0 - \hat{\theta}_n$, then we have that $h \in C_{T_0}$ since it holds that

$$\begin{aligned} 0 \geq \|\theta_0 - h\|_1 - \|\theta_0\|_1 &= \sum_{j\in T_0^c} |h_{T_{0j}^c}| + \sum_{j\in T_0} (|\theta_{0j} - h_{T_{0j}}| - |\theta_{0j}|) \\ &\geq \sum_{j\in T_0^c} |h_{T_{0j}^c}| - \sum_{j\in T_0} |h_{T_{0j}}| \\ &= \|h_{T_0^c}\|_1 - \|h_{T_0}\|_1. \end{aligned}$$

Notice moreover that $\|h\|_1 \leq \|\hat{\theta}_n\|_1 + \|\theta_0\|_1 \leq 2\|\theta_0\|_1$ by the definition of $\hat{\theta}_n$. Now, we use Lemma 4.10 for h to deduce that

$$\begin{aligned} h^T V_n(0; \theta_0) h &\leq \frac{2}{\nu} h^T [\psi_n(0; \hat{\theta}_n) - \psi_n(0; \theta_0)] \\ &\leq \frac{4}{\nu}\gamma_n \|h\|_1 \\ &\leq \frac{8}{\nu}\gamma_n \|\theta_0\|_1 \\ &=: K_2\gamma_n. \end{aligned}$$

Thus it holds that

$$\begin{aligned} h^T J_n h &\leq |h^T (J_n - V_n(0;\theta_0))h| + h^T V_n(0;\theta_0)h \\ &\leq \epsilon_n \|h\|_1^2 + K_2\gamma_n \\ &\leq \epsilon_n \cdot 4\|\theta_0\|_1^2 + K_2\gamma_n \\ &=: K_3\epsilon_n + K_2\gamma_n. \end{aligned}$$

By the definition of the restricted eigenvalue, we have that

$$\begin{aligned} RE^2(T_0; J_n) &\leq \frac{h^T J_n h}{\|\hat{\theta}_n - \theta_0\|_2^2} \\ &\leq \frac{K_2\gamma_n + K_3\epsilon_n}{\|\hat{\theta}_n - \theta_0\|_2^2}. \end{aligned}$$

Noting that $RE^2(T_0; J_n) > 0$ with large probability when n is sufficiently large, we obtain that

$$\|\hat{\theta}_n - \theta_0\|_2^2 \leq \frac{K_2\gamma_n + K_3\epsilon_n}{RE^2(T_0; J_n)},$$

which yields the conclusion in (i). Using the factor $F_\infty(T_0; J_n)$, we obtain the conclusion in (ii) by the similar way.

It follows from the proof of (i) that

$$h^T V_n(0;\theta_0)h \leq K_4\gamma_n \|\hat{\theta}_n - \theta_0\|_1.$$

Noting that $\|b\|_2^2 \leq \|b\|_1^2$ for all $b \in \mathbb{R}^{p_n}$, we have that

$$h^T J_n h \leq \epsilon_n \|\hat{\theta}_n - \theta_0\|_1^2 + K_4\gamma_n \|\hat{\theta}_n - \theta_0\|_1.$$

The definition of $\kappa(T_0; J_n)$ implies that

$$\begin{aligned} \kappa^2(T_0; J_n) &\leq \frac{S h^T J_n h}{\|h_{T_0}\|_1^2} \\ &\leq \frac{S\epsilon_n \|h\|_1^2 + K_4 S\gamma_n \|h\|_1}{\|h_{T_0}\|_1^2}. \end{aligned}$$

Since $\|h\|_1 \leq 2\|h_{T_0}\|_1$, this yields the conclusion in (iii).

On the other hand, using the weak cone invertibility factor for every $q \geq 1$, we have that

$$F_q(T_0; J_n) \leq \frac{S^{\frac{1}{q}}\epsilon_n \|h\|_1^2 + S^{\frac{1}{q}} K_4\gamma_n \|h\|_1}{\|h_{T_0}\|_1 \|h\|_q},$$

which implies that

$$\|\hat{\theta}_n - \theta_0\|_q \leq \frac{2S^{\frac{1}{q}}\epsilon_n\|\hat{\theta}_n - \theta_0\|_1 + 2S^{\frac{1}{q}}K_4\gamma_n}{F_q(T_0; J_n)}.$$

Using the l_1 bound derived above, we obtain the conclusion in (iv). □

4.3 The variable selection consistency of the Dantzig selector

As in the previous chapters, we construct the estimator $\hat{T}_n$ for the support index set T_0 by

$$\hat{T}_n = \{j : |\hat{\theta}_n| > \gamma_n\}.$$

The variable section consistency can be derived from Theorem 4.11 as follows.

Theorem 4.12. *Under Assumptions 4.1 and 4.3, it holds that*

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

Proof. Note that $\|\hat{\theta}_n - \theta_0\|_\infty \leq \|\hat{\theta}_n - \theta_0\|_1$ and that the sparsity S is assumed to be fixed. We have that

$$\lim_{n\to\infty} P\left(\|\hat{\theta}_n - \theta_0\|_\infty > \gamma_n\right) = 0$$

by the l_1 bound from Theorem 4.11 (iii). Therefore, it is sufficient to show that the next inequality

$$\|\hat{\theta}_n - \theta_0\|_\infty \leq \gamma_n$$

implies that

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

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

$$|\theta_0^j| - |\hat{\theta}_n^j| \leq |\hat{\theta}_n^j - \theta_0^j| \leq \gamma_n.$$

We have that

$$|\hat{\theta}_n^j| \geq |\theta_0^j| - \gamma_{n,p_n} > \gamma_n$$

for sufficiently large n, which implies that $T_0 \subset \hat{T}_n$. On the other hand, for every $j \in T_0^c$, we have that

$$|\hat{\theta}_n^j - \theta_0^j| = |\hat{\theta}_n^j| \leq \gamma_n$$

since it holds that $\theta_0^j = 0$. From this fact, we can see that $j \in \hat{T}_n^c$ which implies that $\hat{T}_n \subset T_0$. We thus obtain the conclusion. □

4.4 After variable selection

In this section, we deal with Case 2, *i.e.*, the case where $t_n^n \to \infty$. Hereafter, we assume the following ergodic condition.

Assumption 4.13. *The sub-vector of the covariate process $\{Z_{tT_0}\}_{t\geq 0}$ is ergodic with an invariant measure μ_0, i.e., for every μ_0-integrable function g, it holds that*

$$\frac{1}{T}\int_0^T g(Z_{tT_0})dt \to^p \int_{\mathbb{R}^S} g(z)\mu_0(dz)$$

as $T \to \infty$.

The next lemma is immediately derived from Assumption 4.13.

Lemma 4.14. *Under Assumptions 4.1 and 4.13, it holds that*

$$\|J_{nT_0,T_0} - \mathcal{I}\|_\infty \to^p 0, \quad n \to \infty,$$

where

$$\mathcal{I} = 2\int_{\mathbb{R}^S} zz^\top \mu_0(dz).$$

We assume the non-singularity of matrix $\mathcal{I}$.

Assumption 4.15. *The $S \times S$ matrix $\mathcal{I}$ is positive definite.*

Using the estimator $\hat{T}_n$, we construct the new estimator $\hat{\theta}_n^{(2)}$ by the solution to the following equation:

$$\psi_n(\theta)_{\hat{T}_n} = 0, \quad \theta_{\hat{T}_n^c} = 0. \tag{4.4}$$

Moreover, define the parameter space Ξ_n by

$$\Xi_n := \{\theta \in \mathbb{R}^{p_n} : \theta_{T_0^c} = 0,\ \|\theta\|_q < \infty,\ \forall q \in [1,\infty]\}.$$

We present the l_2 consistency of the estimator $\hat{\theta}_n^{(2)}$ by the next theorem.

Theorem 4.16. *Under Assumptions 4.1, 4.3, 4.13 and 4.15, it holds for every $q \in [1,\infty]$ that*

$$\|\hat{\theta}_n^{(2)} - \theta_0\|_q \to^p 0, \quad n \to \infty.$$

Proof. We have that

$$\|\hat{\theta}_n^{(2)} - \theta_0\|_q \leq \|\hat{\theta}_{nT_0}^{(2)} - \theta_{0T_0}\|_q + \|\hat{\theta}_{nT_0^c}^{(2)}\|_q.$$

It follows from previous lemmas and ergodic property of $\{Z_{tT_0}\}_{t\geq 0}$ that

$$\begin{aligned}
\psi_n(\theta)_{T_0} &= \frac{1}{n}\sum_{k=1}^{n} Z_{t_{k-1}^n T_0} \exp\left(-2\theta_{T_0}^\top Z_{t_{k-1}^n T_0}\right) \\
&\quad \times \left\{\exp\left(2\theta_{0T_0}^\top Z_{t_{k-1}^n T_0}\right) - \exp\left(2\theta_{T_0}^\top Z_{t_{k-1}^n T_0}\right)\right\} + o_p(1) \\
&\to^p \int_{\mathbb{R}^S} z \exp\left(-2\theta_{T_0}^\top z\right)\left\{\exp\left(2\theta_{0T_0}^\top z - \exp\left(2\theta_{T_0}^\top z\right)\right)\right\}\mu_0(dz) \\
&=: \psi(\theta)_{T_0}
\end{aligned}$$

for every $\theta \in \Xi$ in the sense of l_q norm. This pointwisely convergence can be extended to the uniformly convergence over Ξ since we can easily check the condition that

$$\sup_{\theta\in\Xi_n}\left|\frac{\partial}{\partial\theta_j}\psi_n^i(\theta)_{T_0}\right| = O_p(1)$$

for every $i, j \in T_0$. In addition, we have that

$$\begin{aligned}
\psi_n(\hat{\theta}_n^{(2)})_{T_0} &= \psi_n(\hat{\theta}_n^{(2)})_{\hat{T}_n} 1_{\{\hat{T}_n = T_0\}} + \psi_n(\hat{\theta}_n^{(2)})_{T_0} 1_{\{\hat{T}_n \neq T_0\}} \\
&= 0 + o_p(1)
\end{aligned}$$

and

$$\psi(\theta_0) = 0.$$

We therefore obtain that

$$\|\hat{\theta}_{nT_0}^{(2)} - \theta_{0T_0}\|_q \to^p 0, \quad n \to \infty.$$

It is obvious that $\|\hat{\theta}_{nT_0^c}^{(2)}\|_q = o_p(1)$ since we have that

$$\|\hat{\theta}_{nT_0^c}^{(2)}\|_q = \|\hat{\theta}_{nT_n^c}^{(2)}\|_q 1_{\{\hat{T}_n = T_0\}} + \|\hat{\theta}_{nT_0^c}^{(2)}\|_q 1_{\{\hat{T}_n \neq T_0\}}.$$

□

Finally, we can derive the following asymptotic normality.

