Selected Problems for High-Dimensional Data – Quantile and Errors-in-Variables Regressions

by

Seyoung Park

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Xuming He, Co-Chair
Assistant Professor Shuheng Zhou, Co-Chair
Professor Timothy D. Johnson
Professor Kerby A. Shedden

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors Xuming He, Shuheng Zhou and Kerby Shedden whose continuous guidance and support has been the greatest source of encouragement throughout my PhD. I feel very lucky to have them as my advisors, and their warm encouragement and support make me have such a wonderful experience as a PhD student.

I would like to thank my committee member, Professor Timothy Johnson for his valuable time and suggestions on my thesis research.

I also thank Professor Alexandre Belloni and Professor Po-Ling Loh for sharing codes with me.

The research in the thesis is supported by NSF Grants DMS-13-07566, DMS-13-16731, and the Elizabeth Caroline Crosby Research Award from the Advance Program at the University of Michigan.

Last, but not the least, I would like to thank my parents and sister for their unconditional love and endless support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER 1

Introduction

1.1 Multiple Quantile Regression with High Dimensional Covariates

Quantile regression has become a widely used method to evaluate the effect of regressors on the conditional distribution of a response variable (Koenker, 2005). Compared to linear regression analysis, quantile regression is less sensitive to the mis-specification of error distributions and provides more comprehensive information on the relationship between the response variable and the covariates. It is important to study quantile regression in the high-dimensional setting because high-dimensional data arise from many modern application areas such as signal processing and genomics. We focus on the cases where, p, the number of covariates, is greater than n, the sample size.

There has been a line of recent work on variable selection for quantile regression models (Li and Zhu, 2008; Zou and Yuan, 2008a,b; Wu and Liu, 2009). In the high-dimensional setting, the penalization methods with the ℓ_1 penalty (Belloni and Chernozhukov, 2011; Wang, 2013), weighted ℓ_1 penalty (Zheng et al., 2013; Fan et al., 2014a) and smoothly clipped absolute deviation (SCAD) penalty (Wang et al., 2012;

Fan et al., 2014b) have been used to obtain consistent model selection. Belloni and Chernozhukov (2011) establish consistency in parameter estimation with the ℓ_1 penalty. Wang et al. (2012) consider the SCAD penalty, and show that the oracle estimate is one of the local minima of their non-convex optimization problem. Fan et al. (2014a) use the weighted ℓ_1 penalty based on the SCAD penalty function, and establish the model selection consistency and asymptotic normality.

Although the aforementioned work establish nice theoretical properties, empirical evidence shows that the sets of variables selected at two nearby quantiles are often unpleasantly different. The stability of selected variables across quantiles is desirable both for the purpose of interpreting results and for understanding the impact of a particular covariate on the conditional quantile functions. For example, a covariate that is selected at quantiles 0.5 and 0.6 but not at 0.55 would not be much appreciated unless there is a strong reason. The motivation and the main contribution of our work is to show joint modeling across quantiles could lead to stable models. Zou and Yuan (2008a,b), Bang and Jhun (2012), Jiang et al. (2013), Peng et al. (2014), and Volgushev et al. (2014) consider joint quantile regression and provide consistent estimators. He (1997), Dette and Volgushev (2008), Bondell et al. (2010), and Jang and Wang (2015) study non-crossing quantile regression at multiple quantiles. A related piece of work by Zheng et al. (2015) focus on the selection of all the variables that impact one of the quantile functions. In Chapter 2, we aim to identify what impacts each quantile function by allowing subsets of covariates for each quantile to vary smoothly across quantiles.

1.2 Matrix Variate Model

In the second part, we study matrix variate models (Dawid, 1981; Gupta and Varga, 1992) to explain two-way dependencies in data. Recent work on matrix variate models (Dutilleul, 1999; Lu and Zimmerman, 2005; Werner et al., 2008; Efron, 2009; Allen and Tibshirani, 2010; Yin and Li, 2012; Hoff, 2011a) has focused on developing algorithms and theoretical properties for using the Kronecker product covariance models to explain the two-way dependencies in the observational data that arise from diverse areas such as image and signal processing, wireless communication, biology and genomics, and neuroscience. To explain the dependencies in spatiotemporal data (Cressie and Wikle, 2011), Smith et al. (2003) decompose data into functions of time and space. Leng and Tang (2012) consider the Kronecker product model with sparse graphical structure and Zhou (2014) analyzes this sparse Kronecker product model with one matrix variate data. Kalaitzis et al. (2013) use a Kronecker sum model, which is related to our work, to explain the structure of a precision matrix. The Kronecker sums and products of covariance functions describe the additive processes in the context of errors-in-variables models, spatial statistics and spatiotemporal modeling (Carroll et al., 1985; Stefanski, 1985; Hwang, 1986; Iturria et al., 1999; Carroll et al., 2006).

The present work fits a new ensemble of additive covariance models to biological and neuroscience datasets. The baseline Kronecker sum covariance structure has the form of $\Sigma = A \oplus B := A \otimes I_n + I_m \otimes B \in \mathbb{R}^{mn \times mn}$, where $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ are positive definite matrices, and $I_n \in \mathbb{R}^{n \times n}$ is an identity matrix. This Kronecker sum model is motivated by the additive model of $X = X_0 + W \in \mathbb{R}^{m \times n}$, where we

use one covariance component A to describe the covariance among columns of X_0, and the other component B to describe the covariance among rows of W.

The additive covariance model has potential for applications. For example, we may consider brain image data collected over time with additive noise, which yields a grainy appearance. If each row in the data represents the full image at a given time while each column represents a voxel, corresponding to an unique brain region, then the matrix A may show the relationships between brain regions, and the matrix B may uncover the noise pattern over time. If the rows of X are time series measurement at different locations, this model describes the temporal dynamics and the spatial correlation. If the rows of X are repeated trials, with each trial producing a time series, this model describes the temporal dynamics and the trial-wise dependence. In some settings, we may find that one summand in the decomposition $X = X_0 + W$ is primarily "signal" and the other is primarily "noise".

In Chapter 3, we review recent methods for errors-in-variables regression under the Kronecker sum covariance model, and compare Lasso-type and Conic-type estimators used in Rudelson and Zhou (2015). The estimators can be used in node-wise regression procedure to estimate the inverse covariance matrices $\Theta = A^{-1}$ and $\Omega = B^{-1}$ in Chapter 4. We apply the Kronecker sum model to neuromotor control study of hawkmoths (Sponberg et al., 2015), where the data consist of torque measurement (movement) and motor signal. We analyze the temporal and spatial dynamics in the movement data. To assess the goodness of fit of the Kronecker sum model to neuromotor control study, we use a scale-invariant statistic, which shows that the movement data is explained well by the Kronecker sum model. We use measurement error regression techniques from Chapter 3, and analyze the relationship

between neural firing and torque.

CHAPTER 2

Multiple Quantile Regression with High Dimensional Covariates

2.1 Introduction

In this paper, we consider joint quantile regression in the high dimensional setting, where the number of potential covariates as well as the number of quantiles are allowed to increase with n. The penalty we use consists of two components; the first shrinks the magnitudes of the coefficients toward zero; the second controls the rate of changes in coefficients at adjacent quantiles. Both contribute to sparse and stable model selection across quantiles. We propose to minimize the combined penalty in a way that is similar to the Dantzig selector proposed by Candes and Tao (2007). Throughout this paper, the size of set differences of the selected models at adjacent quantiles and the size of the union of the selected covariates across all quantiles of interest will be used to quantify stability of selected models. Moreover, we study a post–selection quantile regression estimate and establish its asymptotic distribution.

The rest of the part is organized as follows. In Section 2.2, we describe the quantile regression model and the proposed Dantzig–type joint quantile regression estimation under consideration. Its theoretical properties are presented in Section

2.3. An implementation of the proposed method is described in Section 2.4, which is shown in Section 2.5 to be consistent in recovering the exact model structure with high probability. Section 2.6 discusses post–selection joint quantile regression and its theoretical properties. The simulation results presented in Section 2.7 demonstrate that the proposed method provides sparse and stable model selection across quantiles. A real data example and some concluding remarks are given in Section 2.8 and Section 2.9, respectively. All technical proofs and the additional simulation study are presented in the Supplementary material.

2.2 Model and Method

Let $X = (x_1, \cdots, x_n)^T$ be an $n \times p$ fixed design matrix and $Y = (y_1, \cdots, y_n)^T \in \mathbb{R}^n$ be an n-dimensional response vector. Consider the following quantile regression model at multiple quantile levels $0 < \tau_1 < \cdots < \tau_{K_n} < 1$, where K_n is allowed to increase with n,

$$Y = X\beta(\tau_k) + \epsilon^{(k)} \quad (k = 1, \ldots, K_n), \tag{2.1}$$

where $\beta(\tau_k) \in \mathbb{R}^p$ is a τ_k-th quantile coefficient vector in the sense that ${x_i}^T\beta(\tau_k)$ is the τ_k-th quantile of y_i evaluated at x_i, which will be called the conditional quantile of y_i given x_i for the sake of convenience, and $\epsilon^{(k)} = (\epsilon_1^{(k)}, \cdots, \epsilon_n^{(k)})^T$ is an n-dimensional vector with mutually independent elements and

$$\mathbb{P}\left[\epsilon_i^{(k)} \leq 0 \mid x_i\right] = \tau_k \quad (i = 1, \ldots, n;\ k = 1, \ldots, K_n).$$

In the special case where we have a linear model with i.i.d. errors, $\epsilon^{(k)}$ would depend on k only through a location shift. Our model assumes that the conditional quantile of y_i given x_i is linear at each τ_k, but no distributional assumptions are made on

$\epsilon^{(k)}$. Let $T^{(k)}$ be the support set of $\beta(\tau_k)$ and $B^{(k)}$ be the indices where the quantile coefficients at the τ_k-th quantile are different from those at the τ_{k-1}-th quantile; that is,

$$\begin{aligned} T^{(k)} &= \{j \in \{1,\ldots,p\} : \beta_j(\tau_k) \neq 0\} \quad (k = 1,\ldots,K_n), \\ B^{(k)} &= \{j \in \{1,\ldots,p\} : \beta_j(\tau_k) \neq \beta_j(\tau_{k-1})\} \quad (k = 2,\ldots,K_n). \end{aligned} \tag{2.2}$$

Let $s_k = |T^{(k)}|$ denote the sparsity level of the model for the τ_k-th quantile. We consider a high dimensional sparse model with $\max(n, K_n) = o(p)$, where $p = o\left(\exp(n^b)\right)$ for some constant $b > 0$. Let $s_0 := \max_k s_k$. Our goal is to recover support sets $T^{(k)}$ $(k = 1,\ldots,K_n)$, $B^{(k)}$ $(k = 2,\ldots,K_n)$, and coefficient vectors $\beta(\tau_k)$ $(k = 1,\ldots,K_n)$.

Let $w^{(k)}$ $(k = 1,\ldots,K_n)$ and $v^{(k)}$ $(k = 2,\ldots,K_n)$ be p-dimensional vectors of nonnegative weights, λ be a regularization parameter, and $r_k > 0$ for $k = 1,\ldots,K_n$ be constraint bounds to be chosen. We consider the following convex optimization problem:

$$\min_{\mathcal{B}=[\beta^{(1)},\cdots,\beta^{(K_n)}]\in\mathbb{R}^{p\times K_n}} \sum_{k=1}^{K_n}\sum_{j=1}^{p} w_j^{(k)}|\beta_j^{(k)}| + \lambda \sum_{k=2}^{K_n}\sum_{j=1}^{p} v_j^{(k)} \frac{|\beta_j^{(k)} - \beta_j^{(k-1)}|}{|\tau_k - \tau_{k-1}|}, \tag{2.3}$$

$$\text{s.t. } \forall k, \quad \beta^{(k)} \in \mathcal{R}^{(k)}(r_k) = \left\{ \beta \in \mathbb{R}^p : \ \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - {x_i}^T\beta) \leq r_k \right\}, \tag{2.4}$$

where $\rho_\tau(t) = t(\tau - 1\{t \leq 0\})$ is the τ-th quantile loss function (Koenker and Basset, 1978).