Theorem 4.17. *Under Assumptions 4.1, 4.3, 4.13 and 4.15, it holds that*

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

Proof. It follows from Taylor expansion and previous lemmas that

$$\sqrt{n}\psi_n(0;\theta_0)_{T_0}1_{\{\hat{T}_n=T_0\}} = V_n(0;\tilde{\theta}_n)_{T_0,T_0}\sqrt{n}\left(\hat{\theta}^{(2)}_{n\hat{T}_n}-\theta_{0T_0}\right)1_{\hat{T}_n=T_0},$$

where $\tilde{\theta}$ is a point between $\hat{\theta}^{(2)}_n$ and θ_0. By the similar way to that in the proof of Theorem 4.16, we have that

$$\begin{aligned} V_n(0;\theta)_{T_0,T_0} &= \frac{2}{n}\sum_{k=1}^{n} Z_{t^n_{k-1}T_0}Z^{\top}_{t^n_{k-1}T_0}\exp\left\{2\left(\theta_0-\theta\right)^{\top}_{T_0}Z_{t^n_{k-1}T_0}\right\}+o_p(1) \\ &\to^p \int_{\mathbb{R}^{\mathbb{S}}} zz^{\top}\exp\left\{2\left(\theta_0-\theta\right)^{\top}_{T_0}z\right\}\mu_0(dz) \\ &=: V(0;\theta) \end{aligned}$$

for every $\theta\in\Xi_n$ as $n\to\infty$ in the sense of l_∞ norm and this convergence can be extended to the uniform convergence over Ξ_n. Combining this fact with Lemmas 4.9, 4.14, and continuous mapping theorem, we have that

$$\|V_n(0;\tilde{\theta}_n)_{T_0,T_0}-\mathcal{I}\|_\infty = o_p(1).$$

since $\|\tilde{\theta}_n-\theta_0\|_q=o_p(1)$ by Theorem 4.16. We therefore have that

$$\sqrt{n}\left(\hat{\theta}^{(2)}_{n\hat{T}_n}-\theta_{0T_0}\right)1_{\{\hat{T}_n=T_0\}} = \mathcal{I}^{-1}\sqrt{n}\psi_n(0;\theta_0)_{T_0}1_{\{\hat{T}_n=T_0\}}+o_p(1).$$

Since $\psi_n(0;\theta_0)_{T_0}$ satisfies that

$$\sqrt{n}\psi_n(0;\theta_0)_{T_0} = \frac{1}{\sqrt{n}\Delta_n}\sum_{k=1}^{n} Z_{t^n_{k-1}T_0}\left\{\left(W_{t^n_k}-W_{t^n_{k-1}}\right)^2-\Delta_n\right\}+o_p(1)$$

and the main term of the right-hand side of the equality is the terminal value of square integrable martingale, we can apply the martingale central limit theorem to deduce that

$$\sqrt{n}\psi_n(0;\theta_0)_{T_0}\to^d N(0,\mathcal{I}),\quad n\to\infty.$$

Noting that $1_{\{\hat{T}_n=T_0\}}\to^p 1$ as $n\to\infty$, we obtain the conclusion by Slutsky's lemma. □

4.5 Concluding remarks

As in Theorem 4.11 (i) and (ii), we have that the rates of convergence of $\|\hat{\theta}_n-\theta_0\|_2$ and $\|\hat{\theta}_n-\theta_0\|_\infty$ are $n^{-\rho}$ for $\rho\in(0,1/4)$ in Case 1. In a low-dimensional

setting, it is known that the rate of convergence of the maximum quasi-likelihood estimator is $n^{-1/2}$. So we can see the influence of a high-dimensional setting in our results. Although the l_1 norm is greater than l_∞ norm, the l_1 and l_q risk bounds for finite q which we derived in Theorem 4.11 (iii) and (iv) tell us that the rates of convergence are $n^{-\rho'}$ for $\rho' \in (0, 1/2)$, which can be faster than l_2 and l_∞ bound. These phenomena are caused by the fact that the sparsity S of the true value is fixed constant and that the covariate process $\{Z_t\}_{t\geq 0}$ is assumed to be uniformly bounded. When the sparsity S depends on n and the covariate process $\{Z_t\}_{t\geq 0}$ is not bounded, we may not be able take the dimension p of the unknown parameter in exponential order of n and the rates of convergence of $\|\hat{\theta}_n - \theta_0\|_1$ and $\|\hat{\theta}_n - \theta_0\|_q$ for finite q would become slower than those we derived.

Chapter 5

A linear model of diffusion processes

Let us consider the following model given by the linear stochastic differential equation:

$$X_t = X_0 + \int_0^t \Theta^\top \phi(X_s) ds + \sigma W_t, \tag{5.1}$$

where $\{W_t\}_{t\geq 0} := \{(W_t^1, \ldots, W_t^p)\}_{t\geq 0}$ is a p-dimensional standard Brownian motion, Θ is a $p \times p$ sparse deterministic matrix, $\sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ is a $p \times p$ diagonal matrix and $\phi(x) = (\phi_1(x_1), \ldots, \phi_p(x_p))^\top$ for $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ is a smooth $\mathbb{R}^p$-valued function. We will propose some estimators for the true values (Θ^0, σ^0) of (Θ, σ) based on the observation of $\{X_t\}_{t\geq 0}$ at $n+1$ equidistant time points $0 =: t_0^n < t_1^n < \ldots < t_n^n$, under the high-dimensional and sparse setting, *i.e.*, $p \gg n$ and the number of nonzero components of the true value Θ^0 is relatively small.

The statistical inference for high-dimensional linear diffusion processes was especially discussed by some researchers. Periera and Ibrahimi (2014) studied various models of multi-dimensional diffusion processes observed continuously in high-dimensional settings including the following p-dimensional linear model:

$$X_t = X_0 + \int_0^t \Theta^\top X_s ds + W_t, \quad t \in [0, T]. \tag{5.2}$$

This model may be useful for various fields such as statistical physics, chemical reactions, finances and network systems. They proposed a Lasso type estimator for the true value Θ^0 of Θ and discussed the support recovery of the estimator when the dimension of the process p and the time interval T tends to ∞ independently. Similarly, Gaiffas and Matulewicz (2017) studied the drift estimation based on the

Lasso-type estimators for the high-dimensional Ornstein-Uhlenbeck processes described by the SDE like (5.2). They derived the oracle properties of the estimator for the sparse drift matrix and showed some applications for financial data. However, there are few previous researches dealing with the estimation problems for these linear models based on discrete observations.

In this chapter, we will apply the Dantzig selector to the linear models of stochastic processes (5.1), which is similar to the model (5.2), to estimate the drift matrix Θ^0 and prove the consistency in the sense of l_q norm for every $q \in [1, \infty]$ and the variable selection consistency under some appropriate conditions. Moreover, using the variable selection consistency, we will construct a new estimator which has an asymptotic normality. We can prove the consistency of the Dantzig selector by the similar way to Bickel et al. (2009). Note that our "high-dimensional setting' is different from that in previous researches such as Periera and Ibrahimi (2014) and Gaiffas and Matulewicz (2017); our estimation procedure is based on discrete observations and we assume that the dimension p depends on the number n of observed time points. We thus need to develop a different type of asymptotic theories.

This chapter is organized as follows. In Section 5.1, we will introduce our model setups and some regularity conditions. The construction of the estimator for the diffusion matrix and its consistency result are described in Section 5.2. The Dantzig selector for the drift matrix and the l_q consistency of the Dantzig selector are presented in Section 5.3. We will prove the variable selection consistency of the Dantzig selector in Section 5.4. Moreover, we will construct a new estimator by using the variable selection consistency and prove the asymptotic normality of the new estimator in the same section. Finally, some concluding remarks are described in Section 5.5.

5.1 Preliminaries

Let $\{W_t^1\}_{t\geq 0}, \{W_t^2\}_{t\geq 0}, \ldots$ be independent standard Brownian motions on a probability space $(\Omega, \mathcal{F}, P)$. Define the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ as follows.

$$\mathcal{F}_t := \mathcal{F}_0 \vee \sigma(W_s^j;\ j = 1, 2, \ldots,\ s \in [0, t]), \quad t \geq 0,$$

where $\mathcal{F}_0$ is a σ-field independent of $\{W_t^j\}_{t\geq 0}, j = 1, 2, \ldots$. We consider the following p-dimensional linear stochastic differential equation (5.1) defined on the stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$:

$$X_t = X_0 + \int_0^t \Theta^\top \phi(X_s) ds + \sigma W_t, \quad t \geq 0$$

where $\{W_t\}_{t\geq 0} := \{(W_t^1, \ldots, W_t^p)\}_{t\geq 0}$ is a p-dimensional standard Brownian motion, Θ is a $p \times p$ deterministic matrix, $\sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ is a $p \times p$ diagonal matrix, and $\phi(x) = (\phi_1(x_1), \ldots, \phi_p(x_p))^\top, x = (x_1, \ldots, x_p)^\top$ is a smooth $\mathbb{R}^p$-valued function. Assume that X_0 is $\mathcal{F}_0$-measurable. Note that $\{X_t^i\}_{t\geq 0}$ for each $i = 1, 2, \ldots, p$ satisfies the following equation.

$$X_t^i = X_0^i + \int_0^t \Theta_i^\top \phi(X_s) ds + \sigma_i W_t^i, \quad t \geq 0,$$

where Θ_i is the i-th row of the matrix Θ. In this paper, we consider the estimation problem of the true value (Θ^0, σ^0) of (Θ, σ). Suppose that we can observe the process $\{X_t\}_{t\geq 0}$ at $n+1$ discrete time points:

$$0 =: t_0^n < t_1^n < \ldots < t_n^n, \quad t_k^n = \frac{k t_n^n}{n}, \quad k = 0, 1, \ldots, n.$$

Write T_0^i for the support of the true value Θ_i^0 for every $i \in \{1, 2, \ldots, p\}$, *i.e.*, $T_0^i = \{j : \Theta_{ij}^0 \neq 0\}$. Let S_i be the number of elements in the index set T_0^i. Hereafter, we assume the following high-dimensional and sparse setting for the true matrix Θ^0.

$$p = p_n \gg n, \qquad \sup_{1\leq i<\infty} S_i =: S^* < \infty;$$

note that $S^* > 0$ is a constant which does not depend on n. We use the quasi-likelihood method which is commonly used in this field to estimate the unknown parameters. The quasi-likelihood function is constructed by discretization of the processes by Euler-Maruyama scheme, which is based on the fact that diffusion processes can be locally approximated by Gaussian random variables. See e.g. Yoshida (1992), Genon-Catalot and Jacod (1993) and Kessler (1997) for details. In this model, the quasi-log-likelihood function is given by

$$\sum_{k=1}^n \left\{ -\frac{1}{2}\log(2\pi\sigma_i^2\Delta_n) - \frac{|X_{t_k^n}^i - X_{t_{k-1}^n}^i - \Theta_i^T \phi(X_{t_{k-1}^n})\Delta_n|^2}{2\sigma_i^2\Delta_n} \right\},$$

where $\Delta_n := t_k^n - t_{k-1}^n = t_n^n/n$. We write $l_n(\Theta_i, \sigma_i)$ for the normalized quasi-log-likelihood, *i.e.*,

$$l_n(\Theta_i, \sigma_i) := \frac{1}{n\Delta_n} \sum_{k=1}^n \left\{ -\frac{1}{2}\log(2\pi\sigma_i^2\Delta_n) - \frac{|X_{t_k^n}^i - X_{t_{k-1}^n}^i - \Theta_i^T \phi(X_{t_{k-1}^n})\Delta_n|^2}{2\sigma_i^2\Delta_n} \right\}.$$

We assume the following conditions.