Let $\widehat{\mathcal{B}} = [\widehat{\beta}^{(1)},\cdots,\widehat{\beta}^{(K)}]$ be any optimum of (2.3) and (2.4), as an estimator of the true parameter $\mathcal{B}^o = [\beta(\tau_1),\ldots,\beta(\tau_{K_n})]$. In (2.3), two types of penalties are required to simultaneously provide sparse and stable models. The first one, a sparsity penalty, aims to obtain a sparse model. The second one, a weighted total variation

penalty (WTV), controls the rate of change in quantile coefficients functions; see the related work by Rudin et al. (1992) and Tibshirani et al. (2005). The feasible set of the optimization problem (2.3) is non-empty for any choices of positive r_ks because there always exists $\beta \in \mathbb{R}^p$ satisfying $Y = X\beta$ provided that the column space of X spans $\mathbb{R}^n$.

Throughout the paper, it is to be understood that the design matrix X is normalized to have column ℓ_2 norm $\sqrt{n}$, and non-stochastic. The quantities p, s_0 and K_n depend on the sample size n. Given a vector $\delta = (\delta_1, \cdots, \delta_p)^T \in \mathbb{R}^p$ and a set of indices $S \subset \{1, \ldots, p\}$, denote by $\delta_S \in \mathbb{R}^p$ the vector with the jth component $\delta_{S,j} = \delta_j I(j \in S)$. Let $\|\delta\|_0$, $\|\delta\|_\infty$ and $\|\delta\|_q$ for any positive integer q be the number of nonzero components, the maximum absolute value and the ℓ_q norm of δ, respectively. Let S^c be the complement set of S. For p-dimensional vectors $\beta^{(1)}, \cdots, \beta^{(K)}$, let $[\beta^{(1)}, \cdots, \beta^{(K)}]$ be the $p \times K$ matrix whose kth column is $\beta^{(k)}$ for $k = 1, \ldots, K$. For two numbers a and b, we also use the notation $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$ and $x_+ = xI(x > 0)$ for $x \in \mathbb{R}$. For sequences $\{a_n\}$ and $\{\zeta_n\}$, we write $a_n = O(\zeta_n)$ to mean that $a_n \leq C\zeta_n$ for a universal constant $C > 0$. Similarly, $a_n = \Omega(\zeta_n)$ when $a_n \geq C'\zeta_n$ for some universal constant $C' > 0$.

We also summarize notations used in the theorems in Table 2.1.

Table 2.1: Notations used in the Chapter

Parameters	Definitions
$\lambda =$	A regularization parameter in (2.3)
$d_{\min} =$	$\min_{k\geq 2} \lvert\tau_k - \tau_{k-1}\rvert$
$W_0 =$	$\max_k \left\| w^{(k)}{}_{\left(T^{(k)}\right)^c} \right\|_\infty \bigvee \max_{k\geq 2} \left\| v^{(k)}{}_{\left(B^{(k)}\right)^c} \right\|_\infty$
$W_1 =$	$\max_k \left\| w^{(k)}{}_{T^{(k)}} \right\|_\infty \bigvee \max_{k\geq 2} \left\| v^{(k)}{}_{B^{(k)}} \right\|_\infty$
$W_2 =$	$\min_k \min_{j\in\{T^{(k)}\}^c} w_j^{(k)} \bigwedge \min_{k\geq 2} \min_{j\in\{B^{(k)}\}^c} v_j^{(k)}$
$c_0 =$	$(d_{\min} W_1 + 2\lambda W) \,/\, (d_{\min} W_2 - 2\lambda W)$
$\psi_\lambda =$	$(d_{\min} + 2\lambda) \,/\, (d_{\min} - 2\lambda)$
$M_n =$	$\max_i \left\| x_{i,\cup_k T^{(k)}} \right\|_\infty$
$d_0 =$	$\lvert T^{(1)}\rvert + \sum_{k=2}^K \lvert B^{(k)} \setminus T^{(k)}\rvert$
$\mathbb{M}(S) =$	Median of a sequence of real number S

2.3 Theoretical Properties

We first define the following cone constraint: for any set $J \subset \{1, \cdots, p\}$ and any positive number c,

$$C(J, c) = \{x \in \mathbb{R}^p \mid x \neq 0,\ \|x_{J^c}\|_1 \leq c\|x_J\|_1\}\,.$$

Define a restricted eigenvalue (RE) condition (Bickel et al., 2009; van de Geer and Buhlmann, 2009): for any integer $0 < s < p$ and any positive number $c > 0$, RE(s, c) means

$$k^2(s, c) := \min_{\substack{J\subseteq\{1,\ldots,p\},\\ |J|\leq s}} \min_{\delta\in C(J,c)} \frac{\delta^T X^T X \delta}{n\|\delta_J\|_2^2} > 0, \tag{2.5}$$

which is imposed on the $p \times p$ sample covariance matrix $X^T X/n$. The RE condition is needed to guarantee consistency of the Lasso and Dantzig selectors (Bickel et al., 2009). This condition also implies that the gram matrix $X^T X/n$ behaves like a positive definite matrix over the cone $C(J, c)$ for any J such that $|J| \leq s$.

Similarly, we introduce a restricted nonlinear impact (RNI) condition, as in Belloni and Chernozhukov (2011): For any integer $0 < s < p$ and any positive number $c > 0$, RNI(s, c) means

$$q(s,c) := \min_{\substack{J\subseteq\{1,\ldots,p\},\\ |J|\le s}} \min_{\delta\in C(J,c)} \frac{\|X\delta\|_2^3}{n^{1/2}\|X\delta\|_3^3} > 0, \tag{2.6}$$

which controls the norm $\|X\delta\|_3$ by $\|X\delta\|_2$ over the cone $C(J, c)$ for any J such that $|J| \le s$. RNI(s, c) can be equivalently written as for $\delta \in C(J, c)$,

$$\left(\frac{1}{n}\sum_{i=1}^{n}|x_i^T\delta|^2\right)^3 \ge q^2(s,c)\left(\frac{1}{n}\sum_{i=1}^{n}|x_i^T\delta|^3\right)^2,$$

which implies that the third sample moment is controlled by the second sample moment. This condition is necessary to control the quantile regression objective function by quadratic terms (Belloni and Chernozhukov, 2011).

Condition 2.3.1. *[On the conditional density] For each $i = 1, \ldots, n$, let $f_i(\cdot)$ denote the probability density function of y_i given x_i. The function $f_i(\cdot)$ has a continuous derivative $f_i'(\cdot)$. For each i, $f_i(\cdot) \le \overline{f}$, $|f_i'(\cdot)| \le \overline{f}$ and $\min_k f_i\left({x_i}^T\beta(\tau_k)\right) \ge \underline{f}$ for some constants $\overline{f}, \underline{f} > 0$.*

Condition 2.3.2. *[On the weights] Let W_0 and W_1 be the maximum weight imposed on the zero components and nonzero components, respectively, and W_2 be the minimum weight imposed on zero components. The weights satisfy*

$$\frac{W_2}{W_0 \vee W_1} \ge \frac{2.5\lambda}{\min_k |\tau_k - \tau_{k-1}|}.$$

Condition 2.3.3. *[On the growth rate of the sparsity] The maximal sparsity s_0 satisfies the growth condition, $s_0 \log p = o(n)$.*

Condition 2.3.1 is the same as Condition D.1 in Belloni and Chernozhukov (2011). For the location model and the location-scale model, Belloni and Chernozhukov (2011, Lemmas 1 and 2) analyze the sufficient conditions of Condition 2.3.1 by specifying the values of $\overline{f}$, $\underline{f}$. Condition 2.3.2 imposes a lower bound on $W_2/(W_0 \vee W_1)$. This condition implies that W_2 must not be too small, which means that zero components must be penalized in the optimization problem (2.3) and (2.4). In Sections 2.3 and 2.4, we show that W_0, W_1 and W_2 can be constructed from an appropriate initial estimator, and W_0 and W_1 are upper bounded and W_2 is lower bounded by some constants. Condition 2.3.3 is necessary for the consistency of our estimators.

Remark 2.3.1. Note that the regular adaptive lasso weights are used in Jiang et al. (2013), where $w_j^{(k)} = 1/|\widetilde{\beta}_j^{(k)}|^q$ and $v_j^{(k)} = 1/|\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}|^q$ with an initial estimate $\widetilde{\beta}^{(k)}$ at quantile level τ_k and $q > 0$. Condition 2.3.2 is not guaranteed for this weight because $W_0 \vee W_1$ can have any arbitrary large number. This motivates us to use a different type of weights, and in Section 3 the derivative of the SCAD penalty function is used for calculating the weights $w_j^{(k)}$ and $v_j^{(k)}$ that satisfy Condition 2.3.2 with high probability, as can be seen in the proof of Theorem 2.5.1.

Throughout this section, for any $\eta \geq 0$, let $\mathbb{E}_\eta$ be the event

$$\mathbb{E}_\eta = \left\{ 0 \leq r_k - \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - {x_i}^T \beta(\tau_k)\right) \leq \eta \quad (k = 1, \ldots, K) \right\}. \tag{2.7}$$

The following theorem shows the consistency of the proposed estimator $\widehat{\mathcal{B}}$.

Theorem 2.3.1. *Suppose that Conditions 2.3.1-2.3.2,* $\mathrm{RE}(2s_0, c_0)$ *and* $\mathrm{RNI}(2s_0, c_0)$ *hold. Let* $\widehat{\mathcal{B}} = [\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K)}]$ *be the solution to* (2.3) *and* (2.4). *Let* $\eta_n = o(1)$ *be any sequence of positive numbers with* $0 \leq \eta_n < 9\underline{f}^3 q^2(2s_0, c_0)/(32\overline{f}^2)$. *Then we have*

with probability at least $1 - 1/n - \mathbb{P}(\mathbb{E}_{\eta_n}^c)$,

$$\max_k \|\widehat{\beta}^{(k)} - \beta(\tau_k)\|_2 \leq \xi_1 \sqrt{\frac{s_0 \log p}{n} + \eta_n}, \tag{2.8}$$

$$\sum_{k=1}^{K} \|\widehat{\beta}^{(k)}_{\{T^{(k)}\}^c}\|_1 \bigvee \lambda \sum_{k=2}^{K} \left\| \frac{\{\widehat{\beta}^{(k)} - \widehat{\beta}^{(k-1)}\}_{\{B^{(k)}\}^c}}{|\tau_k - \tau_{k-1}|} \right\|_1 \leq \xi_3 s_0 K \sqrt{\frac{\log p}{n}} + \xi_3 \sqrt{s_0} K \sqrt{\eta_n}, \tag{2.9}$$

where for some absolute constant $C_1 > 0$,

$$\xi_1 = \frac{2(1+c_0)^2}{k(2s_0, c_0)\sqrt{\underline{f}}} \left\{ 1 + \frac{2C_1}{k(s_0, c_0)} \right\} \quad \text{and} \quad \xi_3 = \xi_1 \frac{W_1}{W_2}. \tag{2.10}$$

The upper bound in (2.8) implies that the estimates $\widehat{\beta}^{(k)}$ for $k = 1, \cdots K_n$ are uniformly consistent when $\eta_n = o(1)$ and $n = \Omega(s_0 \log p)$. The upper bound in (2.8) has two components, where the first component $\sqrt{s_0 \log p/n}$ is within a factor of $\sqrt{\log p}$ of the oracle rate, and the second component $\sqrt{\eta_n}$ characterizes the bias induced by the use of the feasible region $\mathcal{R}^{(k)}(r_k)$ in (2.7). To obtain the consistency rate $\sqrt{s_0 \log p/n}$ for $\widehat{\beta}^{(k)}$ in (2.8), which is an expected bound for high dimensional models (Belloni and Chernozhukov, 2011; Fan et al., 2014a; Zheng et al., 2015), $\eta_n = O(s_0 \log p/n)$ is required. By using a consistent initial estimate, we can choose such η_n with r_k such that the event $\mathbb{E}_{\eta_n}$ holds with a high probability; See (2.16) for details.

As can be seen in (2.8), as η_n increases, the estimation error bound is larger while the probability $\mathbb{P}(\mathbb{E}_{\eta_n}^c)$ becomes smaller. The optimal r_k is $\frac{1}{n}\sum_{i=1}^n \rho_{\tau_k}\left(y_i - x_i{}^T\beta(\tau_k)\right)$, which provides the fastest convergence rate. Therefore, using r_k near this optimal value in (2.4) is a key part of our implementations. We use a proper initial estimate of $\beta^{(k)}$ to estimate the optimal value r_k in (2.11).

Inequality (2.9) shows that the ℓ_1 norm of the quantile coefficients estimates for inactive predictors (with true zero coefficients) converges to zero provided that $W_1/W_2 = o(1)$, $K_n^2\eta_n = o(1)$ and $n = \Omega(K_n^2 s_0 \log p)$. Moreover, the ℓ_1 norm is deceasing as W_1/W_2 becomes smaller, which implies that choosing smaller weights W_1 and larger weights W_2 would improve the rate of convergence, which is consistent with the idea used in adaptive Lasso.

Later in Theorem 2.5.2, we will discuss exact model structure selection by using (2.9) with an additional beta–min condition.

Remark 2.3.2. The quantity ξ_1 in (2.10) depends on n by the term $k(2s_0, c_0)$ and $k(s_0, c_0)$. Consider a simple case that $\tau_k - \tau_{k-1} = 1/K_n$ for all k, and $w_j^{(k)} = v_j^{(k)} = 1$ for all k and j. Then $W_0 = W_1 = W_2 = 1$, and Condition 2.3.2 reduces to $\lambda \leq 2/(5K_n)$. If $\lambda = 2/(5K_n)$, then the condition of c_0 in Theorem 2.3.1 is equivalent to $c_0 \geq 9$. Specifically, if $c_0 = 9$, then ξ_1 is less than some universal constant given that $k(2s_0, 9)$ is lower bounded by some universal constant.

Remark 2.3.3. Our formulation (2.3) enables us to use r_k as a tunning parameter, and the scale of r_k is more interpretable than a tuning parameter in the Lagrangian formulation. Letting the weights of the quantile loss functions for all quantile levels to be equal in the dual problem is proposed by Jiang et al. (2013) under the fixed p setting, which includes fewer regularization parameters. But it is not clear whether model selection consistency holds for such estimators in the high dimensional setting. Moreover, our empirical work shows that, in terms of model selection, our proposed method outperforms the implementation based on the equal weights in the dual problem. See Section 2.7 for details.

2.4 Implementation

We provide a specific realization for the Dantzig–type joint quantile regression introduced in Section 2.2. This procedure involves the derivative of the SCAD penalty function (Fan and Li, 2001):

$$P_\zeta(x) = I(x \le \zeta) + \frac{(3.7\zeta - x)_+}{2.7\zeta} I(x > \zeta)$$

with a regularization parameter $\zeta \ge 0$. We now specify the multi-step procedure.

Step 1. Obtain initial estimates. We obtain initial estimates following Belloni and Chernozhukov (2011). Let $\widetilde{\lambda} = 1.1\ \Pi(0.9)$ be a regularization parameter, where $\Pi(0.9)$ is defined in Remark 2.4.1,

$$\widetilde{\beta}^{(k)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - {x_i}^T \beta \right) + \widetilde{\lambda} \|\beta\|_1 \quad (k = 1, \ldots, K_n). \tag{2.11}$$

Step 2. Solve the Dantzig-type optimization. To solve the optimization (2.3) and (2.4), we use the following specifications.

Step 2a: For the parameters in the objective function (2.3), we use the following specifications. Let $\widetilde{s} = \max_k \|\widetilde{\beta}^{(k)}\|_0$.

$$\zeta_n = 0.1\sqrt{\widetilde{s} \log p / n}, \tag{2.12}$$

$$w_j^{(k)} = P_{\zeta_n}\left(|\widetilde{\beta}_j^{(k)}| \right) \ (k = 1, \ldots, K_n), \tag{2.13}$$

$$v_j^{(k)} = P_{\zeta_n}\left(|\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}| \right) \ (k = 2, \ldots, K_n), \tag{2.14}$$

$$\lambda = 0.4 \min_{k \ge 2} |\tau_k - \tau_{k-1}|. \tag{2.15}$$

Step 2b Let $h > 0$ denote a scaling parameter to be chosen and $\Lambda_k^{(h)} \ge 0$ $(k = 1, \ldots, K_n)$ be regularization parameters taken to be $\Lambda_k^{(h)} = \mathbb{M}(R_k)h$, where $\mathbb{M}$ is

defined in Table 1 and $R_k = \left\{|y_i - {x_i}^T\widetilde{\beta}^{(k)}| : i = 1, \ldots, n\right\}$. For the parameter r_k in the constraint (2.4), we use

$$r_k^{(h)} = \frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_k}\left(y_i - {x_i}^T\widetilde{\beta}^{(k)}\right) + \Lambda_k^{(h)}\frac{\widetilde{s}\log p}{n} \quad (k = 1, \ldots, K_n). \tag{2.16}$$

Step 3. Choose h. We use 5-fold cross validation to minimize the sum of the quantile loss functions over all quantiles of interest. More specifically, we randomly split the data into five roughly equal parts $X^{(1)}, \cdots, X^{(5)} \in \mathbb{R}^{[n/5]\times p}$ and $y^{(1)}, \cdots, y^{(5)} \in \mathbb{R}^{[n/5]\times 1}$, respectively. For $t = 1, \cdots, 5$, let $X^{(t)} = \left[x_1^{(t)}, \cdots, x_{[n/5]}^{(t)}\right]^T$. Let $\widehat{\beta}_t^{(k)}(h)$ $(k = 1, \cdots, K_n)$ be the solution to the (2.3) and (2.4) following Step 1 and Step 2 for the data X and Y excluding the tth fold. Let the CV score function

$$\text{score}(h) := \sum_{t=1}^{5}\sum_{k=1}^{K_n}\sum_{i=1}^{[n/5]}\rho_{\tau_k}\left(y_i^{(t)} - (x_i^{(t)})^T\widehat{\beta}_t^{(k)}(h)\right).$$

We choose h^0 from the set $S := \{0.01, 0.02, \cdots, 4\}$ that minimizes the score, that is,

$$h^o := \arg\min_{h\in S}\ \text{score}(h).$$

The Dantzig-type estimate $\widehat{\beta}^{(k)}$ is the solution to (2.3) and (2.4) using the aforementioned specifications with $h = h^o$, $\Lambda_k := \Lambda_k^{(h^o)}$, and $r_k := r_k^{(h^0)}$.

In Step 2 (b), $\Lambda_k^{(h)}$ plays the role of scaling to achieve scale equivariance of the method. It is obvious that those choices of the regularization parameters do not give the best results for any given models, but they lead to good empirical results in a variety of settings and could help understand how the proposed Dantzig–type penalization performs with reasonable choices of these tuning parameters.

Remark 2.4.1. Following Belloni and Chernozhukov (2011), define

$$\Pi := \max_{1\le k\le K_n}\max_{1\le j\le p}\frac{1}{n}\left|\sum_{i=1}^{n}\frac{x_{ij}\left(\tau_k - I(u_i \le \tau_k)\right)}{\sqrt{\tau_k(1-\tau_k)}}\right|,$$

where $u_1, \cdots, u_n$ are independent and identically distributed from the uniform distribution on $(0, 1)$ and independent of x_is, and x_{ij} is the jth component of the design x_i for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $\Pi(0.9)$ be the 0.9th quantile of Π that can be computed using simulated Π. As seen in Step 1, we use $\widetilde{\lambda} = 1.1\ \Pi(0.9)$, where the constant factor 1.1 differs from the recommendation made in Belloni and Chernozhukov (2011), giving us initial estimates with low false negative rates.

2.5 Theoretical Properties (continued)

Let $\widehat{\mathcal{B}} = [\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K_n)}]$ be any optimum of (2.3) and (2.4), where $w_j^{(k)}$s, $v_j^{(k)}$s and r_ks are defined in (2.13), (2.14) and (2.16). Define an event for the initial estimates $\widetilde{\beta}^{(k)}$s for $k = 1, \ldots, K_n$ as follows: for some positive constants C_2, C_3 and C_4,

$$E_1 = \left\{ \widetilde{\lambda} \leq C_2 \sqrt{\frac{\log p}{n}},\ \max_k \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \leq C_3 \sqrt{\frac{s_0 \log p}{n}},\ \max_k \|\widetilde{\beta}^{(k)}\|_0 \leq C_4 s_0 \right\}, \tag{2.17}$$

Denote by $\gamma_n := \mathbb{P}(E_1^c)$ the probability that the event E_1 does not occur.

Belloni and Chernozhukov (2011) prove that their estimators and the corresponding regularization parameters as stated in (2.11) satisfy condition E_1 with probability close to 1.

For theoretical properties of the estimates detailed in Section 2.4, we assume the following conditions.

Condition 2.5.1. *[On the regularization parameters] Assume that*

$$\min_k \Lambda_k \geq 6\sqrt{C_4 + 1} C_3 \quad \text{and} \quad \zeta_n \geq 2C_3 \sqrt{s_0 \log p / n}.$$

Condition 2.5.2. *[On the non-zero coefficients] The following beta-min conditions hold for some positive constants C_5 and C_6,*

$$\min_k \min_{j\in T^{(k)}} |\beta_j(\tau_k)| > C_5\sqrt{\frac{s_0\log p}{n}}, \tag{2.18}$$

$$\min_{k\geq 2} \min_{j\in B^{(k)}} \frac{|\beta_j(\tau_k)-\beta_j(\tau_{k-1})|}{|\tau_k-\tau_{k-1}|} > C_6K_n\sqrt{\frac{s_0\log p}{n}}, \tag{2.19}$$

where we assume $n=\Omega(K^2 s_0\log p)$.

Our multi-step Dantzig–type joint quantile estimator $\widehat{\mathcal{B}}$ is consistent as shown in the following theorems.

Theorem 2.5.1. *Suppose Conditions 2.3.1,2.3.3,2.5.1,* $\mathrm{RE}(2s_0,\psi_\lambda)$ *and* $\mathrm{RNI}(2s_0,\psi_\lambda)$ *hold. Then with probability at least $1-2/n-\gamma_n$, $\widehat{\mathcal{B}}$ satisfies*

$$\max_k \|\widehat{\beta}^{(k)}-\beta(\tau_k)\|_2 \leq \xi_2\sqrt{\frac{s_0\log p}{n}},$$

where for some absolute constant $C>0$, $\xi_2=\frac{C}{k(2s_0,\psi_\lambda)\sqrt{\underline{f}}}\sqrt{1+\max_k\Lambda_k}$.

The following Theorem 2.5.2 shows that $\widehat{\mathcal{B}}$ recovers the exact model structure under appropriate conditions.

Theorem 2.5.2. *Suppose that the conditions of Theorem 2.5.1 and Condition 2.5.2 hold. Then*

$$\mathbb{P}\left(\left\{\widehat{T}^{(k)}=T^{(k)} \text{ and } \widehat{B}^{(k)}=B^{(k)} \text{ for all } k\right\}\right)\geq 1-\frac{2}{n}-\gamma_n.$$

Theorem 2.5.1 follows from (2.8) in Theorem 2.3.1 and shows that our multi-step Dantzig–type joint quantile estimator $\widehat{\mathcal{B}}$ is consistent when $n=\Omega(s_0\log p)$ under appropriate conditions. Theorem 2.5.1 requires the lower bound of Λ_k for the feasible

regions (2.4) to include the true parameter $\mathcal{B}^o$ with high probability. In simulations, our estimator still worked quite well even if Λ_k is set to zero so Condition 2.5.1 is violated.

Theorem 2.5.2 implies that the true parameter $\mathcal{B}^o$ belongs to the set of optimal solutions with high probability and $\widehat{\mathcal{B}}$ recovers the true model structure with high probability, which also satisfies the exact model selection property (Zhao and Yu, 2006; Wainwright, M., 2009; Fan et al., 2014a).

Remark 2.5.1. The beta-min condition (2.18) imposes a lower bound of the nonzero coefficients. While Condition (2.18) has been studied in high dimensional analysis to establish the exact model selection property (Meinshausen and Bühlmann, 2006; van de Geer et al., 2011; Bühlmann and van de Geer, 2011), the beta-min condition (nonzero rate of change in interquantile coefficients) that provides a lower bound on the nonzero interquantile differences rate has not been considered elsewhere.