Assumption 5.1. (i) *The following conditions hold true;*

$$p_n \to \infty, \quad \log p_n / \sqrt{n\Delta_n} \to 0,$$

and $\Delta_n = \Delta n^{-\alpha}$*, for some* $\alpha \in (1/2, 1)$ *and positive constant* Δ*. Especially, the last condition implies that* $n\Delta_n = t_n^n \to \infty$*,* $\Delta_n \to 0$ *and that* $n\Delta_n^2 \to 0$ *as* $n \to \infty$*.*

(ii) *The functions* ϕ_i*'s are uniformly bounded and satisfy the global Lipschitz condition, i.e., there exist positive constants* L *and* L' *such that*

$$\sup_{1\le i<\infty} \sup_{x\in\mathbb{R}} |\phi_i(x)| \le L$$

and that

$$\sup_{1\le i<\infty} |\phi_i(x) - \phi_i(y)| \le L'|x-y|, \quad \forall x, y \in \mathbb{R}.$$

(iii) *For every* $\nu \ge 1$*, there exists a positive constant* $\tilde{C}_\nu$ *such that*

$$\sup_{1\le i<\infty} \sup_{t\in[0,\infty)} E\left[|X_t^i|^\nu\right] \le \tilde{C}_\nu.$$

Note that this assumption implies that

$$\sup_{t\in[0,\infty)} E\left[\sup_{1\le i\le p_n} |X_t^i|^\nu\right] \le p_n \tilde{C}_\nu, \quad \forall n \in \mathbb{N}.$$

(iv) *There exist some positive constants* K_1*,* K_2 *and* K_3 *such that*

$$\sup_{1\le i,j<\infty} |\Theta_{ij}^0| < K_1,$$

$$K_2 < \inf_{1\le i<\infty} \sigma_i^0 \le \sup_{1\le i<\infty} |\sigma_i^0| < K_3.$$

(v) *For every* $i \in \mathbb{N}$*, the* $\mathbb{R}^{S_i}$*-valued process* $\{X_{tT_0^i}\}_{t\in[0,T_n]}$ *is ergodic for* $\Theta = \Theta^0$ *and* $\sigma = \sigma^0$ *with invariant measure* μ_0^i*.*

Remark 5.2. *Let us discuss on Assumption 5.1-(v). Put* $\tilde{T}_0^i = \bigcup_{j\in T_0^i} T_0^j$*. It follows from a direct calculation that*

$$X_{t\tilde{T}_0^i} = X_{0\tilde{T}_0^i} + \int_0^t \left(\Theta^0_{\tilde{T}_0^i,\tilde{T}_0^i}\right)^\top \phi(X_{s\tilde{T}_0^i})_{\tilde{T}_0^i} ds + \sigma^0_{\tilde{T}_0^i,\tilde{T}_0^i} W_{t\tilde{T}_0^i}, \quad t \ge 0. \tag{5.3}$$

If this process is ergodic, then Assumption 5.1-(v) is valid because the process $\{X_{tT_0^i}\}$ *is a marginal process of* $\{X_{t\tilde{T}_0^i}\}_{t\ge0}$*. The conditions for the ergodic property of multidimensional diffusion processes can be checked by, for instance, Assumptions (D) and (E) in Gobet (2002).*

Remark 5.3. *As in Remark 1, if Θ^0 and $\phi(\cdot)$ satisfy Assumptions in Gobet (2002), then $\{X_t\}_{t\geq 0}$ has a stationary invariant measure in fixed dimensional settings. However, it is not obvious that Assumption 5.1-(iii) is verified in our high-dimensional setting since we need the uniformly boundedness in i. Even if high-dimensional matrix Θ^0 satisfies Assumptions in Gobet (2002), we can only observe that arbitrary finite dimensional marginal of $\{X_t\}_{t\geq 0}$ is stationary, which may not imply Assumption 5.1-(iii). Note moreover that for the process $\{X_t\}_{t\geq 0}$ which satisfies (5.1), Assumption 5.1-(iii) implies that for every $\nu \geq 1$, there exists a constant $C_\nu > 0$ such that for any n, $i = 1, 2, \ldots, p_n$ and $k = 1, 2, \ldots, n$,*

$$E\left[\sup_{s\in[t^n_{k-1},t^n_k]} |X^i_s - X^i_{t^n_{k-1}}|^\nu\right] \leq C_\nu \Delta_n^{\frac{\nu}{2}}.$$

Note moreover that uniformly bounded condition for ϕ_i's in Assumption 5.1-(ii) is just a technical condition to deal with high-dimensional settings. We may consider the estimation problems without this condition, however, in such settings, the condition for the dimension p and sample size n in Assumption 5.1-(i) should be more strong.

5.2 Estimators for diffusion coefficients

It is well-known that we can ignore the influence of drift coefficients when we estimate the diffusion coefficients (see e.g. Yoshida (1992)). We thus define the estimator for σ_i^0 by the solution $\hat{\sigma}_{n,i}$ to the equation

$$\frac{\partial}{\partial \sigma_i} l_n(0, \sigma_i) = 0, \quad i = 1, 2, \ldots, p_n,$$

by letting $\Theta = 0$. Note that $\hat{\sigma}_{n,i}$ can be written explicitly in the following way:

$$\hat{\sigma}_i^2 := \hat{\sigma}_{n,i}^2 = \frac{1}{n\Delta_n} \sum_{k=1}^n |X^i_{t^n_k} - X^i_{t^n_{k-1}}|^2.$$

The next theorem asserts the consistency of $\hat{\sigma}_i$ uniformly in i.

Theorem 5.4. *Under Assumption 5.1, it holds that*

$$\sup_{1\leq i\leq p_n} |\hat{\sigma}_i^2 - (\sigma_i^0)^2| \to^p 0, \quad n \to \infty.$$

Proof. It is clear that

$$\begin{aligned}
\hat{\sigma}_i^2 &= \frac{1}{n\Delta_n}\sum_{k=1}^{n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds+\sigma_i^0(W_{t_k^n}-W_{t_{k-1}^n})\right|^2 \\
&= \frac{1}{n\Delta_n}\sum_{k=1}^{n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right|^2 \\
&\quad +\frac{2\sigma_i^0}{n\Delta_n}\sum_{k=1}^{n}\left(\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right)\left(W_{t_k^n}-W_{t_{k-1}^n}\right) \\
&\quad +\frac{(\sigma_i^0)^2}{n\Delta_n}\sum_{k=1}^{n}(W_{t_k^n}-W_{t_{k-1}^n})^2.
\end{aligned}$$

Thus we have that

$$\hat{\sigma}_i^2-(\sigma_i^0)^2=(I)+(II)+(III),$$

where

$$(I)=\frac{1}{n\Delta_n}\sum_{k=1}^{n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right|^2,$$

$$(II)=\frac{2\sigma_i^0}{n\Delta_n}\sum_{k=1}^{n}\left(\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right)\left(W_{t_k^n}-W_{t_{k-1}^n}\right)$$

and

$$(III)=\frac{(\sigma_i^0)^2}{n\Delta_n}\sum_{k=1}^{n}\{(W_{t_k^n}^i-W_{t_{k-1}^n}^i)^2-\Delta_n\}.$$

Using Markov's and Schwartz's inequalities, we can evaluate (I) for every $\delta>0$

uniformly in i as follows

$$
\begin{aligned}
& P\left(\sup_{1\le i\le p_n}\frac{1}{n\Delta_n}\sum_{k=1}^{n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right|^2\ge\delta\right)\\
&\le \frac{1}{n\Delta_n\delta}\sum_{k=1}^{n}E\left[\sup_{1\le i\le p_n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right|^2\right]\\
&\le \frac{1}{n\Delta_n\delta}\sum_{k=1}^{n}E\left[\sup_{1\le i\le p_n}\Delta_n\int_{t_{k-1}^n}^{t_k^n}\left|(\Theta_i^0)^\top\phi(X_s)\right|^2ds\right]\\
&\le \frac{1}{n\delta}\sum_{k=1}^{n}\int_{t_{k-1}^n}^{t_k^n}E\left[\sup_{1\le i\le p_n}\|\Theta_i^0\|_1^2\max_{l\in T_0^i}|\phi_l(X_s^l)|^2\right]ds\\
&\le \frac{1}{\delta}\sup_{1\le i\le p_n}\|\Theta_i^0\|_1^2S^*L\Delta_n.
\end{aligned}
$$

The right-hand side of this inequality converges to 0 for every $\delta>0$. This yields that $(I)\to^p 0$ uniformly in i.

To evaluate (II), we use Φ_2-Orilicz norm introduced in the end of Introduction. We have that Φ_2-Orlicz norm for a standard normal random variable X is

$$\|X\|_{\Phi_2}=\sqrt{\frac{8}{3}}. \tag{5.4}$$

From (5.4) and Lemma 2.4 for Φ_p-Orlicz norm $\|\cdot\|_{\Phi_p}$, we can evaluate (II) for any

$\delta > 0$ uniformly in i as follows:

$$
\begin{aligned}
& P\left(\sup_{1\le i\le p_n}\frac{2\sigma_i^0}{n\Delta_n}\sum_{k=1}^{n}\left|\left(\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right)\left(W^i_{t_k^n}-W^i_{t_{k-1^n}}\right)\right|\ge\delta\right)\\
&\le \frac{2\sup_i\sigma_i^0}{n\Delta_n\delta}\sum_{k=1}^{n}E\left[\sup_{1\le i\le p_n}\left|\left(\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right)\left(W^i_{t_k^n}-W^i_{t_{k-1}^n}\right)\right|\right]\\
&\le \frac{2\sup_i\sigma_i^0}{n\Delta_n\delta}\sum_{k=1}^{n}\left(E\left[\sup_{1\le i\le p_n}\left|\int_{t_{k-1}^n}^{t_k^n}(\Theta_i^0)^\top\phi(X_s)ds\right|^2\right]\right)^{\frac12}\\
&\qquad\qquad\times\left\|\sup_{1\le i\le p_n}\left|W^i_{t_k^n}-W^i_{t_{k-1}^n}\right|\right\|_{L_2}\\
&\le \frac{2\sup_i\sigma_i^0}{n\Delta_n\delta}\sum_{k=1}^{n}(S^*L\Delta_n^2)^{\frac12}\sup_{1\le i\le p_n}\|\Theta_i^0\|_1K\sqrt{\log(1+p_n)}\left\|\left|W^i_{t_k^n}-W^i_{t_{k-1}^n}\right|\right\|_{\Phi_2}\\
&\le \frac{2\sup_i\sigma_i^0}{n\Delta_n\delta}\sum_{k=1}^{n}(S^*L\Delta_n^2)^{\frac12}\sup_{1\le i\le p_n}\|\Theta_i^0\|_1K\sqrt{\log(1+p_n)}\left(\frac{8\Delta_n}{3}\right)^{\frac12},
\end{aligned}
$$

where K is a positive constant which does not depend on n. The right-hand side of this inequality converges to 0 for every $\delta > 0$. So we have that $(II) \to^p 0$ uniformly in i.

(III) is a terminal value of an $\{\mathcal{F}_{t_k^n}\}_{k\ge0}$-martingale. We will apply Bernstein's inequality for martingales (See van de Geer (1995), Lemma 8.9.) to the following processes:

$$M_n^i := \sum_{k=1}^{n}\{(W^i_{t_k^n}-W^i_{t_{k-1}^n})^2-\Delta_n\}.$$

To do this, we shall evaluate the m-th moment for every integer $m \ge 2$;

$$\frac{1}{n}\sum_{k=1}^{n}E[|(W^i_{t_k^n}-W^i_{t_{k-1}^n})^2-\Delta_n|^m|\mathcal{F}_{t_{k-1}^n}].$$

Noting that $W^i_{t^n_k} - W^i_{t^n_{k-1}}$ is independent of $\mathcal{F}_{t^n_{k-1}}$, we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} E[|(W^i_{t^n_k} - W^i_{t^n_{k-1}})^2 - \Delta_n|^m | \mathcal{F}_{t^n_{k-1}}] \\
&= \frac{1}{n}\sum_{k=1}^{n} E[|(W^i_{t^n_k} - W^i_{t^n_{k-1}})^2 - \Delta_n|^m] \\
&\leq \frac{1}{n}\sum_{k=1}^{n}\sum_{r=0}^{m} \binom{m}{r} \Delta_n^{m-r} E[(W^i_{t^n_k} - W^i_{t^n_{k-1}})^{2r}] \\
&= \Delta_n^m + \sum_{r=1}^{m} \binom{m}{r} \Delta_n^{m-r} (2r-1)!! \Delta_n^r \\
&= \Delta_n^m + \sum_{r=1}^{m} \frac{(2r-1)!!}{r!(m-r)!} m! \Delta_n^m \\
&< \sum_{r=0}^{m} 2^r m! \Delta_n^m \\
&< \frac{m!}{2} (2\Delta_n)^{m-2} 4\Delta_n^2.
\end{aligned}$$

So it follows from Bernstein's inequality that for every $\epsilon > 0$,

$$P(|M^i_n| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2(2\Delta_n\epsilon + 4n\Delta_n^2)}\right).$$

Apply Lemma 2.5 to deduce that there exist a constant $L_1 > 0$ depending only on the function $\Phi_1(x) = e^x - 1$ such that

$$\left\| \sup_{1\leq i\leq p_n} |M^i_n| \right\|_{\Phi_1} \leq L_1\{2\Delta_n \log(1+p_n) + \sqrt{4n\Delta_n^2 \log(1+p_n)}\}.$$