The beta–min condition (2.19) can be demonstrated by the following example. For simplicity, we consider equally-spaced quantile levels τ_k $(k = 1, \ldots, K_n)$ with $\tau_k - \tau_{k-1} \asymp 1/K_n$. Consider a location–scale model, as used in Example 2.7.2 in Section 2.7, $y_i = {x_i}^T\beta + {x_i}^T r\epsilon_i$, where the design x_i and the vector $r \in \mathbb{R}^p$ have nonnegative components with $x_i^T r > 0$ for all i. Then (2.19) holds as long as the components of r satisfy $r_j 1\{r_j \neq 0\} \succ K_n\sqrt{s_0 \log p/n}$ $(j = 1, \ldots, p)$, where r_j is the jth component of r.

2.6 Post–Selection Joint Quantile Regression

We consider a post–selection joint quantile regression that minimizes the sum of quantile loss functions over all quantiles of interest based on the model struc-

ture $\widehat{T}^{(k)}$ $(k = 1, \ldots, K_n)$ and $\widehat{B}^{(k)}$ $(k = 2, \ldots, K_n)$ of the multi-step Dantzig–type joint quantile estimator $\widehat{\mathcal{B}} = [\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K)}]$ as described in Section 2.4. The post–selection joint quantile estimator (POST JQR) denoted by $\widehat{\mathcal{B}}^{po}$ is a minimizer of

$$\min_{\mathcal{B}=[\beta^{(1)},\cdots,\beta^{(K_n)}]\in G} \sum_k \sum_i \rho_{\tau_k}\left(y_i - {x_i}^T \beta^{(k)}\right), \quad \text{where} \tag{2.20}$$

$$G = \left\{ \mathcal{B} = [\beta^{(1)}, \ldots, \beta^{(K_n)}] \in \mathbb{R}^{p \times K_n} : {\beta^{(k)}}_{\{\widehat{T}^{(k)}\}^c} = 0,\ {\beta^{(k)}}_{\{\widehat{B}^{(k)}\}^c} = {\beta^{(k-1)}}_{\{\widehat{B}^{(k)}\}^c} \right\}$$

is a set of matrices whose induced model structure is the same as the structure of $\widehat{\mathcal{B}}$. Throughout this section, we assume that $\widehat{T}^{(k)} = T^{(k)}$ $(k = 1, \ldots, K_n)$ and $\widehat{B}^{(k)} = B^{(k)}$ $(k = 2, \ldots, K_n)$, which holds with probability tending to 1. As can be seen in the proof of Theorem 2.6.1 in the Supplementary material, there is a one-to-one mapping T between G and $\mathbb{R}^{d_0}$, where d_0 is the effective dimension of the parameter for the selected model as defined in Table 4.1. In other words, the set $G \subset \mathbb{R}^{p \times K_n}$ in (2.20) can be embedded in $\mathbb{R}^{d_0}$. We use $T(\widehat{\mathcal{B}}^{po})$ to estimate $T(\mathcal{B}^o)$, which is a d_0–dimensional vector that consists of the active components of $\mathcal{B}^o$.

To establish the theoretical properties of $T(\widehat{\mathcal{B}}^{po})$, we redefine POST JQR. As defined in the proof of Theorem 2.6.1 in the Supplementary material, there exist new design variables $z_i^{(k)}$ $(i = 1, \ldots, n;\ k = 1, \ldots, K_n)$ such that

$$T(\widehat{\mathcal{B}}^{po}) = \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_k \sum_i \rho_{\tau_k}\left(y_i - (z_i^{(k)})^T \beta\right). \tag{2.21}$$

Now to establish the asymptotic convergence rate and asymptotic normality of $T(\widehat{\mathcal{B}}^{po})$, we use the following sparse eigenvalue condition: For $0 < s < p$,

$$\text{Sparse}(s):\ \phi(s) = \max_{\|\delta\|_0 \le s} \frac{\|X\delta\|_2^2}{n\|\delta\|_2^2} < \infty. \tag{2.22}$$

Sparse(s) means that the maximal s-sparse eigenvalue of the gram matrix $X^T X/n$ is bounded by some constant (Rudelson and Zhou, 2013; Belloni et al., 2015; Zheng

et al., 2015). We use the following conditions to show the theoretical properties of the estimator.

Condition 2.6.1(a). *[On the sample size]*

$$n = \Omega\left(d_0 s_0^3 (\log n)^6 \vee M_n^4 d_0 (\log n)^2\right).$$

Condition 2.6.1(b). $n = \Omega\left(d_0^5 s_0^3 (\log n)^6 \vee M_n^2 d_0^3 s_0\right).$

Condition 2.6.1(a) is used to show the asymptotic oracle consistency of the estimator in Theorem 2.6.1, and Condition 2.6.1(b) is required for showing the asymptotic normality of the estimator in Theorem 2.6.2. These conditions involve d_0, s_0, M_n and n. If the entries in x_i are uniformly bounded, and d_0 and s_0 grow slowly with n, Condition 2.6.1(a) is quite mild. The POST JQR enjoys the asymptotic oracle consistency rate as follows:

Theorem 2.6.1. *Suppose the conditions of Theorem 2.5.2 together with Condition 2.6.1(a) and* Sparse(s_0). *Then*

$$\|T(\widehat{\mathcal{B}}^{po}) - T(\mathcal{B}^o)\|_2 = O_p\left(\sqrt{d_0/n}\right). \tag{2.23}$$

Theorem 2.6.2. *Suppose that the conditions of Theorem 2.6.1 and Condition 2.6.1(b) hold. Then, for any sequence of vectors $\alpha_n \in R^{d_0}$ with $\|\alpha_n\|_2 = 1$, $T(\widehat{\mathcal{B}}^{po})$ is asymptotically normal,*

$$\alpha_n^T \sqrt{n} ({A_n}^{-1} B_n {A_n}^{-1})^{-\frac{1}{2}} \left(T(\widehat{\mathcal{B}}^{po}) - T(\mathcal{B}^o)\right) \to^d N(0,1),$$

where

$$A_n = \sum_{k=1}^{K_n} \sum_{i=1}^{n} \frac{1}{n} f_i\left({x_i}^T \beta(\tau_k)\right) z_i^{(k)} \left(z_i^{(k)}\right)^T,$$

$$B_n = \sum_{i=1}^{n} \sum_{k,k'=1,\ldots,K_n} \frac{1}{n} z_i^{(k)} \left(z_i^{(k')}\right)^T \left(\tau_k \wedge \tau_{k'} - \tau_k \tau_{k'}\right),$$

with $z_i^{(k)}$ given in the proof of Theorem 2.6.1 in the Supplementary material.

Theorem 2.6.2 provides sufficient conditions for the asymptotic normality of POST JQR, which relies on the exact model structure property as defined in Theorem 2.5.2. This property is typically fragile without beta-min condition, and is not uniformly valid (Leeb and Pötscher, 2005). Leeb and Pötscher (2003) and Belloni et al. (2015) considered the post-model-selection estimator conditional on selecting an incorrect model, and established the uniform asymptotic distribution of the estimator. Establishing an asymptotic distribution without beta-min condition in our setting will also be of interest in a follow-up work.

2.7 Numerical Studies

Our optimization problem (2.3) is equivalent to a linear programming problem with the aid of slack variables, and can be solved by existing optimization packages in a way that is similar to the problem of Jiang et al. (2013). For the other estimators, we use $\widetilde{\beta}^{(k)}$ as an initial estimate at τ_k-th quantile. More specifically, ALasso at τ_k is

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - {x_i}^T \beta) + \lambda_{ad,k} \sum_{j=1}^{p} |\beta_j| / |\widetilde{\beta}_j^{(k)}|,$$

where $\lambda_{ad,k}$ is the regularization parameter to be chosen by 5-fold cross validation to minimize the τ_k-th quantile loss function, and FAL (Jiang et al., 2013) finds

$$\begin{aligned} \underset{[\beta^{(1)},\cdots,\beta^{(K_n)}] \in \mathbb{R}^{p \times K}}{\arg\min} \quad & \frac{1}{n} \sum_{k=1}^{K_n} \sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - {x_i}^T \beta^{(k)}\right) \\ + \quad & \lambda_a \left(\sum_{k=1}^{K_n} \sum_{j=1}^{p} w_j^{(k)} |\beta_j^{(k)}| + \sum_{k=2}^{K_n} \sum_{j=1}^{p} v_j^{(k)} |\beta_j^{(k)} - \beta_j^{(k-1)}| \right), \end{aligned}$$

where $w_j^{(k)} = 1/|\widetilde{\beta}_j^{(k)}|$ and ${v^{(k)}}_j = 1/|\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}|$ and λ_a is the regularization parameter to be chosen by 5-fold cross validation to minimize the sum of quantile loss functions over all quantiles of interest. Our proposed estimator Dantzig is described in Section 2.4.

To assess the performances of the competing methods, the following performance measures were calculated based on 100 Monte Carlo replications.

1. "FP_k", the number of false positives in the selected model at τ_k, i.e., $|\widehat{T}^{(k)} \backslash T^{(k)}|$;
2. "FN_k", the number of false negatives in the selected model at τ_k, i.e., $|T^{(k)} \backslash \widehat{T}^{(k)}|$;
3. "SD_k", the size of the set difference of the selected models for adjacent quantile levels, τ_k and τ_{k-1}, i.e., $|\widehat{T}^{(k)} \triangle \widehat{T}^{(k-1)}|$ for $k = 2, \ldots, K_n$;
4. "FP_U", the number of false positives in the union of the selected models across all quantile levels, i.e., $|\cup_k \widehat{T}^{(k)} \backslash \cup_k T^{(k)}|$;
5. "FN_U", the number of false negatives in the union of the selected models across all quantile levels, i.e., $|\cup_k T^{(k)} \backslash \cup_k \widehat{T}^{(k)}|$.

In the following examples, we consider five different models, a location model, a location-scale model and a random coefficient model.

Example 2.7.1. *Consider the linear regression model with* $(n, p, K_n, s_0) = (100, 500, 5, 6)$ *and* $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = (0.30, 0.40, 0.50, 0.60, 0.70)$*:*

$$y_i = {x_i}^T \beta + \epsilon_i, \quad \beta = (1.0, 0.8, 0.0, 0.9, 0.5, 0.0, 0.3, 0.7, 0.0, \cdots, 0.0)^T,$$

where ϵ_i*s are independent and identically distributed from the standard normal distribution and independent of* x_i*s. The regressors are* $x_i = (1, z_i)^T$*, where* $z_{ij} \sim N(0, \Sigma)$

is generated from the autoregressive model, AR(1), with correlation 0.5*, that is,* $\Sigma_{(i,j)} = 0.5^{|i-j|}$.

Example 2.7.2. *Consider the following location-scale model with* $(n, p, K_n, s_0) = (100, 500, 5, 7)$ *and* $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = (0.30, 0.40, 0.50, 0.60, 0.70)$*:*

$$y_i = x_1 + 0.8x_2 + 0.9x_4 + 0.5x_5 + 0.3x_7 + 0.75x_8 + (0.5x_2 + x_3 + 0.5x_8)\epsilon_i,$$

where ϵ_i*s are independent and identically distributed from the standard normal distribution and independent of* x_i*s. The regressors are generated in two steps, following Wang et al. (2012). First generate* $\widetilde{x}_{ij} \sim N(0, \Sigma_x)$ *from the AR(1) model, with correlation* 0.5*, and then* $x_{ij} = \Phi(\widetilde{x}_{ij})$ $(j = 2, 3, 8)$ *and* $x_{ij} = \widetilde{x}_{ij}$ $(j \neq 2, 3, 8)$*, where* Φ *is the cumulative distribution function of the standard normal distribution.*

Example 2.7.3. *Consider the random coefficient model with* $(n, p, K_n, s_0) = (100, 500, 5, 6)$ *and* $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = (0.70, 0.75, 0.80, 0.85, 0.90)$*:*

$$y_i = {x_i}^T \beta(u_i), \quad \beta(u_i) = (\beta_1(u_i), \cdots, \beta_p(u_i))^T,$$

where $u_1, \cdots, u_n$ *are independent and identically distributed from the uniform distribution on* $(0, 1)$ *and independent of* x_i*, and* $\beta_1(u) = 1.7 + \Phi^{-1}(u)$*,* $\beta_2(u) = 0.35$*,* $\beta_3(u) = 3(u - 0.8)_+$*,* $\beta_5(u) = 0.5 + 0.5 \times 2^u$*,* $\beta_6(u) = 0.5 + u$*,* $\beta_{10}(u) = 0.4 + \sqrt{u}$ *and* $\beta_j(u) = 0$ $(j \neq 1, 2, 3, 5, 6, 10)$*. The regressors are generated in the same way as in Example 2.7.2.*

Example 2.7.4. *Consider the model, which is same as Example 1 in the main paper except that* ϵ_i*s follow the standard Cauchy distribution.*

Example 2.7.5. *Consider the model, which is same as Example 1 in the main paper except that* ϵ_i*s follow the standard Laplace distribution.*

Figure 2.1 shows the performance measures defined in Subsection 2.7.1 for Examples 2.7.1–2.7.3. The first, second and third rows correspond to Example 2.7.1, Example 2.7.2 and Example 2.7.3, respectively. Each row consists of three sub-figures. The first and the second sub-figures show the number of false positives and the number of false negatives for each of the five quantile levels, respectively, which explains the quality and the sparsity of the selected models. The last sub-figure shows the size of the set differences of the selected models at adjacent quantile levels, and the number of false positives and false negatives of the union of the selected covariates over the five quantile levels, which explains the stability of the model. Across all figures, the largest standard errors for the false positives, the false negatives and the size of set differences are less than 0.9, 0.1 and 0.5, respectively.

As seen in Figures 2.1 and 2.2, Dantzig includes smaller number of false positives with more false negatives compared to the other methods. But this increase in false negatives is relatively small considering the decrease in false positives. Dantzig has a smaller size of set difference for two neighboring quantiles, and fewer false positives than other methods for the union of the selected variables across the five quantile levels. This indicates that Dantzig shares many common variables across different quantiles, and provides more stable models. Overall, at each quantile, Dantzig provides sparser model than other competitors in all the examples. In terms of stability of the selected models across quantiles, Dantzig outperforms the others.

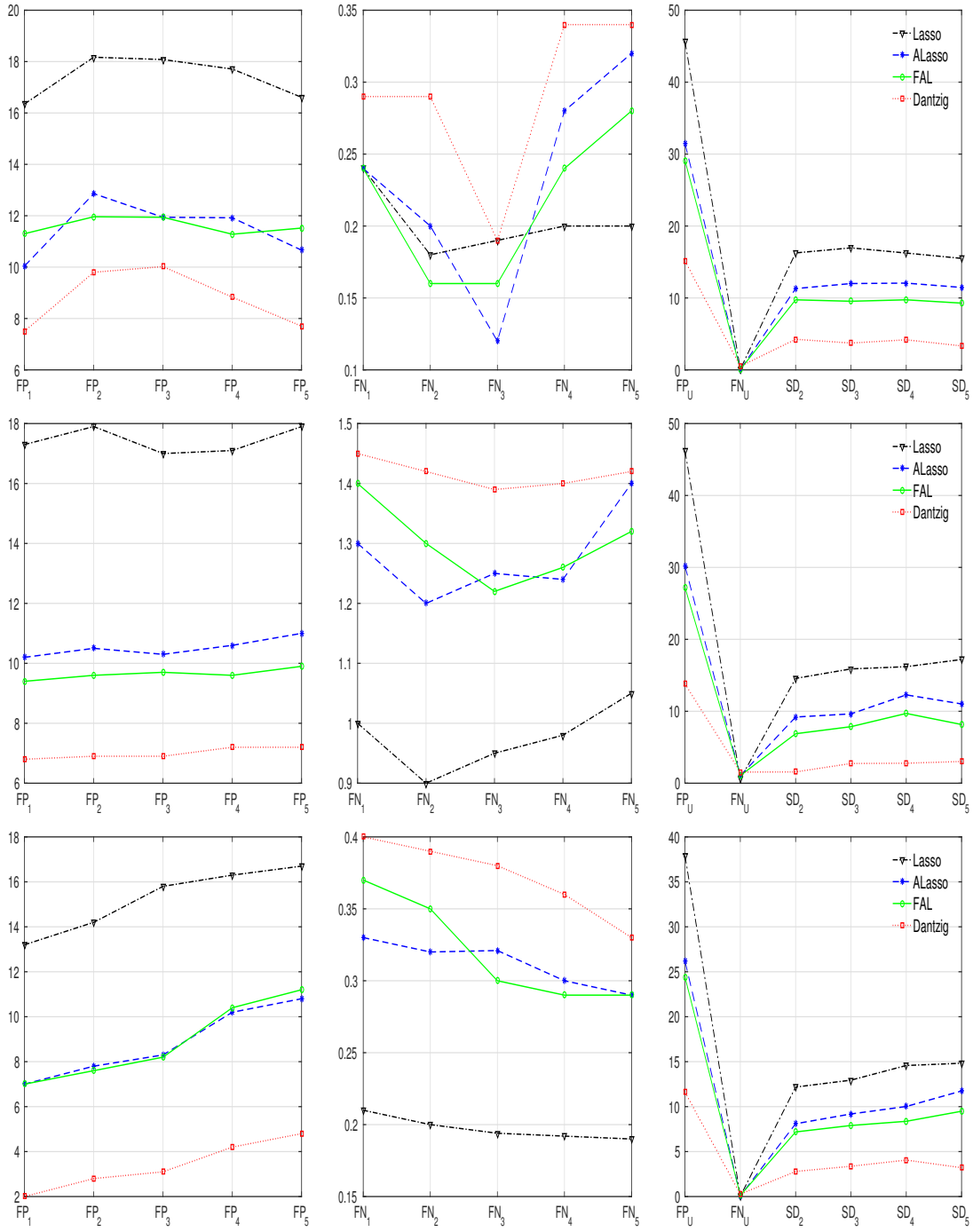

Figure 2.1: Results for Example 2.7.1 (top), 2.7.2 (middle) and 2.7.3 (below): Include false positives(left), false negatives (middle) and the stability measures (right). Four competing procedures are evaluated: Lasso, ALasso, FAL and Dantzig.

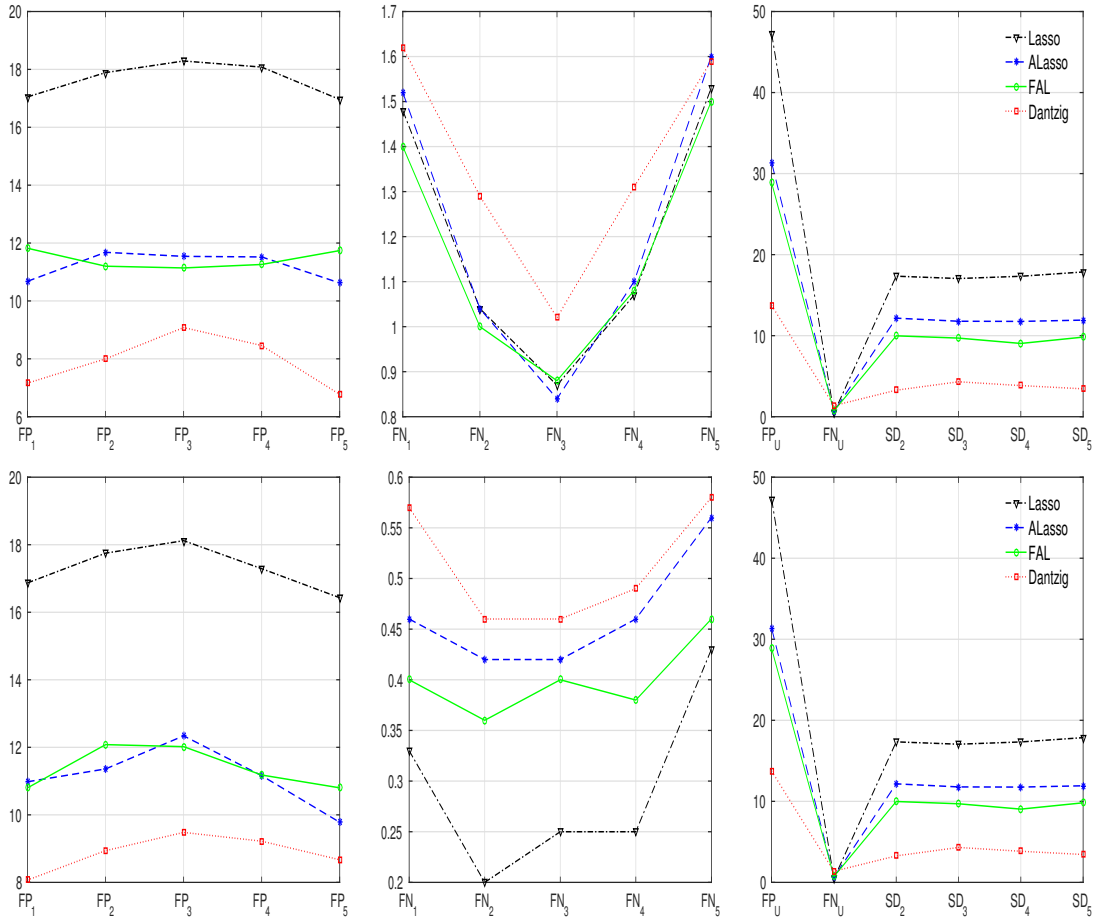

Figure 2.2: Results for Example 2.7.4 (top) and 2.7.5 (below): Include false positives (left), false negatives (middle) and the stability measures (right). Four competing procedures are evaluated: Lasso, ALasso, FAL and Dantzig.

2.8 Application

We consider the proposed Dantzig–type joint quantile regression method in an application to a genetic data set used in Scheetz et al. (2006). This data set consists of the expression values of 31042 probe sets for 120 rats. As in Huang et al. (2008), Kim et al. (2008) and Wang et al. (2012), we are interested in finding genes that are related to gene TRIM32, which is known for causing Bardet-Biedl syndrome.

The model selection approach is applied to 300 probe sets that pass an initial screening. See Huang et al. (2008) for details of the screening steps. We apply

Dantzig, Lasso, ALasso and FAL, which are defined in Section 2.7, and SCAD (Wang et al., 2012) on these 300 probe sets ($p = 300$) with 120 rats ($n = 120$). SCAD is a single quantile regression method, which uses the SCAD penalty function to penalize quantile coefficients. We consider two sets of five quantile levels $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$ as $(0.48, 0.49, 0.50, 0.51, 0.52)$ and $(0.81, 0.82, 0.83, 0.84, 0.85)$, representing interests in the middle and the upper tail of the distribution of the target gene expressions. To select a tuning parameter for each method, we use 5-fold cross validation. See Subsections 2.4.1 and 2.7.1 for details.

We report the number of nonzero coefficients ("SIZ") selected by each method at each quantile level. The size of the set difference of the selected models at adjacent quantile levels ("DIF") and the size of the union of the selected covariates over five quantile levels ("TOT") are considered to investigate the stability of the selected models. As can be seen in Table 2.2, the Dantzig–type estimators, Dantzig and Dantzig0, consistently provide sparser model than other methods. Dantzig also provides the most stable model as we expected.

We also randomly divide the data set into a training set and a test set; the training set includes 80 rats and the test set includes 40 rats. We estimate the models with each method, by using the training set, and record "SIZ", "DIF" and "TOT". The prediction error ("PRE") is calculated over the test set as the quantile loss for each quantile level τ_k. We repeat this random experiments 100 times and report the average value of "SIZ", "DIF", "TOT" and "PRE" over the 100 repetitions for each method in Table 2.3. As seen in Table 2.3, all of the six methods are similar in terms of prediction error. In terms of the sparsity of the selected models, all of the methods except Lasso are similar. But in terms of the stability of models, Dantzig

outperforms other competitors as we expected. In Table 2.3, the largest standard errors for the columns corresponding to SIZ, DIF, PRE and TOT are less than 0.7, 0.3, 0.05 and 1.2, respectively.