So we obtain from Markov's inequality that

$$\begin{aligned}
&P\left(\sup_{1\leq i\leq p_n} |M^i_n| \geq \epsilon\right) \\
&\leq \Phi_1\left(\frac{\epsilon}{L_1\{2\Delta_n \log(1+p_n) + \sqrt{4n\Delta_n^2 \log(1+p_n)}\}}\right)^{-1}.
\end{aligned}$$

For every $\epsilon > 0$, the right-hand side of the above inequality converges to 0. Noting that

$$P\left(\sup_{1\leq i\leq p_n} |(III)| \geq \frac{(\sigma^0_i)^2\epsilon}{n\Delta_n}\right) = P\left(\sup_{1\leq i\leq p_n} |M^i_n| \geq \epsilon\right),$$

we obtain the conclusion. □

Note that Theorem 5.4 and Assumption 5.1 imply that there exists a constant $\tilde{K}_1$ such that

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \ge \tilde{K}_1\right) = 0.$$

5.3 Estimators for drift coefficients

In this section, we define the estimator for Θ_i by plugging $\hat{\sigma}_i$ in quasi-log-likelihood l_n. Hereafter, we write $\psi_n(\Theta_i)$ for the gradient of $l_n(\Theta_i, \hat{\sigma}_i)$ with respect to Θ_i, and V_n^i for Hessian of $-l_n(\Theta_i, \hat{\sigma}_i)$, *i.e.*,

$$\psi_n(\Theta_i) := \frac{1}{n\Delta_n \hat{\sigma}_i^2} \sum_{k=1}^{n} \phi(X_{t_{k-1}^n})(X^i_{t_k^n} - X^i_{t_{k-1}^n} - \Theta_i^T \phi(X_{t_{k-1}^n})\Delta_n),$$

$$V_n^i := \frac{1}{n\hat{\sigma}_i^2} \sum_{k=1}^{n} \phi(X_{t_{k-1}^n})\phi(X_{t_{k-1}^n})^\top.$$

Note that the Hessian matrix does not depend on Θ. Define the Dantzig selector type estimator $\hat{\Theta}_{n,i}$ for Θ_i^0 by

$$\hat{\Theta}_{n,i} := \hat{\Theta}_i := \underset{\Theta_i \in \mathcal{C}_n^i}{\arg\min} \|\Theta_i\|_1, \quad \mathcal{C}_n^i := \{\Theta_i \in \mathbb{R}^{p_n} : \|\psi_n(\Theta_i)\|_\infty \le \gamma_n^i\},$$

where γ_n^i is a tuning parameter. Hereafter, we assume the following condition about γ_n^i

Assumption 5.5. *γ_n^i satisfies the following equality for some positive constants c^i's which are uniformly bounded in i:*

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

where $\tilde{\gamma}_n := (\log p_n / n\Delta_n)^{1/4}$.

We define the quantity γ_n by

$$\gamma_n = \sup_{1\le i\le p_n} \gamma_n^i.$$

Under Assumption 5.5, it is obvious that there exists a constant $c \in (0, \infty)$ such that

$$\frac{\gamma_n}{\tilde{\gamma}_n} = \sup_{1\le i\le p_n} c^i \le c.$$

Some remarks about the choice of c^i's are described in Chapter 6. The goal of this section is to prove the consistency of $\hat{\Theta}_i$.

5.3.1 Some discussions on the gradient

Let us prove that

$$\sup_{1\le i\le p_n}\|\psi_n(\Theta_i^0)\|_\infty \le \gamma_n$$

with probability tending to 1. To do this, we decompose

$$\psi_n^j(\Theta_i^0) = A_n^{i,j} + B_n^{i,j},$$

where

$$A_n^{i,j} := \frac{1}{n\Delta_n\hat{\sigma}_i^2}\sum_{k=1}^n \phi_j(X^j_{t^n_{k-1}})\int_{t^n_{k-1}}^{t^n_k}(\Theta_i^0)^\top(\phi(X_s)-\phi(X_{t^n_{k-1}}))ds$$

and

$$B_n^{i,j} := \frac{\sigma_i^0}{n\Delta_n\hat{\sigma}_i^2}\sum_{k=1}^n \phi_j(X^j_{t^n_{k-1}})(W^i_{t^n_k}-W^i_{t^n_{k-1}}).$$

Lemma 5.6. *Under Assumptions 5.1 and 5.5, it holds that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i,j\le p_n}|A_n^{i,j}|\ge\gamma_n \text{ and } \sup_{1\le i\le p_n}\hat{\sigma}_i^{-2}\le\tilde{K}_1\right)=0.$$

Proof. We have that

$$\begin{aligned}
&P\left(\sup_{1\le i,j\le p_n}|A_n^{i,j}|\ge\gamma_n \text{ and } \sup_{1\le i\le p_n}\hat{\sigma}_i^{-2}\le\tilde{K}_1\right)\\
&\le \frac{\tilde{K}_1 L}{n\Delta_n\gamma_n}\sum_{k=1}^n E\left[\sup_{1\le i\le p_n}\left|\int_{t^n_{k-1}}^{t^n_k}(\Theta_i^0)^\top(\phi(X_s)-\phi(X_{t^n_{k-1}}))ds\right|\right]\\
&\le \frac{\tilde{K}_1 L}{n\Delta_n\gamma_n}\sum_{k=1}^n\int_{t^n_{k-1}}^{t^n_k} E\left[\sup_{1\le i\le p_n}\|\Theta_i^0\|_1\sup_{l\in T_0^i}|\phi_l(X_s^l)-\phi_l(X^l_{t^n_{k-1}})|\right]ds\\
&\le \frac{\tilde{K}_1 LL'\sup_{1\le i\le p_n}\|\Theta_i^0\|_1}{n\Delta_n\gamma_n}\sum_{k=1}^n\int_{t^n_{k-1}}^{t^n_k} E\left[\sup_{l\in T_0^i}|X_s^l-X^l_{t^n_{k-1}}|\right]ds\\
&\le \frac{\tilde{K}_1 LL'\sup_{1\le i\le p_n}\|\Theta_i^0\|_1}{n\Delta_n\gamma_n}\cdot n\cdot S^*\Delta_n^{\frac{3}{2}}.
\end{aligned}$$

The right-hand side of the above inequality converges to 0 under our assumptions. So we obtain the conclusion. □

Lemma 5.7. *Under Assumptions 5.1 and 5.5, it holds that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i,j\le p_n}|B_n^{i,j}|\ge\gamma_n \text{ and } \sup_{1\le i\le p_n}\hat{\sigma}_i^{-2}\le\tilde{K}_1\right)=0.$$

Proof. We apply Bernstein's inequality for martingales to the following terminal value of martingale :

$$\tilde{M}_n^{i,j} = \sum_{k=1}^{n} \phi_j(X^j_{t^n_{k-1}})(W^i_{t^n_k} - W^i_{t^n_{k-1}}).$$

For any integer $m \geq 2$, it holds that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} E\left[|\phi_j(X^j_{t^n_{k-1}})|^m |W^i_{t^n_k} - W^i_{t^n_{k-1}}|^m | \mathcal{F}_{t^n_{k-1}}\right] \\
&= \frac{1}{n}\sum_{k=1}^{n} |\phi_j(X^j_{t^n_{k-1}})|^m E[|W^i_{t^n_k} - W^i_{t^n_{k-1}}|^m] \\
&\leq L^m \Delta_n^{\frac{m}{2}} \frac{2^{\frac{m}{2}}\Gamma(\frac{m+1}{2})}{\pi^{\frac{1}{2}}} \\
&\leq \frac{m!}{2}(L\sqrt{2\Delta_n})^{m-2} L^2(2\Delta_n).
\end{aligned}$$

Put

$$K := L\sqrt{2\Delta_n}, \quad R^2 := L^2(2\Delta_n).$$

It follows from Bernstein's inequality that for any $\epsilon > 0$

$$P(|\tilde{M}_n^{i,j}| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2(\epsilon K + nR^2)}\right).$$

Using Lemma 2.5, we have that there exists a constant $L_2 > 0$ depending only on Φ_1 such that

$$\left\| \sup_{1\leq i,j\leq p_n} |\tilde{M}_n^{i,j}| \right\|_{\Phi_1} \leq L_2\{K\log(1+p_n^2) + \sqrt{nR^2\log(1+p_n^2)}\}.$$

For $\epsilon = n\Delta_n\gamma_n/(\sigma_i^0\tilde{K}_1)$, we obtain that

$$\begin{aligned}
&P\left(\sup_{1\leq i,j\leq p_n} |B_n^{i,j}| \geq \gamma_n \text{ and } \sup_{1\leq i\leq p_n} \hat{\sigma}_i^{-2} \leq \tilde{K}_1\right) \\
&\leq P\left(\sup_{1\leq i,j\leq p_n} |\tilde{M}_n^{i,j}| \geq \epsilon\right) \\
&\leq \Phi_1\left(\frac{\epsilon}{L_2\{K\log(1+p_n^2) + \sqrt{nR^2\log(1+p_n^2)}\}}\right)^{-1} \\
&\to 0.
\end{aligned}$$

□

Using the above lemmas, we obtain the following theorem.

Theorem 5.8. *Under Assumptions 5.1 and 5.5, it holds that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n\right) = 0.$$

Proof. It is obvious that Lemma 5.6 and 5.7 imply that

$$P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n \text{ and } \sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \le \tilde{K}_1\right) \to 0.$$

Noting that

$$\begin{aligned} & P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n\right) \\ = \;& P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n \text{ and } \sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \le \tilde{K}_1\right) \\ & + P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n \text{ and } \sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \ge \tilde{K}_1\right) \end{aligned}$$

and that

$$P\left(\sup_{1\le i\le p_n} \|\psi_n(\Theta_i^0)\|_\infty \ge 2\gamma_n \text{ and } \sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \ge \tilde{K}_1\right) \le P\left(\sup_{1\le i\le p_n} \hat{\sigma}_i^{-2} \ge \tilde{K}_1\right),$$

we obtain the conclusion. □

5.3.2 Some discussions on the Hessian

We introduce the following factors for V_n^i to deduce the l_q consistency of $\hat{\Theta}_i$.

Definition 5.9. *For every index set $T \subset \{1,\ 2,\ \cdots,\ p_n\}$ and $h \in \mathbb{R}^{p_n}$, h_T is an $\mathbb{R}^{|T|}$ dimensional sub-vector of h constructed by extracting the components of h corresponding to the indices in T. Define the set C_T by*

$$C_T := \{h \in \mathbb{R}^{p_n} : \|h_{T^c}\|_1 \le \|h_T\|_1\}.$$

(i) **Compatibility factor**

$$\kappa(T_0^i, V_n^i) := \inf_{0\ne h\in C_{T_0^i}} \frac{S_i^{\frac{1}{2}}(h^T V_n^i h)^{\frac{1}{2}}}{\|h_{T_0^i}\|_1}.$$

(*ii*) **Weak cone invertibility factor**

$$F_q(T_0^i, V_n^i) := \inf_{0 \neq h \in C_{T_0^i}} \frac{S_i^{\frac{1}{q}} h^T V_n^i h}{\|h_{T_0^i}\|_1 \|h\|_q}, \quad q \in [1, \infty).$$

$$F_\infty(T_0^i, V_n^i) := \inf_{0 \neq h \in C_{T_0^i}} \frac{(h^T V_n^i h)^{\frac{1}{2}}}{\|h\|_\infty}.$$

(*iii*) **Restricted eigenvalue**

$$RE(T_0^i, V_n^i) := \inf_{0 \neq h \in C_{T_0^i}} \frac{(h^T V_n^i h)^{\frac{1}{2}}}{\|h\|_2}.$$

We assume that $\kappa(T_0^i, V_n^i)$ satisfies the following condition.