Table 2.2: Performance results of whole dataset

Method	SIZ	DIF	TOT	Method	SIZ	DIF	TOT
Lasso (0.48)	37			Lasso (0.81)	37		
Lasso (0.49)	38	9		Lasso (0.82)	41	8	
Lasso (0.50)	35	15		Lasso (0.83)	39	8	
Lasso (0.51)	36	5		Lasso(0.84)	36	7	
Lasso (0.52)	37	3	45	Lasso (0.85)	38	4	49
SCAD (0.48)	24			SCAD (0.81)	20		
SCAD (0.49)	24	0		SCAD (0.82)	25	9	
SCAD (0.50)	20	6		SCAD (0.83)	16	9	
SCAD (0.51)	14	7		SCAD (0.84)	29	13	
SCAD (0.52)	18	5	25	SCAD (0.85)	27	4	35
ALasso (0.48)	27			ALasso (0.81)	25		
ALasso (0.49)	17	14		ALasso (0.82)	24	3	
ALasso (0.50)	20	7		ALasso (0.83)	22	4	
ALasso (0.51)	14	6		ALasso (0.84)	21	3	
ALasso (0.52)	15	1	29	ALasso (0.85)	28	7	34
FAL (0.48)	21			FAL (0.81)	25		
FAL (0.49)	21	0		FAL (0.82)	25	1	
FAL (0.50)	22	2		FAL (0.83)	26	2	
FAL (0.51)	21	3		FAL (0.84)	25	2	
FAL (0.52)	21	2	25	FAL (0.85)	25	2	26
Dantzig (0.48)	21			Dantzig (0.81)	21		
Dantzig (0.49)	19	2		Dantzig (0.82)	20	1	
Dantzig (0.50)	20	1		Dantzig (0.83)	21	1	
Dantzig (0.51)	21	3		Dantzig (0.84)	22	2	
Dantzig (0.52)	20	1	22	Dantzig (0.85)	21	1	24

Table 2.3: Performance results of 100 random partitions of the data

Method	SIZ	DIF	PRE	TOT	Method	SIZ	DIF	PRE	TOT
Lasso (0.48)	30.94		1.79		Lasso (0.81)	32.94		1.33	
Lasso (0.49)	31.10	3.35	1.79		Lasso (0.82)	33.04	4.22	1.30	
Lasso (0.50)	31.76	4.38	1.78		Lasso (0.83)	33.00	6.36	1.26	
Lasso (0.51)	31.92	4.66	1.78		Lasso (0.84)	32.88	4.20	1.23	
Lasso (0.52)	32.60	4.78	1.78	37.73	Lasso (0.85)	32.78	4.34	1.21	40.28
SCAD (0.48)	22.04		1.79		SCAD (0.81)	20.90		1.32	
SCAD (0.49)	22.82	5.10	1.78		SCAD (0.82)	20.32	6.46	1.27	
SCAD (0.50)	21.86	5.44	1.78		SCAD (0.83)	21.02	6.62	1.27	
SCAD (0.51)	21.38	4.52	1.78		SCAD (0.84)	22.10	6.12	1.23	
SCAD (0.52)	21.66	5.40	1.79	28.74	SCAD (0.85)	20.50	5.16	1.20	28.86
ALasso (0.48)	19.96		1.82		ALasso (0.81)	19.98		1.34	
ALasso (0.49)	19.70	2.98	1.79		ALasso (0.82)	19.22	4.04	1.31	
ALasso (0.50)	19.32	3.46	1.80		ALasso (0.83)	20.04	5.34	1.26	
ALasso (0.51)	19.08	3.40	1.80		ALasso (0.84)	19.92	3.36	1.25	
ALasso (0.52)	19.64	3.76	1.80	24.56	ALasso (0.85)	19.44	3.60	1.21	25.78
FAL (0.48)	19.75		1.85		FAL (0.81)	20.95		1.37	
FAL (0.49)	20.70	1.28	1.82		FAL (0.82)	21.71	2.34	1.32	
FAL (0.50)	20.72	1.94	1.88		FAL (0.83)	20.33	3.90	1.27	
FAL (0.51)	20.18	2.40	1.82		FAL (0.84)	20.18	2.76	1.25	
FAL (0.52)	19.94	2.59	1.83	23.63	FAL (0.85)	21.74	2.20	1.22	24.55
Dantzig (0.48)	20.20		1.84		Dantzig (0.81)	21.94		1.33	
Dantzig (0.49)	20.06	0.98	1.84		Dantzig (0.82)	21.72	1.02	1.31	
Dantzig (0.50)	19.98	1.82	1.82		Dantzig (0.83)	21.98	2.70	1.27	
Dantzig (0.51)	20.70	2.01	1.81		Dantzig (0.84)	21.60	1.78	1.25	
Dantzig (0.52)	21.02	2.52	1.80	22.90	Dantzig (0.85)	21.42	1.18	1.22	23.86

2.9 Conclusion

Model selection stability across quantile levels adds credibility and interpretability of the selected models in applications. If the selected models vary significantly from one quantile to the next when the quantile levels used are very close to each other, it could be an undesirable feature of model selection. The proposed Dantzig–type approach leads to a much more stable selection without a noticeable sacrifice on the prediction error. We adopt a Dantzig–type optimization problem and establish the uniform non-asymptotic error bounds and model selection consistency

under appropriate conditions. By using the selected model structure, we also study post–selection joint quantile regression and establish its asymptotic distributions. The simulation study and real data analysis show that the proposed method consistently provides sparse and stable models, and reduces the noisy component in model selection at single quantile levels for both homogeneous and heterogeneous cases.

2.10 Supplementary Material

Let F_i denote the conditional distribution of y_i given x_i for $i = 1, \ldots, n$, that is $F_i(x) = \mathbb{P}[y_i \leq x \mid x_i]$ for all $x \in R$. Define the diagonal matrices

$$H_k = \text{diag}\left[f_1\left({x_1}^T\beta(\tau_k)\right), \cdots, f_n\left({x_n}^T\beta(\tau_k)\right)\right] \quad (k = 1, \ldots, K_n),$$

where $f_1, \cdots, f_n$ are defined in Condition 2.3.1 of the main paper. Then for any vector $\delta \in \mathbb{R}^p$, we define an intrinsic norm as in Belloni and Chernozhukov (2011),

$$\|\delta\|_{k,2} = \sqrt{\delta^T \frac{X^T H_k X}{n} \delta} \quad (k = 1, \ldots, K_n). \tag{2.24}$$

For any positive constant c and the sets $T^{(k)}$ $(k = 1, \ldots, K_n)$ defined in (2.2), let

$$A^{(k)}(c) = \left\{\delta : \delta \neq 0, \delta \in \mathbb{R}^p,\ \|\delta_{\{T^{(k)}\}^c}\|_1 \leq c\|\delta_{T^{(k)}}\|_1\right\}.$$

Define the function as follows: for $k = 1, \ldots, K_n$,

$$\mathbb{Q}_n^{(k)}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - {x_i}^T\beta),$$

where the subdifferential of $\mathbb{Q}_n^{(k)}(\beta)$ at β is the following set of vectors (Wang et al., 2012):

$$\partial\mathbb{Q}_n^{(k)}(\beta) = \left\{\delta \in \mathbb{R}^p \mid \delta_j = -\frac{\tau}{n}\sum_i x_{ij} I(y_i > {x_i}^T\beta) + \frac{1-\tau}{n}\sum_i x_{ij} I(y_i < {x_i}^T\beta) - \frac{1}{n}\sum_i x_{ij} v_i \right\},$$

where x_{ij} is the jth component of x_i, and $v_i = 0$ if $y_i \neq {x_i}^T\beta$ and $v_i \in [\tau - 1, \tau]$ otherwise. For simplicity, for any $\mathcal{B} = [\beta^{(1)}, \cdots, \beta^{(K)}] \in \mathbb{R}^{p \times K_n}$, let

$$G(\mathcal{B}) = \sum_{k=1}^{K_n} \sum_{j=1}^{p} w_j^{(k)} |\beta_j^{(k)}| + \lambda \sum_{k=2}^{K_n} \frac{1}{|\tau_k - \tau_{k-1}|} \sum_{j=1}^{p} v_j^{(k)} |\beta_j^{(k)} - \beta_j^{(k-1)}|, \tag{2.25}$$

which is the objective function of our optimization problem as defined in (2.3), where $w^{(k)}$ $(k = 1, \ldots, K_n)$ and $v^{(k)}$ $(k = 2, \ldots, K_n)$ are p-dimensional wight vectors.

For any square matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be the maximum eigenvalue and the minimum eigenvalue of A, respectively.

Preliminary Results

The following Lemma 2.10.1 controls the empirical error over all vectors in $A^{(k)}(c_0)$ for all $k = 1, \ldots, K_n$ and is analogous to Lemma 5 of the Belloni and Chernozhukov (2011).

Lemma 2.10.1. *Let c_0 and $t_1, \cdots, t_{K_n}$ be positive numbers. Suppose Condition 2.3.1 and RE($2s_0, c_0$) hold. Let*

$$\widetilde{\mathbb{Q}}^{(k)}(v) = E\left[\mathbb{Q}_n^{(k)}\{\beta(\tau_k) + v\} - \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\}\right] - \mathbb{Q}_n^{(k)}\{\beta(\tau_k) + v\} + \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\}.$$

for any $v \in \mathbb{R}^p$. Then we have

$$\mathbb{P}\left\{\sup_{v \in A^{(k)}(c_0), \|v\|_{k,2} \leq t_k} \left|\widetilde{\mathbb{Q}}^{(k)}(v)\right| > C_1 \frac{1 + c_0}{k(s_0, c_0)} t_k \sqrt{\frac{s_0 \log p}{n}} \ (k = 1, \ldots, K_n)\right\} \leq \frac{1}{n} \tag{2.26}$$

for some absolute constant $C_1 > 0$.

Proof of Theorem 2.3.1

We begin by providing several lemmas that will be used for the proof.

Lemma 2.10.2. *Let c_0 be a positive number. Suppose* $\mathrm{RE}(2s_0, c_0)$ *holds. Then we have for all $k = 1, \ldots, K_n$,*

$$\|\delta\|_1 \leq \sqrt{s_0}\frac{1+c_0}{\sqrt{\underline{f}}k(s_0,c_0)}\|\delta\|_{k,2}, \quad \|\delta\|_2 \leq \frac{1+c_0}{\sqrt{\underline{f}}k(2s_0,c_0)}\|\delta\|_{k,2}$$

for all $\delta \in A^{(k)}(c_0)$.

The following Lemma 2.10.3 is a fixed design version of (3.7) in Belloni and Chernozhukov (2011), which provides the lower bound of the difference of the expected values of quantile loss function over all vectors in the cone $A^{(k)}(c_0)$ for all $k = 1, \ldots, K_n$.

Lemma 2.10.3. *Let c_0 be a positive number. Suppose Condition 2.3.1 and RNI($2s_0, c_0$) hold. Then we have for all $k = 1, \ldots, K_n$,*

$$E\left[\mathbb{Q}_n^{(k)}\{\beta(\tau_k)+\delta\} - \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\}\right] \geq \frac{3\underline{f}^{3/2}q(2s_0,c_0)}{8\bar{f}}\|\delta\|_{k,2} \wedge \frac{1}{4}\|\delta\|_{k,2}^2 \tag{2.27}$$

for all $\delta \in A^{(k)}(c_0)$.

The following Lemma 2.10.4 shows that $\widehat{\beta}^{(k)} - \beta(\tau_k)$ is included in the specific cone for all k.

Lemma 2.10.4. *Let η be any positive number. Let $[\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K_n)}]$ be an optimum of (2.3) and (2.4) in the main paper. Suppose Condition 2.3.2 holds. Then on event $\mathbb{E}_\eta$ defined in* (2.7) *in the main paper, we have*

$$\widehat{\beta}^{(k)} - \beta(\tau_k) \in A^{(k)}\left(\frac{d_{\min}W_1 + 2\lambda(W_0 \vee W_1)}{d_{\min}W_2 - 2\lambda(W_0 \vee W_1)}\right) \quad (k = 1, \ldots, K_n),$$

where W_1 and W_2 are defined in Table 2.1.

We fix any c_0 and η, which satisfy the conditions in Theorem 2.3.1. Let $\delta_k = \widehat{\beta}^{(k)} - \beta(\tau_k)$ $(k = 1, \ldots, K_n)$. Let E_2 be the event

$$\sup_{v \in A^{(k)}(c_0), \|v\|_{k,2} \le \|\delta_k\|_{k,2}} \left| \widetilde{\mathbb{Q}}^{(k)}(v) \right| \le C_1 \frac{1 + c_0}{k(s_0, c_0)} \|\delta_k\|_{k,2} \sqrt{\frac{s_0 \log p}{n}} \ (k = 1, \ldots, K_n),$$

where C_1 is the constant used in Lemma 2.10.1 and $P(E_2) \ge 1 - 1/n$ by Lemma 2.10.1.

Proof of (5) in Theorem 2.3.1. Throughout the proof, we assume $E_2 \cap \mathbb{E}_{\eta_n}$ holds. Lemma 2.10.4 implies that δ_k is in $A^{(k)}(c_0)$ for $k = 1, \ldots, K_n$. By Lemma 2.10.3, we have that for $k = 1, \ldots, K_n$,

$$\begin{aligned}
& \frac{\|\delta_k\|_{k,2}^2}{4} \wedge \frac{3\underline{f}^{3/2} q(2s_0, c_0)}{8\bar{f}} \|\delta_k\|_{k,2} \\
\le \ & \mathbb{Q}^{(k)}\{\widehat{\beta}^{(k)}\} - \mathbb{Q}^{(k)}\{\beta(\tau_k)\} \\
= \ & \mathbb{Q}_n^{(k)}\{\widehat{\beta}^{(k)}\} - \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + [\mathbb{Q}^{(k)}\{\widehat{\beta}^{(k)}\} - \mathbb{Q}^{(k)}\{\beta(\tau_k)\} - \mathbb{Q}_n^{(k)}\{\widehat{\beta}^{(k)}\} + \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\}] \\
\le \ & \eta_n + [\mathbb{Q}^{(k)}\{\widehat{\beta}^{(k)}\} - \mathbb{Q}^{(k)}\{\beta(\tau_k)\} - \mathbb{Q}_n^{(k)}\{\widehat{\beta}^{(k)}\} + \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\}] \\
\le \ & \eta_n + C_1 \frac{1 + c_0}{k(s_0, c_0)} \sqrt{\frac{s_0 \log p}{n}} \|\delta_k\|_{k,2}, \qquad (2.28)
\end{aligned}$$

where C_1 is the absolute constant stated in Lemma 2.10.1.

Notice that (2.28) implies that the first term in the left hand side must be less than the second term. Suppose otherwise, that is, $\|\delta_k\|_{k,2} \ge 3\underline{f}^{3/2} q(2s_0, c_0)/(2\bar{f})$. Then we have

$$\frac{3\underline{f}^{3/2} q(2s_0, c_0)}{8\bar{f}} \|\delta_k\|_{k,2} \le \eta_n + C_1 \frac{1 + c_0}{k(s_0, c_0)} \sqrt{\frac{s_0 \log p}{n}} \|\delta_k\|_{k,2},$$

which contradicts the assumption that $0 \le \eta_n < 9\underline{f}^3 q^2(2s_0, c_0)/(32\bar{f}^2)$, Thus, we conclude

$$\frac{\|\delta_k\|_{k,2}^2}{4} \le \eta + C_1 \frac{1 + c_0}{k(s_0, c_0)} \sqrt{\frac{s_0 \log p}{n}} \|\delta_k\|_{k,2} \quad (k = 1, \ldots, K_n),$$

which yields

$$\|\delta_k\|_{k,2} \leq 4C_1 \frac{1+c_0}{k(s_0,c_0)} \sqrt{\frac{s_0 \log p}{n}} + 2\sqrt{\eta_n} \quad (k = 1, \ldots, K_n). \tag{2.29}$$

By Lemma 2.10.2 and (2.29), we have

$$\|\delta_k\|_2 \leq 4C_1 \frac{(1+c_0)^2}{k(2s_0,c_0)k(s_0,c_0)\sqrt{\underline{f}}} \sqrt{\frac{s_0 \log p}{n}} + 2\frac{1+c_0}{k(2s_0,c_0)\sqrt{\underline{f}}}\sqrt{\eta_n} \quad (k = 1, \ldots, K_n),$$

which implies

$$\|\widehat{\beta}^{(k)} - \beta(\tau_k)\|_2 \leq \frac{(1+c_0)^2}{k(2s_0,c_0)\sqrt{\underline{f}}} \left\{2 + \frac{4C_1}{k(s_0,c_0)}\right\} \sqrt{\frac{s_0 \log p}{n} + \eta_n} = \xi_1 \sqrt{\frac{s_0 \log p}{n} + \eta_n}, \tag{2.30}$$

where

$$\xi_1 = \frac{(1+c_0)^2}{k(2s_0,c_0)\sqrt{\underline{f}}} \left\{2 + \frac{4C_1}{k(s_0,c_0)}\right\}.$$

This completes the proof. □

Proof of (2.8) *in Theorem 2.3.1.* Throughout the proof, we assume $E_2 \cap \mathbb{E}_{\eta_n}$ holds. The main idea is to compare the objective functions of our optimization problem as stated in (2.3) at $\widehat{\mathcal{B}}$ and $\mathcal{B}^o$. Since $\mathcal{B}^o$ is feasible, $G(\widehat{\mathcal{B}})$ must not be greater than $G(\mathcal{B}^o)$, where the function $G(\cdot)$ is defined in (2.25). So we have

$$\begin{aligned}
0 \;\leq\; & G(\mathcal{B}^o) - G(\widehat{\mathcal{B}}) = \sum_{k=1}^{K} \sum_{j \in T^{(k)}} w_j^{(k)} |\beta_j(\tau_k)| + \sum_{k=2}^{K} \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j \in B^{(k)}} v_j^{(k)} |\beta_j(\tau_k) - \beta_j(\tau_{k-1})| \\
& - \sum_{k=1}^{K_n} \sum_{j \in T^{(k)}} w_j^{(k)} |\widehat{\beta}_j^{(k)}| + \sum_{k=2}^{K_n} \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j \in B^{(k)}} v_j^{(k)} |\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k-1)}| + \sum_{k=1}^{K_n} \sum_{j \in \{T^{(k)}\}^c} w_j^{(k)} |\widehat{\beta}_j^{(k)}| \\
& + \sum_{k=2}^{K_n} \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j \in \{B^{(k)}\}^c} v_j^{(k)} |\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k-1)}|.
\end{aligned}$$

By the triangle inequality with the definition of W_1 and (2.30), the above inequality

implies

$$
\begin{aligned}
&\sum_{k=1}^{K_n}\sum_{j\in\{T^{(k)}\}^c} w_j^{(k)}|\widehat{\beta}_j^{(k)}| + \sum_{k=2}^{K_n}\frac{\lambda}{|\tau_k-\tau_{k-1}|}\sum_{j\in\{B^{(k)}\}^c} v_j^{(k)}|\widehat{\beta}_j^{(k)}-\widehat{\beta}_j^{(k-1)}| \\
\leq\ & \sum_{k=1}^{K_n}\sum_{j\in T^{(k)}} w_j^{(k)}|\widehat{\beta}_j^{(k)}-\beta_j(\tau_k)| + \sum_{k=2}^{K_n}\frac{\lambda}{|\tau_k-\tau_{k-1}|}\sum_{j\in B^{(k)}} v_j^{(k)}|\widehat{\beta}_j^{(k)}-\widehat{\beta}_j^{(k-1)}-\beta_j(\tau_k)+\beta_j(\tau_{k-1})| \\
\leq\ & W_1\sum_{k=1}^{K_n}\|\{\widehat{\beta}^{(k)}-\beta(\tau_k)\}_{T^{(k)}}\|_1 + W_1\sum_{k=2}^{K_n}\frac{\lambda}{|\tau_k-\tau_{k-1}|}\|\{\widehat{\beta}^{(k)}-\beta(\tau_k)\}_{B^{(k)}}\|_1 \\
&+W_1\sum_{k=2}^{K_n}\frac{\lambda}{|\tau_k-\tau_{k-1}|}\|\{\widehat{\beta}^{(k-1)}-\beta(\tau_{k-1})\}_{B^{(k)}}\|_1 \\
\leq\ & W_1\sqrt{K_n}\sqrt{s_0}\sqrt{\sum_{k=1}^{K_n}\|\{\widehat{\beta}^{(k)}-\beta(\tau_k)\}\|_2^2} \qquad (2.31) \\
+\ & 2W_1\frac{\lambda}{\min_{k\geq 2}|\tau_k-\tau_{k-1}|}\sqrt{K_n}\sqrt{2s_0}\sqrt{\sum_{k=1}^{K_n}\|\{\widehat{\beta}^{(k)}-\beta(\tau_k)\}\|_2^2} \\
\leq\ & \xi_1(W_1+\sqrt{2}W_1)\sqrt{s_0}K_n\sqrt{\frac{s_0\log p}{n}+\eta_n}, \qquad (2.32)
\end{aligned}
$$

where the third inequality comes from the Cauchy-Schwarz inequality with $|T^{(k)}| \leq s_0$ and $|B^{(k)}| \leq 2s_0$. Applying (2.32) and the definition of W_2, we complete the proof.

□

Proofs of Theorem 2.5.1

We begin by providing the following lemmas that will be used for the proof of Theorem 2.5.1. Lemma 2.10.5 is only used to show Lemma 2.10.6.

Lemma 2.10.5. *For an $n\times p$ design matrix $X=(x_1,\cdots,x_n)^T$, which is normalized to have column ℓ_2 norm $\sqrt{n}$, we have with probability at least $1-1/n$,*

$$\max_k \Big\| \sum_{i=1}^n x_i[\tau_k - I\{y_i \leq {x_i}^T\beta(\tau_k)\}]/n\Big\|_\infty \leq 3\sqrt{\frac{\log p}{n}}. \qquad (2.33)$$

Recall on event E_1 defined in (2.17) in the matin paper, we have for all k,

$$\widetilde{\lambda} \leq C_2\sqrt{\log p/n}, \quad \|\widetilde{\beta}^{(k)}-\beta(\tau_k)\|_2 \leq C_3\sqrt{s_0\log p/n}, \quad \|\widetilde{\beta}^{(k)}\|_0 \leq C_4 s_0. \qquad (2.34)$$

Now we have the following lemma, which implies that we can find proper η_n on event E_1.

Lemma 2.10.6. *Let E_1 be the event as defined in* (2.17) *in the main paper. Suppose the conditions of Theorem 2.5.1 hold. Then we have $P(\mathbb{E}_{\eta_n^*} \mid E_1) \geq 1 - 1/n$, where $\eta_n^* = \left(C_2 C_3 \sqrt{C_4 + 1} + C_4 \max_k \Lambda_k\right) s_0 \log p / n$.*

Lemma 2.10.6 implies

$$\mathbb{P}(\mathbb{E}_{\eta_n^*} \cap E_1) = \mathbb{P}(E_1)\mathbb{P}(\mathbb{E}_{\eta_n^*} \mid E_1) \geq (1 - \mathbb{P}(E_1^c))\,(1 - 1/n) \geq 1 - \frac{1}{n} - \mathbb{P}(E_1^c).$$

Let $\delta_k = \widehat{\beta}^{(k)} - \beta(\tau_k)$ $(k = 1, \ldots, K_n)$. On event E_3, we have

$$\sup_{v \in A^{(k)}(\psi_\lambda), \|v\|_{k,2} \leq \|\delta_k\|_{k,2}} \left|\widetilde{\mathbb{Q}}^{(k)}(v)\right| \leq C_1 \frac{1 + \psi_\lambda}{k(s_0, \psi_\lambda)} \|\delta_k\|_{k,2} \sqrt{\frac{s_0 \log p}{n}} \ (k = 1, \ldots, K_n),$$

where $\psi_\lambda = (d_{\min} + 2\lambda)/(d_{\min} - 2\lambda)$ as defined in Theorem 2.5.1, and $\mathbb{P}(E_3) \geq 1 - 1/n$ by Lemma 2.10.1.