Assumption 5.10. *For every $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,*

$$P\left(\inf_{1 \leq i \leq p_n} \kappa(T_0^i, V_n^i) > \delta\right) \geq 1 - \epsilon.$$

Noting that $\|h_{T_0^i}\|_1^q \geq \|h_{T_0^i}\|_q^q$ for all $q \geq 1$, we can see that $\kappa(T_0^i; V_n^i) \leq 2\sqrt{S_i} RE(T_0^i; V_n^i)$, and that $\kappa(T_0^i; V_n^i) \leq F_q(T_0^i; V_n^i)$. So under Assumption 5.10, the two factors $RE(T_0^i; V_n^i)$ and $F_q(T_0^i; V_n^i)$ also satisfy the corresponding conditions. See van de Geer and Bühlmann (2009) for the details of the matrix conditions to deal with the sparsity.

5.3.3 The consistency of the drift estimator

The following theorem give the l_q consistency of $\hat{\Theta}_i$ uniformly in i for every $q \in [1, \infty]$.

Theorem 5.11. *Under Assumptions 5.1, 5.5 and 5.10, the following (i)-(iv) hold true.*

(*i*) *It holds that*

$$\lim_{n \to \infty} P\left(\sup_{1 \leq i \leq p_n} \|\hat{\Theta}_i - \Theta_i^0\|_2^2 \geq \frac{4 \sup_{1 \leq i \leq p_n} \|\Theta_i^0\|_1 \gamma_n}{\inf_{1 \leq i \leq p_n} RE^2(T_0^i, V_n^i)}\right) = 0.$$

In particular, it holds that $\sup_{1 \leq i \leq p_n} \|\hat{\Theta}_i - \Theta_i^0\|_2 \to^p 0$ as $n \to \infty$.

(*ii*) *It holds that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} \|\hat{\Theta}_i - \Theta_i^0\|_\infty^2 \ge \frac{4\sup_{1\le i\le p_n}\|\Theta_i^0\|_1\gamma_n}{\inf_{1\le i\le p_n} F_\infty^2(T_0^i, V_n^i)}\right) = 0.$$

In particular, it holds that $\sup_{1\le i\le p_n}\|\hat{\Theta}_i - \Theta_i^0\|_\infty \to^p 0$ *as* $n\to\infty$.

(*iii*) *It holds that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} \|\hat{\Theta}_i - \Theta_i^0\|_1 \ge \frac{8S^*\gamma_n}{\inf_{1\le i\le p_n}\kappa^2(T_0^i, V_n^i)}\right) = 0.$$

In particular, it holds that $\sup_{1\le i\le p_n}\|\hat{\Theta}_i - \Theta_i^0\|_2 \to^p 0$ *as* $n\to\infty$.

(*iv*) *It holds for every* $q\in(1,\infty)$ *that*

$$\lim_{n\to\infty} P\left(\sup_{1\le i\le p_n} \|\hat{\Theta}_i - \Theta_i^0\|_q \ge \frac{4S^{*\frac{1}{q}}\gamma_n}{\inf_{1\le i\le p_n} F_q(T_0^i, V_n^i)}\right) = 0.$$

In particular, it holds that $\sup_{1\le i\le p_n}\|\hat{\Theta}_i - \Theta_i^0\|_q \to^p 0$ *as* $n\to\infty$.

Proof. (i) and (ii): It is sufficient to show that

$$\sup_{1\le i\le_n} \|\psi_n(\Theta_i^0)\|_\infty \le \gamma_n$$

implies that

$$\sup_{1\le i\le p_n} \|\hat{\Theta}_i - \Theta_i^0\|_2^2 \le \frac{4\sup_{1\le i\le p_n}\|\Theta_i^0\|_1\gamma_n}{\inf_{1\le i\le p_n} RE^2(T_0^i, V_n^i)}.$$

By the definition of $\hat{\Theta}_i$, we have that

$$\sup_{1\le i\le p_n} \|\psi_n(\hat{\Theta}_i)\|_\infty \le \sup_{1\le i\le p_n}\gamma_n^i = \gamma_n.$$

It follows from the triangle inequality that

$$\sup_{1\le i\le p_n} \|\psi_n(\hat{\Theta}_i) - \psi_n(\Theta_i^0)\|_\infty \le 2\gamma_n.$$

Put $h_i := \hat{\Theta}_i - \Theta_i^0$. Noting that $\hat{\Theta}_i$ is the minimizer of the l_1 norm, we can easily check that $h_i \in C_{T_0^i}$. In fact, it holds that

$$\begin{aligned} 0 \ge \|\Theta_i^0 + h_i\|_1 - \|\Theta_i^0\|_1 &= \sum_{j\in(T_0^i)^c}|h_{ij}| + \sum_{j\in T_0}(|\Theta_{ij}^0 + h_{ij}| - |\Theta_{ij}^0|) \\ &\ge \sum_{j\in(T_0^i)^c}|h_{ij}| - \sum_{j\in T_0^i}|h_{ij}| \\ &= \|h_{i(T_0^i)^c}\|_1 - \|h_{iT_0^i}\|_1. \end{aligned}$$

By the Taylor expansion, we have that

$$h_i^T[\psi_n(\hat{\Theta}_i) - \psi_n(\Theta_i^0)] = h_i^T V_n^i h_i.$$

Noting that $\|h_i\|_1 \leq \|\hat{\Theta}_i\|_1 + \|\Theta_i^0\|_1 \leq 2\|\Theta_i^0\|_1$ and that $\sup_{1\leq i\leq p_n} \|\psi_n(\hat{\Theta}_i)\|_\infty \leq \gamma_n$, we have that

$$\begin{aligned}
\sup_{1\leq i\leq p_n} h_i^T V_n^i h_i &= \sup_{1\leq i\leq p_n} h_i^T[\psi_n(\hat{\Theta}_i) - \psi_n(\Theta_i^0)] \\
&\leq \sup_{1\leq i\leq p_n} \|h_i\|_1 \|\psi_n(\hat{\Theta}_i) - \psi_n(\Theta_i^0)\|_\infty \\
&\leq 2 \sup_{1\leq i\leq p_n} \|\Theta_i^0\|_1 \left\{ \|\psi_n(\Theta_i^0)\|_\infty + \|\psi_n(\hat{\Theta}_i)\|_\infty \right\} \\
&\leq 4 \sup_{1\leq i\leq p_n} \|\Theta_i^0\|_1 \gamma_n.
\end{aligned}$$

By the definition of $RE(T_0^i, V_n^i)$, we obtain that:

$$\begin{aligned}
RE^2(T_0^i, V_n^i)\|h_i\|_2^2 &\leq h_i^T V_n^i h \\
\sup_{1\leq i\leq p_n} RE^2(T_0^i, V_n^i)\|h_i\|_2^2 &\leq 4 \sup_{1\leq i\leq p_n} \|\Theta_i^0\|_1 \gamma_n \\
\sup_{1\leq i\leq p_n} \|\hat{\Theta}_i - \Theta_i^0\|_2^2 &\leq \frac{4 \sup_{1\leq i\leq p_n} \|\Theta_i^0\|_1 \gamma_n}{\inf_{1\leq i\leq p_n} RE^2(T_0^i, V_n^i)}.
\end{aligned}$$

These facts yield our conclusion in (i). Using the factor $F_\infty(T_0^i, V_n^i)$ in place of $RE(T_0^i, V_n^i)$, we obtain the conclusion in (ii) by the similar way.

(iii) and (iv): It follows from the proof of (i) and (ii) that

$$\sup_{1\leq i\leq p_n} h_i^T V_n^i h_i \leq 2 \sup_{1\leq i\leq p_n} \|h_i\|_1 \gamma_n.$$

So by the definition of $\kappa(T_0^i, V_n^i)$, we have that

$$\begin{aligned}
\kappa^2(T_0^i, V_n^i) \sup_{1\leq i\leq p_n} \|h_i\|_1^2 &\leq 4S^* \sup_{1\leq i\leq p_n} h_i^T V_n^i h_i \\
&\leq 8S^* \sup_{1\leq i\leq p_n} \|h_i\|_1 \gamma_n.
\end{aligned}$$

We therefore obtain that

$$\sup_{1\leq i\leq p_n} \|h_i\|_1 \leq \frac{8S^*\gamma_n}{\inf_{1\leq i\leq p_n} \kappa^2(T_0^i, V_n^i)}.$$

This yields our conclusion in (iii).

On the other hand, by the definition of the factor $F_q(T_0^i, V_n^i)$, we have that

$$F_q(T_0^i, V_n^i) \leq \frac{4S^{*\frac{1}{q}}\gamma_n}{\|h_i\|_q}.$$

This yields our conclusion in (iv). □

Remark 5.12. *The asymptotic rate $\tilde{\gamma}_n$ of the tuning parameter $\gamma_n^i = c^i\tilde{\gamma}_n$ needs to satisfy the following condition;*

$$\frac{\sqrt{\Delta_n}}{\tilde{\gamma}_n} \to 0, \tag{5.5}$$

$$\frac{\sqrt{\Delta}_n \log p_n}{n\Delta_n\tilde{\gamma}_n} \to 0 \tag{5.6}$$

and

$$\frac{\sqrt{n\Delta_n \log p_n}}{n\Delta_n\tilde{\gamma}_n} \to 0 \tag{5.7}$$

as $n \to \infty$, which can be seen in the proof of Lemmas 5.6 and 5.7. Moreover, to obtain the consistency of the estimators, it is necessary that $\tilde{\gamma}_n \to 0$. Since the left-hand side of (5.7) is larger than those of (5.5) and (5.6), we can choose the $\tilde{\gamma}_n$ by a solution to the following equation

$$\tilde{\gamma}_n = \frac{\sqrt{n\Delta_n \log p_n}}{n\Delta_n\tilde{\gamma}_n}.$$

We therefore obtain that

$$\tilde{\gamma}_n = (\log p_n / n\Delta_n)^{1/4}.$$

We discuss the choice of the constant c^i in Section 6.

Remark 5.13. *Theorem 5.11 implies that the estimator $\hat{\Theta}$ converges to the drift matrix Θ^0, i.e.,*

$$\|\hat{\Theta} - \Theta^0\|_\infty \to^p 0$$

as $n \to \infty$. For each $i \in \{1, 2, \ldots, p_n\}$, we have that

$$\lim_{n\to\infty} P\left(\|\hat{\Theta}_i - \Theta_i^0\|_\infty^2 \geq \frac{4\|\Theta_i^0\|_1\gamma_n^i}{F_\infty^2(T_0^i, V_n^i)}\right) = 0$$

or

$$\lim_{n\to\infty} P\left(\|\hat{\Theta}_i - \Theta_i^0\|_1 \geq \frac{8S_i\gamma_n^i}{\kappa^2(T_0^i, V_n^i)}\right) = 0.$$

We can observe from these facts that the rate of convergence of the l_∞ and l_1 norm is $\tilde{\gamma}_n^{1/2}$ and $\tilde{\gamma}_n$ respectively. The similar error bounds holds for other norms. Although the l_1 norm is always greater than l_∞ norm, the rate of convergence of l_1 norm is $\tilde{\gamma}_n$, which is faster than that of l_∞ norm. This is caused by the fact that the sparsity S_i is fixed. If S_i is assumed to be dependent on n, then, the rate of convergence would become slower than that of l_∞ norm.

5.4 Variable selection by the Dantzig selector

5.4.1 Estimator for the support index set of the drift coefficients

In this subsection, we propose the estimator of the support index set T_0^i of the true value Θ_i^0 as follows.

$$\hat{T}_n^i := \{j : |\hat{\Theta}_{ij}| > \gamma_n^i\}.$$

We shall prove that $\hat{T}_n^i = T_0^i$ for sufficiently large n with probability tending to 1.

Theorem 5.14. *Under Assumptions 5.1, 5.5 and 5.10, it holds that*

$$\lim_{n\to\infty} P\left(\hat{T}_n^i = T_0^i \text{ for all } i \in \{1, 2, \ldots, p_n\}\right) = 1.$$

Proof. We have that

$$\lim_{n\to\infty} P\left(\|\hat{\Theta}_i - \Theta_i^0\|_1 > \gamma_n^i \text{ for all } i \in \{1, 2, \ldots, p_n\}\right) = 0$$

by Theorem 5.11. So it is sufficient to show that for every i, the next inequality

$$\|\hat{\Theta}_i - \Theta_i^0\|_1 \leq \gamma_n^i, \text{ for all } i \in \{1, 2, \ldots, p_n\}$$

implies that

$$\hat{T}_n^i = T_0^i, \text{ for all } i \in \{1, 2, \ldots, p_n\}.$$

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

$$|\Theta_{ij}^0| - |\hat{\Theta}_{ij}| \leq |\hat{\Theta}_{ij} - \Theta_{ij}^0| \leq \|\hat{\Theta}_i - \Theta_i^0\|_1 \leq \gamma_n^i.$$

Then, we have that

$$|\hat{\Theta}_{ij}| \geq |\Theta_{ij}^0| - \gamma_n^i > \gamma_n^i$$

for sufficiently large n, which implies that $T_0^i \subset \hat{T}_n^i$ for every $i \in \{1, \ldots, p_n\}$. On the other hand, for every $j \in (T_0^i)^c$, we have that

$$|\hat{\Theta}_{ij} - \Theta_{ij}^0| = |\hat{\Theta}_{ij}| \leq \gamma_n^i$$

since it holds that $\Theta_{ij}^0 = 0$. Thus, we can see that $j \in (\hat{T}_n^i)^c$ which implies that $\hat{T}_n^i \subset T_0^i$ for every $i \in \{1, \ldots, p_n\}$. We have obtained the conclusion. □

5.4.2 New estimator for drift coefficients after variable selection

We construct the new estimator $\hat{\Theta}_i^{(2)}$ by the solution to the next equation

$$\psi_n(\Theta_i)_{\hat{T}_n^i} = 0, \quad \Theta_{i(\hat{T}_n^i)^c} = 0. \tag{5.8}$$

We will prove the asymptotic normality of the estimator $\hat{\Theta}_{i\hat{T}_n^i}^{(2)}$ for every $i \in \{1, 2, \ldots, p_n\}$. In order to consider the asymptotic distribution, we assume the following condition concerning with the Fisher information matrix.

Assumption 5.15. *Define the $S_i \times S_i$ matrix $Q_{T_0^i, T_0^i}^i$ by*

$$Q_{T_0^i, T_0^i}^i := \frac{1}{(\sigma_i^0)^2} \int_{\mathbb{R}^{S_i}} \phi(x)_{T_0^i} \phi(x)_{T_0^i}^\top \mu_0^i(dx).$$

It holds that $Q_{T_0^i, T_0^i}^i$ is invertible for every $i = 1, 2, \ldots, p_n$.

The next lemma states that $V_{nT_0^i, T_0^i}^i$ is approximated by $Q_{T_0^i, T_0^i}^i$ with probability tending to 1 as $n \to \infty$.

Lemma 5.16. *Define the random sequence ϵ_n^i by*

$$\epsilon_n^i := \|V_{nT_0^i, T_0^i}^i - Q_{T_0^i, T_0^i}^i\|_\infty, \quad i \in \{1, 2, \ldots, p_n\}.$$

Under Assumption 5.1, it holds that $\epsilon_n^i \to^p 0$ as $n \to \infty$ for every i.

Proof. Note that

$$V_{nT_0^i, T_0^i}^i = \frac{1}{n\hat{\sigma}_i^2} \sum_{k=1}^n \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i}^\top.$$

It holds that

$$\epsilon_n^i \leq (I) + (II) + (III),$$

where

$$(I) := \left\| V^i_{nT^i_0,T^i_0} - \frac{1}{T_n\hat{\sigma}_i^2}\int_0^{T_n} \phi(X_{tT^i_0})_{T^i_0}\phi(X_{tT^i_0})^\top_{T^i_0}dt \right\|_\infty ,$$

$$(II) := \left\| \frac{1}{T_n\hat{\sigma}_i^2}\int_0^{T_n} \phi(X_{tT^i_0})_{T^i_0}\phi(X_{tT^i_0})^\top_{T^i_0}dt - \frac{1}{\hat{\sigma}_i^2}\int_{\mathbb{R}^{S_i}} \phi(x)_{T^i_0}\phi(x)^\top_{T^i_0}\mu^i_0(dx) \right\|_\infty$$

and

$$(III) := \left\| \frac{1}{\hat{\sigma}_i^2}\int_{\mathbb{R}^{S_i}} \phi(x)_{T^i_0}\phi(x)^\top_{T^i_0}\mu^i_0(dx) - Q^i_{T^i_0,T^i_0} \right\|_\infty .$$

It is obvious that (II) and (III) are $o_p(1)$ by Assumption 5.1 and Theorem 5.4. To complete the proof, it is sufficient to prove that

$$P\left((I) \geq \delta \text{ and } \sup_{1\leq i\leq p_n} \hat{\sigma}_i^{-2} \leq \tilde{K}_1 \right) \to 0$$

as $n \to \infty$ for every $\delta > 0$. Using Markov's inequality, we can see that

$$\begin{aligned} & P\left((I) \geq \delta \text{ and } \sup_{1\leq i\leq p_n} \hat{\sigma}_i^{-2} \leq \tilde{K}_1 \right) \\ \leq\ & \frac{\tilde{K}_1}{n\Delta_n\delta}\sum_{k=1}^n \int_{t^n_{k-1}}^{t^n_k} E\left[\sup_{j,l\in T^i_0} |\phi_j(X^j_t)\phi_l(X^l_t) - \phi_j(X^j_{t^n_{k-1}})\phi_l(X^l_{t^n_{k-1}})| \right] dt. \end{aligned}$$

Moreover, it follows from Schwartz's inequality that

$$\begin{aligned} & E\left[\sup_{j,l\in T^i_0} |\phi_j(X^j_t)\phi_l(X^l_t) - \phi_j(X^j_{t^n_{k-1}})\phi_l(X^l_{t^n_{k-1}})| \right] \\ \leq\ & E\left[\sup_{j,l\in T^i_0} |\phi_l(X^l_t)(\phi_j(X^j_t) - \phi_j(X^j_{t^n_{k-1}}))| \right] \\ & + E\left[\sup_{j,l\in T^i_0} |\phi_j(X^j_{t^n_{k-1}})(\phi_l(X^l_t) - \phi_l(X^l_{t^n_{k-1}}))| \right] \\ \leq\ & \left(E\left[\sup_{l\in T^i_0} |\phi_l(X^l_t)|^2 \right] \right)^{\frac{1}{2}} \left(E\left[\sup_{j\in T^i_0} |\phi_j(X^j_t) - \phi_j(X^j_{t^n_{k-1}})|^2 \right] \right)^{\frac{1}{2}} \\ & + \left(E\left[\sup_{j\in T^i_0} |\phi_j(X^j_{t^n_{k-1}})|^2 \right] \right)^{\frac{1}{2}} \left(E\left[\sup_{l\in T^i_0} |\phi_l(X^l_t) - \phi_l(X^l_{t^n_{k-1}})|^2 \right] \right)^{\frac{1}{2}} \\ \leq\ & 2S^*LL'\Delta_n^{\frac{1}{2}}. \end{aligned}$$

We thus have that

$$P\left((I) \geq \delta \text{ and } \sup_{1\leq i\leq p_n} \hat{\sigma}_i^{-2} \leq \tilde{K}_1\right) \leq \frac{2\tilde{K}_1 LL' S^*}{\delta} \cdot \Delta_n^{\frac{1}{2}}.$$

The right-hand-side of this inequality converges to 0 for every $\delta > 0$, which means that $(I) = o_p(1)$. □

Remark 5.17. *By Assumption 5.15 and Lemma 5.16, the solution to the equation (5.8) exists with probability tending to* 1*, i.e., the estimator* $\hat{\Theta}_i^{(2)}$ *is well-defined with large probability. In fact, under the condition that* $\hat{\Theta}^{(2)}_{i(\hat{T}_n^i)^c} = 0$*, we have that*

$$V^i_{n\hat{T}_n^i,\hat{T}_n^i}\hat{\Theta}^{(2)}_{i\hat{T}_n^i} = \frac{1}{n\Delta_n\hat{\sigma}_i^2}\sum_{k=1}^{n}\phi(X_{t^n_{k-1}\hat{T}_n^i})_{\hat{T}_n^i}(X^i_{t^n_k} - X^i_{t^n_{k-1}}).$$

Therefore, under Assumption 5.15, $\hat{\Theta}_i^{(2)}$ *exists with probability tending to* 1 *since the matrix* $V^i_{n\hat{T}_n^i,\hat{T}_n^i}$ *converges to a nonsingular matrix in probability by Lemma 5.14 and 5.16.*

Now, we are ready to prove the asymptotic normality of $\hat{\Theta}^{(2)}_{i\hat{T}_n^i}$ in the following sense.

Theorem 5.18. *Under Assumptions 5.1, 5.5, 5.10 and 5.15, it holds for every* $i \in \mathbb{N}$ *that*

$$\sqrt{n\Delta_n}(\hat{\Theta}^{(2)}_{i\hat{T}_n^i} - \Theta^0_{iT_0^i})1_{\{\hat{T}_n^i = T_0^i\}} \to^d N\left(0, \left(Q^i_{T_0^i,T_0^i}\right)^{-1}\right)$$

as $n \to \infty$*. Note that for every* $i \in \mathbb{N}$*, it holds that* $i < p_n$ *for sufficiently large* n*.*

Proof. Using the Taylor expansion, we have that

$$\psi_n(\hat{\Theta}_i)_{\hat{T}_n^i} = \psi_n(\Theta_i^0)_{\hat{T}_n^i} - V^i_{n\hat{T}_n^i\hat{T}_n^i}(\hat{\Theta}^{(2)}_{n\hat{T}_n^i} - \Theta^0_{i\hat{T}_n^i}).$$

It follows from the definition of the estimator $\hat{\Theta}_i^{(2)}$ that

$$\sqrt{n\Delta_n}V^i_{n\hat{T}_n^i\hat{T}_n^i}(\hat{\Theta}^{(2)}_{i\hat{T}_n^i} - \Theta^0_{iT_0^i})1_{\{\hat{T}_n^i = T_0^i\}} = \sqrt{n\Delta_n}\psi_n(\Theta_i^0)_{T_0^i}1_{\{\hat{T}_n^i = T_0^i\}}.$$

We decompose $\sqrt{t_n^n}\psi_n(\Theta^0_{iT_0^i})_{T_0^i} = (I) + (II) + (III)$, where

$$(I) = \frac{\sigma_i^0}{\sqrt{n\Delta_n}\hat{\sigma}_i^2}\sum_{k=1}^{n}\phi(X_{t^n_{k-1}T_0^i})_{T_0^i}\int_{t^n_{k-1}}^{t^n_k}(\Theta^0_{iT_0^i})^\top\{\phi(X_{sT_0^i})_{T_0^i} - \phi(X_{t^n_{k-1}T_0^i})_{T_0^i}\}ds,$$

$$(II) = \left(\frac{\sigma_i^0}{\sqrt{n\Delta_n}\hat{\sigma}_i^2} - \frac{1}{\sqrt{n\Delta_n}\sigma_i^0} \right) \sum_{k=1}^{n} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} (W_{t_k^n}^i - W_{t_{k-1}^n}^i)$$

and

$$(III) = \frac{1}{\sqrt{n\Delta_n}\sigma_i^0} \sum_{k=1}^{n} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} (W_{t_k^n}^i - W_{t_{k-1}^n}^i)$$

We can show that $(I) = o_p(1)$ by the similar way to the proof of Lemma 5.6. Next, we will apply the martingale central limit theorem for (III). Define the martingale differences $\{\xi_k\}_{k=1,2,\ldots,n}$ by

$$\xi_k := \frac{1}{\sqrt{n\Delta_n}\sigma_i^0} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} (W_{t_k^n}^i - W_{t_{k-1}^n}^i).$$

It holds for every $j, l \in T_0^i$ that

$$\begin{aligned}
&\sum_{k=1}^{n} E\left[\frac{1}{n\Delta_n (\sigma_i^0)^2} \phi_j(X_{t_{k-1}^n}^j) \phi_l(X_{t_{k-1}^n}^l) (W_{t_k^n}^i - W_{t_{k-1}^n}^i)^2 | \mathcal{F}_{t_{k-1}^n} \right] \\
&= \frac{1}{n\Delta_n (\sigma_i^0)^2} \sum_{k=1}^{n} \phi_j(X_{t_{k-1}^n}^j) \phi_l(X_{t_{k-1}^n}^l) E[(W_{t_k^n}^i - W_{t_{k-1}^n}^i)^2] \\
&= \frac{1}{n(\sigma_i^0)^2} \sum_{k=1}^{n} \phi_j(X_{t_{k-1}^n}^j) \phi_l(X_{t_{k-1}^n}^l).
\end{aligned}$$

We can see that right-hand side converges to the (j,l)-component of the matrix $Q_{T_0^i T_0^i}^i$ in probability by the same way of the proof of Lemma 5.16. Moreover, we can check Lyapnov's condition:

$$\sum_{k=1}^{n} E\left[\|\xi_k\|_2^{2+\delta} | \mathcal{F}_{t_{k-1}^n} \right] \to^p 0$$

for $\delta = 2$, which implies Lindeberg's condition:

$$\sum_{k=1}^{n} E\left[\|\xi_k\|_2^2 1_{\{\|\xi_k\|_2 > \epsilon\}} | \mathcal{F}_{t_{k-1}^n} \right] \to^p 0,$$

for every $\epsilon > 0$. Thus, we obtain that

$$\frac{1}{\sqrt{n\Delta_n}\sigma_i^0} \sum_{k=1}^{n} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} (W_{t_k^n}^i - W_{t_{k-1}^n}^i) \to^d N(0, Q_{T_0^i T_0^i}^i)$$

by the martingale central limit theorem. Noting that

$$(II) = \left(\frac{(\sigma_i^0)^2}{\hat{\sigma}_i^2} - 1 \right) \frac{1}{\sqrt{n\Delta_n}\sigma_i^0} \sum_{k=1}^{n} \phi(X_{t_{k-1}^n T_0^i})_{T_0^i} (W_{t_k^n}^i - W_{t_{k-1}^n}^i)$$

and $(III) = O_p(1)$, we obtain that $(II) = o_p(1)$ since $\hat{\sigma}_i$ is a consistent estimator for σ_i^0. Using the above results and Lemma 5.16, we have that

$$\sqrt{n\Delta_n}(\hat{\Theta}^{(2)}_{i\hat{T}_n^i} - \Theta^0_{iT_0^i}) 1_{\{\hat{T}_n^i = T_0^i\}} = \left(Q^i_{T_0^i T_0^i} \right)^{-1} \sqrt{n\Delta_n} \psi_n(\Theta_i^0)_{T_0^i} 1_{\{\hat{T}_n^i = T_0^i\}} + o_p(1).$$

Since it holds that $1_{\{\hat{T}_n^i = T_0^i\}} \to^p 1$ by Theorem 5.14, we can use Slutsky's theorem to derive our conclusion. □

5.5 Concluding remarks

In summary, we have been able to construct the consistent and asymptotically normal estimator for the model (5.1) even in high-dimensional settings if the sparsity of the parameter is fixed or bounded. If the sparsity is not bounded, we may not reduce the dimension of the parameter. In such cases, the asymptotically normal estimator can not be constructed by the equation (5.8).

To construct an asymptotically efficient estimators for this model in high-dimensional settings includes some remaining problems since the theoretical properties of the estimators strongly depend on the choice of the tuning parameter which works for the variable selection. Therefore, to discuss such problems, we have to construct the "optimal" choice of the tuning parameter which achieve the "optimal" variable selection.

In this paper, we have assumed that the diffusion coefficients σ_i's are constants. However, it may be possible to consider the case when each σ_i has more complicated structures. For example, we can consider the following model:

$$X_t^i = X_0^i + \int_0^t \Theta_i^T \phi(X_s) ds + \int_0^t \exp(\beta_i^T \varphi(X_s)) dW_s^i, \quad i = 1, 2, \ldots, p,$$

where $\beta_i \in \mathbb{R}^p$ and $\varphi(\cdot)$ is an appropriate smooth function. According to Fujimori and Nishiyama (2017b), we can construct estimators for β_i by the Dantzig selector and prove the l_q consistency of the estimators for every $q \in [1, \infty]$. Therefore, we may prove the same asymptotic properties of Θ even for the above model which has high-dimensional parameters in diffusion coefficients.

The variable selection consistency of the estimator of drift matrix is important for applications such as graphical modeling as it can be seen in Ravikumar et al. (2010),

Periera and Ibrahimi (2014) and Gobet and Matulewicz (2017). In the future, we would like to consider such applications and present some numerical results.

Chapter 6

Numerical studies

In this chapter, we demonstrate the finite sample performance of the Dantzig selector for a linear regression model and Cox's proportional hazards model in high-dimensional and sparse settings. We have proved that the Dantzig selectors satisfy the l_q consistency when we choose "good" tuning parameter. Moreover, the consistency results enable us to construct the estimators for the support index sets of the true values by using the thresholding method obtained by "good" tuning parameter. Therefore, we have to discuss about how to choose "good" tuning parameters to ensure the performance of the Dantzig selector. For *i.i.d.* models in high-dimensional and sparse settings, the methods to choose the tuning parameters for penalized estimators or the Dantzig selector have been discussed by many authors. For example, Bayesian information criterion discussed in Wang et al. (2009) is widely used. In addition, the jointly estimation for the regression parameters and noise level which is determines the tuning parameter is studied in Sun and Zhang (2012). One of the most famous method is cross validation which is studied and used in several literatures dealing with estimation problems for high-dimensional and sparse settings. It may be possible to apply them to models of stochastic processes. However, the theoretical properties and practical performance of them for our models of stochastic processes have not yet been verified. Instead of these methods, we propose an intuitive algorithm to choose tuning parameters in Section 6.1. In Sections 6.2 and 6.3, we present the numerical results of the Dantzig selectors for a linear regression model and Cox's proportional hazards model respectively. We focus on the l_1 consistency result compared with the classical estimators such as least square estimator (LSE) and maximum partial likelihood estimator (MPLE) and the variable selection results. Since the estimators after selection are the classical Z-estimators and their performances have been well-studied, we omit this part.

6.1 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. Generally, the Dantzig selector $\hat{\theta}_n$ for several regression models with unknown parameter θ is defined by the following form

$$\hat{\theta}_n := \arg\min_{\theta \in \mathcal{C}_n} \|\theta\|_1, \quad \mathcal{C}_n = \{\theta \in \mathbb{R}^{p_n} : \|\Psi_n(\theta)\|_\infty \leq \gamma_n\},$$

where $\gamma_n \geq 0$ is a tuning parameter and $\Psi_n(\cdot)$ is the score function for the model. If we can evaluate the rate of convergence $\tilde{\gamma}_n$ of $\|\Psi_n(\theta_0)\|_\infty$, where θ_0 is the true value of the unknown parameter θ, we can define the γ_n as follows:

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

where $c \geq 0$ is a constant not depending on n. To ensure the l_q consistency and the variable selection consistency for finite sample, c has to satisfy that

$$\frac{\|\Psi_n(\theta_0)\|_\infty}{\tilde{\gamma}_n} \leq c \leq \frac{\inf_{j \in T_0} |\theta_0^j|}{\tilde{\gamma}_n}, \tag{6.1}$$

where T_0 is the support index set of the true value θ_0. The problem is how to choose c which satisfies (6.1). Note that

$$\frac{\|\Psi_n(\theta_0)\|_\infty}{\tilde{\gamma}_n} = O_p(1)$$

and that

$$\frac{\inf_{j \in T_0^i} |\theta_0^j|}{\tilde{\gamma}_n} \to \infty$$

as $n \to \infty$. When we choose the small tuning parameter satisfying the inequality (6.1), 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:

Step 1. Let $c^{[1]}$ be a positive prefixed constant.

Step 2. Calculate the Dantzig selector for $j \geq 1$ by using $c^{[j]}$;

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

Step 3. Put

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

Step 4. Repeat Step 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^0 has to be chosen large enough to ensure that

$$\frac{\|\Psi_n(\theta_0)\|_\infty}{\tilde{\gamma}_n} \leq c^0.$$

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

$$C := \frac{\|\Psi_n(\theta_0)\|_\infty}{\tilde{\gamma}_n},$$

with probability tending to 1 as $n \to \infty$ since it holds that $\|\hat{\theta}^{[j]} - \theta_0\|_1 \to^p 0$. In addition, for each sufficiently large $n \in \mathbb{N}$, we can also verify that the sequence $\{c^{[j]}\}_{j\in\mathbb{N}}$ is nonincreasing 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 be holds 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:

$$\Psi_n(\theta) = 0.$$

Even though we can easily observe that $c^{[j]}$ converges to a positive constant and works well for variable selection for the usual linear regression model numerically, we have not proved that performance of this method theoretically and numerically for our models of stochastic processes.

6.2 Linear regression models

In this section, we demonstrate the finite sample performance of the Dantzig selector for the following linear regression model:

$$Y_i = Z_i^\top \beta + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where Z_i's are $i.i.d.$ $[-2,2]$-valued uniform random variables and ϵ_i's are $i.i.d.$ standard normal random variables. The tuning parameter γ_n is determined by the algorithm in Section 6.1 with a sufficiently large initial value $\tilde{\gamma}_n = n^{-0.3}\log p$ to obtain the decreasing sequence of tuning parameter by the algorithm for each case. We put $p = 50$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{50}$$

in Case 1, $p = 100$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{100}$$

in Case 2 and $p = 100$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{150}$$

in Case 3. We apply the Dantzig selector to this model, which can be calculated by the algorithm proposed by Candés and Tao (2007) when $n = 50$ and $n = 100$ for 1000 times.

For these 1000 estimators, we use the variable selection criterion proposed in Section 2.4. Tables 6.1 and 6.2 show the proportion of successes of the variable selection among all 1000 estimators. We can see that the Dantzig selector selects the support index set T_0 correctly. This result enables us to reduce the dimension correctly even when $n = 50$.

Tables 6.3 and 6.4 show the l_1 errors of estimators LSE, the Dantzig selector $\hat{\beta}_n$ and the second estimator $\hat{\beta}_n^{(2)}$ after dimension reduction by using $\hat{T}_n$. Note that we cannot construct LSE in when $p \geq n$ since the rank of the optimization problem is deficient. We can observe that for all cases, the Dantzig selectors work better than LSE and even for the higher dimensional case, the l_1 errors keep small values. Moreover, we can verify that the behaviors of the second estimators are the best since the selections work well as observed in Tables 6.1 and 6.2.

Case	$p = 50$	$p = 100$	$p = 150$
$\hat{T}_n = T_0$	93.1 %	92.0 %	90.7 %

Table 6.1: Variable selection results ($n = 50$).

Case	$p = 50$	$p = 100$	$p = 150$
$\hat{T}_n = T_0$	99.4 %	99.9 %	99.6 %

Table 6.2: Variable selection results ($n = 100$).

Case	$p = 50$	$p = 100$	$p = 150$
LSE	214.2453	NA	NA
$\hat{\beta}_n$	1.815687	2.261749	2.620261
$\hat{\beta}_n^{(2)}$	0.5189128	0.535665	0.5718938

Table 6.3: l_1 errors of estimators ($n = 50$).

Case	$p = 50$	$p = 100$	$p = 150$
LSE	3.789864	173.243	NA
$\hat{\beta}_n$	0.9972455	1.113642	1.264328
$\hat{\beta}_n^{(2)}$	0.3551882	0.3320217	0.3541766

Table 6.4: l_1 errors of estimators ($n = 100$).

6.3 Cox's proportional hazards model

In this section, we will verify the l_1 consistency and the variable selection consistency of the Dantzig selector numerically. We omit the asymptotic normalities of the estimators obtained after variable selection since these are the consequences of the variable selection consistency and the asymptotic normalities of the maximum partial likelihood estimator (MPLE) and the Bleslow estimator. We consider the following deigns for the simulation studies. For all cases, the sample size $n = 100$, the covariates $Z_1, \ldots, Z_{100}$ are $i.i.d.$ Bernoulli random vectors 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 is generated to have roughly 10% censoring. The tuning parameter γ_n is determined by the algorithm in Section 6.1 with a sufficiently large initial value $\tilde{\gamma}_n = n^{-0.3} \log p$ to obtain the decreasing sequence of tuning parameter by the algorithm for each case. We put $p = 50$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{50}$$

in Case 1, $p = 100$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{100}$$

in Case 2 and $p = 100$ and

$$\beta_0 = (2, 2, 2, -2, -2, 0, \ldots, 0)^\top \in \mathbb{R}^{150}$$

in Case 3. Note that all of these cases satisfy the regularity conditions and the matrix condition Assumption 3.3 theoretically (See Section 3.6.). We apply the Dantzig selector to the proportional hazards model, which can be calculated by the algorithm proposed by Antoniadis et al. (2010) to these data for 1000 times.

For these 1000 estimators, we use the variable selection criterion proposed by in Section 3.3 when $n = 50$ and $n = 100$. Tables 6.5 and 6.6 show that the proportion of successes of the variable selection for all 1000 estimators. We can see that the selection results become better as n becomes larger, which supports our theoretical results.

Tables 6.7 and 6.8 show the l_1 errors of estimators MPLE, the Dantzig selector $\hat{\beta}_n$ and the second estimator $\hat{\beta}_n^{(2)}$ after dimension reduction by using $\hat{T}_n$ when $n = 50$ and $n = 100$ respectively. Note that we cannot construct MPLE when $p \geq n$ since the rank of the optimization problem is deficient as well as the linear regression case. We can observe that for all cases, the Dantzig selectors work better than MPLE and the performance keeps better for higher dimensional case. However, the

l_1 errors become worse when n is larger. In such cases, the estimated values for nonzero coefficients become larger and those for zero coefficients become smaller which may mean that the estimators work well for the variable selection. This may be caused by the fact that the optimization problem is more complicated than the linear regression case since we need some linear approximations to calculate the Dantzig selector for the proportional hazards model which is a nonlinear model.

Though the error of the Dantzig selector becomes larger, we can obtain the second estimator who performs better when n is larger, since the selection method works well when the sample size is larger as we can see in Table 6.6.

Case	$p = 50$	$p = 100$	$p = 150$
$\hat{T}_n = T_0$	71.8 %	39.3 %	48.5 %

Table 6.5: Variable selection results ($n = 50$)

Case	$p = 50$	$p = 100$	$p = 150$
$\hat{T}_n = T_0$	97.3 %	97.9 %	96.0 %

Table 6.6: Variable selection results ($n = 100$)

Case	$p = 50$	$p = 100$	$p = 150$
MPLE	1649.526	NA	NA
$\hat{\beta}_n$	10.12196	9.914827	10.34296
$\hat{\beta}_n^{(2)}$	3.79501	6.414769	6.119591

Table 6.7: l_1 error of estimators ($n = 50$)

Case	$p = 50$	$p = 100$	$p = 150$
MPLE	10.45088	7321.964	NA
$\hat{\beta}_n$	16.34308	16.43016	15.7336
$\hat{\beta}_n^{(2)}$	1.416038	1.462315	1.45552

Table 6.8: l_1 error of estimators ($n = 100$)

Bibliography

Andersen, P. K. and Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, no. 4, 1100-1120. (1982).

Antoniadis, A., Fryzlewicz, P. and Letué, F. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37**, no.4, 531-552. (2010).

Belomestny, D. and Trabs, M. Low-rank diffusion matrix estimation for high-dimensional time-changed Lévy processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **54**, no.3, 1583-1621, (2018).

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, no. 4, 1705-1732. (2009).

Bradic, J. Fan, J. and Jiang, J. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, no. 6, 3092-3120. (2011).

Candés, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, no.6, 2313-2351. (2007).

Cox, D. R. Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser B.* **34** 187-220. (1972).

Fan, Y., Gai, Y. and Zhu, L. Asymtotics of Dantzig selector for a general single-index model. *J. Syst. Sci. Complex.* **29**, no.4, 1123-1144. (2016).

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

Freedman, D. A. On tail probability for martingales. *Ann. Probab.* **3**, no.1, 100-118. (1975).

Fujimori, K. and Nishiyama, Y. The l_q consistency of the Dantzig selector for Cox's proportional hazards model. *J. Statist. Plann. Inference.* **181**, 62-70. (2017a).

Fujimori, K. and Nishiyama, Y. The Dantzig selector for diffusion processes with covariates. *J. Japan Statist. Soc.* **47**, no.1, 59-73. (2017b).

Fujimori, K. Cox's proportional hazards model with a high-dimensional and sparse regression parameter. arXiv:1710.10416[math.ST]. (2017).

Fujimori, K. The Dantzig selector for a linear model of diffusion processes. *To appear in Stat. Inference Stoch. Process.* (2018).

Genon-Catalot, V. and Jacod, J. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincaré Probab. Statist.* **29**, no.1, 119-151. (1993).

Gaiffas, S. and Matulewicz, G. Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process. *Preprint.* (2017).

Gobet, E. LAN property for ergodic diffusions with discrete observations. *Ann. Inst. H. Poincaré Probab. Statist.* **38**, no.5, 711-737. (2002).

Gobet, M. and Matulewicz, G. Parameter estimation of Ornstein-Uhlenbeck process generating a stochastic graph. *Stat. Inference Stoch. Process.* **20**, no.2, 211-235. (2017).

De Gregorio, A. and Iacus, S. M. Adaptive LASSO-type estimation for multivariate diffusion processes. *Econometric Theory* **28**, no.4, 838-860. (2012).

Hjort, N. L. and Pollard, D. Asymptotics for minimizers of convex processes. arXiv:1107.3806. (1993).

Honda, T. and Härdle, W. K. Variable selection in Cox regression model with varying coefficients. *J. Statist. Plann. Inference* **148**, 67-81. (2013).

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C-H. Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41**, no. 3, 1142-1165. (2013).

Kessler, M. Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, no.2, 211-229. (1997).

Masuda, H. Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stochastic Process. Appl.* **117** no.1, 35-56. (2007).

Masuda, H. and Shimizu, Y. Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. *Math. Methods Statist.* **26** no.2, 81-110. (2017).

Periera, J. B. A. and Ibrahimi, M. Support recovery for the drift coefficient of high-dimensional diffusions. *IEEE Trans. Inform. Theory* **60**, no.7, 4026-4049. (2014).

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. High-dimensional Ising model selection using l_1-regularized logistic regression. *Ann. Statist.* **38**, no.3, 1287-1319. (2010).

Sun, T. and Zhang, C-H. Scaled sparse linear regression. *Biometrika.* **99**, no.4, 879-898. (2012).

Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, no.1, 267-288. (1996).

Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385-395. (1997).

Uchida, M. and Yoshida, N. Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Process. Appl.* **122**, no.8, 2885-2924. (2012).

van de Geer, S. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23**, no. 5, 1779-1801. (1995).

van de Geer, S. A. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics, **6**. (2000).

van de Geer, S. A. and Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3**, 1360-1392. (2009).

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

Wang, Y. and Zou, J. Vast volatility matrix estimation for high-frequency financial data. *Ann. Statist.* **38**, no.2, 943-978.(2010).

Wang, H., Li, B. and Leng, C. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, no.3, 671-683. (2009)

Yoshida, N. Estimation for diffusion processes from discrete observation. *J. Multivariate Anal.* **41**, no.2, 220-242. (1992).

Yoshida, N. Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.* **63**, no.3, 431-479. (2011).

Yu, Y. High-dimensional Variable Selection in Cox Model with Generalized Lasso-type Convex Penalty. *Preprint.* (2010).

早稲田大学　博士（理学）　学位申請　研究業績書

氏 名　　藤森 洸　　印

（2019 年　　1 月　現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
1. 共著論文	○The lq consistency of the Dantzig selector for Cox’s proportional hazards model, Journal of Statistical Planning and Inference, 181, p.62-70. Feb. 2017. Kou Fujimori and Yoichi Nishiyama
2. 共著論文	○The Dantzig selector for diffusion processes with covariates, Journal of Japan Statistical Society, 47, no.1, p.59-73. June. 2017, Kou Fujimori and Yoichi Nishiyama
3. 単著論文	○The Dantzig selector for a linear model of diffusion processes, To appear in Statistical Inference for Stochastic Processes, 2018 (掲載決定) Kou Fujimori
4. 国際会議	The lq consistency of the Dantzig selector for Cox’s proportional hazards model, Kou Fujimori and Yoichi Nishiyama Oct. 2016 Hokkaido International Symposium (Hokkaido)
5. 国際会議	The Dantzig selector for diffusion processes with covariates, Kou Fujimori and Yoichi Nishiyama, Feb. 2017, Waseda International Symposium (Tokyo)
6. 国際会議	The Dantzig selector for statistical models of stochastic processes, Kou Fujimori and Yoichi Nishiyama, July 2017, European meeting of Statisticians (Helsinki, Finland)
7. 国際会議	Cox’s proportional hazards model with a high-dimensional and sparse regression parameter, Kou Fujimori, Mar. 2018, Kouchi International Seminar “Recent Developments of Quantile Method, Causality and High Dim Statistics ”(Kouchi)
8. 国際会議	Cox’s proportional hazards model with a high-dimensional and sparse regression parameter, Kou Fujimori, Mar. 2018, International Workshop at Waseda University 2018 (IWAWU2018) Topics in statistical inference and stochastics. (Tokyo)
9. 国際会議	The Dantzig selector for a linear model of diffusion processes, Kou Fujimori, Oct. 2018, Waseda International Symposium, Introduction of General Causality to Various Data & its Innovation of the optimal inference. (Tokyo)

早稲田大学　博士（理学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
10. 講演	Cox 比例ハザードモデルにおける Dantzig selector の一致性, 藤森洸, 西山陽一　2016 年 9 月, 2016 年度統計関連学会連合大会（石川）
11. 講演	Cox 比例ハザードモデルにおける Dantzig selector の一致性, 藤森洸, 西山陽一　2016 年 9 月,　2016 年度日本数学会秋季総合分科会（大阪）
12. 講演	Cox 比例ハザードモデルにおける Dantzig selector の一致性, 藤森洸, 西山陽一 2016 年 12 月　数理統計ひこね 2016（滋賀）
13. ポスター発表	The Dantzig selector for diffusion processes with covariates, 藤森洸、西山陽一　2017 年 3 月　第 11 回日本統計学会春季集会（東京）
14. 講演	The Dantzig selector for diffusion processes with covariates, 藤森洸, 西山陽一　2017 年 3 月　2017 年度日本数学会年会（東京）
15. 講演	高次元・スパースなパラメータを含む拡散過程に対する Dantzig selector, 藤森洸, 西山陽一　2017 年 9 月　2017 年度統計関連学会連合大会（愛知）
16. 講演	The Dantzig selector for high-dimensional linear models of diffusion processes, 藤森洸 2017 年 9 月　2017 年度日本数学会秋季総合分科会（山形）
17. 講演	The Dantzig selector for diffusion processes with high-dimensional parameters, 藤森洸、西山陽一 2017 年 11 月　多様な分野における統計科学の総合的研究（新潟）
18. 講演	The Dantzig selector for Cox’s proportional hazards model, 藤森洸、西山陽一 2017 年 12 月　大規模データの理論と方法論,及び,関連分野への応用（茨城）
19. 講演	Cox’s proportional hazards model with a high-dimensional and sparse regression parameter, 藤森洸 2018 年 3 月 2018 年度日本数学会年会（東京）
20. 講演	The variable selection by the Dantzig selector for Cox’s proportional hazards model, 藤森洸　2018 年 9 月 2018 年度統計関連学会連合大会（東京）
21. 講演	The variable selection consistency by the Dantzig selector for Cox’s proportional hazards model, 藤森洸 2018 年 9 月 2018 年度日本数学会秋季総合分科会（岡山）