Proof of Theorem 2.5.1. Throughout the proof, we assume $\mathbb{E}_{\eta_n^*} \cap E_1 \cap E_3$, where $\mathbb{P}(\mathbb{E}_{\eta_n^*} \cap E_1 \cap E_3) \geq 1 - 2/n - \mathbb{P}(E_1^c)$. To exploit the results of Theorem 2.3.1, we first show that the conditions stated in Theorem 2.3.1 hold and then find a constant c_0 in the current settings. Note that we have $W_0 \vee W_1 = W_2 = 1$ because the maximum absolute value of $P_{a,\zeta_n}(\cdot)$ is at most 1, and $P_{a,\zeta_n}\left(\widetilde{\beta}_j^{(k)}\right) = 1$ $(j \in \{T^{(k)}\}^c)$ and $P_{a,\zeta_n}\left(\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}\right) = 1$ $(j \in \{B^{(k)}\}^c)$, which follows from

$$\begin{aligned}
|\widetilde{\beta}_j^{(k)}| &\leq \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \leq C_3 \sqrt{\frac{s_0 \log p}{n}} < \zeta_n \quad (j \in \{T^{(k)}\}^c), \\
|\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}| &\leq \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 + \|\widetilde{\beta}^{(k-1)} - \beta(\tau_{k-1})\|_2 \\
&\leq 2C_3 \sqrt{\frac{s_0 \log p}{n}} \leq \zeta_n \quad (j \in \{B^{(k)}\}^c).
\end{aligned}$$

Therefore, Condition 2.3.2 holds and we have

$$\frac{d_{\min}W_1 + 2\lambda(W_0 \vee W_1)}{d_{\min}W_2 - 2\lambda(W_0 \vee W_1)} \leq \psi_\lambda.$$

By using the growth condition of Theorem 2.5.1, where

$$\widetilde{C}_2 := \frac{9\underline{f}^3 q^2(2s_0, \psi_\lambda)}{32\bar{f}^2}\left\{\frac{k^2(s_0,\psi_\lambda)}{8C_1^2(1+\psi_\lambda)^2} \wedge \frac{1}{C_2C_3\sqrt{C_4+1}+C_4\max_k \Lambda_k}\right\},$$

we see that the conditions of Theorem 2.3.1 hold with $c_0 = \psi_\lambda$ and $\eta = \eta_n^*$. Hence we can use the results of Theorem 2.3.1 with $\eta = \eta_n^*$ and $c_0 = \psi_\lambda$. Hence we have

$$\begin{aligned}
&\|\widehat{\beta}^{(k)} - \beta(\tau_k)\|_2 \\
\leq\ & \frac{4d_{\min}^2}{(d_{\min}-2\lambda)^2 k(2s_0,\psi_\lambda)\sqrt{\underline{f}}}\sqrt{\frac{s_0\log p}{n} + \{C_2C_3\sqrt{C_4+1}+C_4\max_k \Lambda_k\}\frac{s_0\log p}{n}} \\
\leq\ & \xi_2\sqrt{\frac{s_0\log p}{n}}\ (k=1,\ldots,K_n), \qquad (2.35)
\end{aligned}$$

where

$$\xi_2 = \frac{4d_{\min}^2}{(d_{\min}-2\lambda)^2 k(2s_0,\psi_\lambda)\sqrt{\underline{f}}}\sqrt{1 + C_2C_3\sqrt{C_4+1}+C_4\max_k \Lambda_k}.$$

This completes the proof. □

Proofs of Theorem 2.5.2

Let $C_5 = \{(a\alpha + C_3) \vee \xi_2\}$ and $C_6 = \{(a\alpha + 2C_3) \vee 2\xi_2\}/(K_n d_{\min})$, where $\alpha = \zeta_n(s_0 \log p/n)^{-0.5}$. We first state the following lemma, which is useful to prove Theorem 2.5.2.

Lemma 2.10.7. *Suppose the conditions of Theorem 2.5.2 hold. Then on event E_1, we have $W_1 = 0$, where W_1 is defined in Subsection 2.3.2.*

Proof of Theorem 2.5.2. Throughout the proof, we assume $\mathbb{E}_{\eta_n^*} \cap E_1 \cap E_3$. By Lemma 2.10.7, we have

$$G(\mathcal{B}^o) = \sum_{k=1}^{K_n} \sum_{j \in \{T^{(k)}\}^c} w_j^{(k)} |\beta_j(\tau_k)| + \sum_{k=2}^{K_n} \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j \in \{B^{(k)}\}^c} v_j^{(k)} |\beta_j(\tau_k) - \beta_j(\tau_{k-1})| = 0,$$

where $G(\cdot)$ is the objective function of our optimization problem as defined in (2.25) and $\mathcal{B}^o$ is the true parameter, which shows that $\mathcal{B}^o$ becomes one of optimal solutions.

To show the second part of Theorem 2.5.2, notice that the proof of Theorem 2.3.1 and the result of Theorem 2.5.1 shows that (3.6) in the main paper holds with $\eta = \eta_n^*$ and $c_0 = \psi_\lambda$. Then the equation (3.6) with $W_1 = 0$ implies

$$\widehat{\beta}^{(k)}_{\{T^{(k)}\}^c} = 0 \quad (k = 1, \ldots, K_n), \quad \{\widehat{\beta}^{(k)} - \widehat{\beta}^{(k-1)}\}_{\{B^{(k)}\}^c} = 0 \quad (k = 2, \ldots, K_n). \tag{2.36}$$

We also have

$$\begin{aligned} \min_k \min_{j \in T^{(k)}} |\widehat{\beta}_j^{(k)}| &\geq \min_k \min_{j \in T^{(k)}} |\beta_j(\tau_k)| - \max_k \|\widehat{\beta}^{(k)} - \beta(\tau_k)\|_2 \\ &> \xi_2 \sqrt{\frac{s_0 \log p}{n}} - \xi_2 \sqrt{\frac{s_0 \log p}{n}} = 0, \end{aligned} \tag{2.37}$$

where the second inequality holds from the beta-min condition as stated in Theorem 2.5.2. Similarly,

$$\begin{aligned} \min_{k \geq 2} \min_{j \in B^{(k)}} |\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k-1)}| &\geq \min_{k \geq 2} \min_{j \in B^{(k)}} |\beta_j(\tau_k) - \beta_j(\tau_{k-1})| - 2 \max_k \|\widehat{\beta}^{(k)} - \beta(\tau_k)\|_2 \\ &> 2\xi_2 \sqrt{\frac{s_0 \log p}{n}} - 2\xi_2 \sqrt{\frac{s_0 \log p}{n}} = 0. \end{aligned} \tag{2.38}$$

By (2.36), (2.37) and (2.38), we have that $\widehat{\mathcal{B}}$ provides the exact model structure, which completes the proof. □

Proofs of Theorem 2.6.1

Here, we define the map T and new design matrix $z_i^{(k)}$, which are stated in Section 2.5. We first define a map $M : \{1, \ldots, p\} \times \{1, \ldots, K_n\} \to \mathbb{R}^{d_0}$ as follows:

1. If $\widehat{\beta}_j^{(k)} = 0$, then $M(j,k) = 0$.

2. if $\widehat{\beta}_j^{(k)} = \widehat{\beta}_j^{(k-1)}$, then $M(j,k) = M(j,k-1)$.

3. If $\widehat{\beta}_j^{(k)} \neq 0$, $\widehat{\beta}_{j'}^{(k')} = 0$ $(k' = 1,\ldots,K_n;\ j' = 1,\ldots,j-1)$ and $\widehat{\beta}_j^{(k')} = 0$ $(k' = 1,\ldots,k-1)$, then $M(j,k) = 1$.

4. If $\widehat{\beta}_j^{(k)} \neq 0$ and $\widehat{\beta}_j^{(k)} \neq \widehat{\beta}_j^{(k-1)}$, then

$$M(j,k) = 1 + \max(M_1, M_2),$$

where $M_1 := \{M(j',k') :\ k' = 1,\ldots,K_n;\ j' = 1,\ldots,j-1\}$ and

$M_2 := \{M(j,k') :\ k' = 1,\ldots,k-1\}$.

5. If $\widehat{\beta}_j^{(1)} \neq 0$ for $j \geq 2$, then

$$M(j,1) = 1 + \max\{M(j',k') :\ k' = 1,\ldots,K_n;\ j' = 1,\ldots,j-1\}.$$

By using the map M, we arrive at a new design matrix denoted by $z_i^{(k)} \in \mathbb{R}^{d_0}$ $(i = 1,\ldots,n;\ k = 1,\ldots,K_n)$. First let $M(T^{(k)},k) = \{M(j,k) : j \in T^{(k)}\}$ for $k = 1,\ldots,K_n$, where the elements in $M(T^{(k)},k)$ are in ascending order. Then let ${z^{(k)}}_{i,M(T^{(k)},k)} = x_{i,T^{(k)}}$ and $z_{i,j}^{(k)} = 0$ for $j \in \{1,\cdots,d_0\} \setminus M(T^{(k)},k)$.

We also define the map T as follows. Let $\mathrm{IM} = \{(j,k) : M(j,k) \neq 0,\ M(j,k) \neq M(j,k-1)\}$, which is the location indices account for effective components. Then for any $\mathcal{B} \in G$, $T(\mathcal{B}) \in \mathbb{R}^{d_0}$, where $T(\mathcal{B})_i = \mathcal{B}_{j,k}$ $(i = 1,\ldots,d_0)$ for i satisfying $M(j,k) = i$ and $(j,k) \in \mathrm{IM}$. We can see that for any $\mathcal{B} \in G$, $T(\mathcal{B})$ is the d_0–dimensional vector, which can construct $\mathcal{B}$ given its structure.

Remark 2.10.1. Illustrative example of the M and T.

Suppose that we consider the model, where $p = 5$, $n = 10$ and $K_n = 3$ with the three quantile levels τ_1, τ_2 and τ_3. Assume that we obtain the following estimates from our Dantzig-type optimization problem:

$$\widehat{B} = \begin{bmatrix} 0.9 & 0.9 & 0.0 \\ 1.1 & 1.5 & 1.5 \\ 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 1.0 \\ 0.0 & 0.2 & 0.2 \end{bmatrix} = \left[\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}, \widehat{\beta}^{(3)}\right].$$

Then M is the function such that

$$M(j,k) = \widetilde{M}_{j,k} \quad \text{for} \quad j = 1, \cdots, 5, \quad \text{and} \quad k = 1, 2, 3,$$

where

$$\widetilde{M} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 3 \\ 0 & 0 & 0 \\ 4 & 0 & 5 \\ 0 & 6 & 6 \end{bmatrix}$$

is the indices matrix that uses the model structure of $\widehat{B}$.

Here $d_0 = 6$, and

$$T(\widehat{B}) = [0.9, 1.1, 1.5, 0.5, 1.0, 0.2]^T.$$

Lemma 2.10.8. *Assume* $d_0 M_n^4 (\log n)^2 = o(n)$. *Let* $\Delta > 0$ *and* $\Theta = \{\theta \in R^{d_0} : \|\theta\|_2 \leq \Delta\}$. *For any* $\theta \in \Theta$, *let*

$$I_2(\theta) = \frac{1}{\|\theta\|_2} \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} I(\epsilon_i^{(k)} \leq x) - I(\epsilon_i^{(k)} \leq 0) dx,$$

where $\epsilon_i^{(k)} = y_i - {x_i}^T\beta(\tau_k) = y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o)$. *Then with probability at least* $1 - n^{-9d_0 \log n} - 2\bar{f}K_n/n$,

$$\sup_{\theta \in \Theta_n} |I_2(\theta) - E[I_2(\theta)]| \leq 2n^{-1/4} K_n \Delta^{1/2} s_0^{3/4} (d_0)^{5/4} (\log n)^{3/2}.$$

Proof of Theorem 2.6.1. We will show that for any constant $\epsilon > 0$, there exists a sufficiently large constant $\Delta > 0$ satisfying

$$\mathbb{P}\Big[\inf_{\|\theta\|_2 = \Delta,\ \theta \in R^{d_0}} L_n\Big(T(\mathcal{B}^o) + \sqrt{\frac{d_0}{n}}\theta\Big) > L_n(T(\mathcal{B}^o))\Big] \geq 1 - \epsilon, \tag{2.39}$$

where $L_n(\theta) = \sum_k \sum_i \rho_{\tau_k}[y_i - \{z_i^{(k)}\}^T\theta]$ for any $\theta \in R^{d_0}$. Since the objective function L_n is a strict convex function over $\theta \in R^{d_0}$, (2.39) implies that the global minimum $T(\widehat{\mathcal{B}})$ lies within the ball whose center is $T(\mathcal{B}^o)$ and the radius is $\Delta\sqrt{d_0/n}$ with probability at least $1 - \epsilon$, which proves the theorem. Let

$$G_n(\theta) = L_n\left(T(\mathcal{B}^o) + \sqrt{\frac{d_0}{n}}\theta\right) - L_n\left(T(\mathcal{B}^o)\right).$$

By using the Knight's identity,

$$\begin{aligned}
G_n(\theta) &= \sum_k \sum_i \rho_{\tau_k}\Big[y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o) - \sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta\Big] - \rho_{\tau_k}\Big[y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o)\Big] \\
&= \sqrt{\frac{d_0}{n}} \sum_k \sum_i \{z_i^{(k)}\}^T\theta\{I(\epsilon_i^{(k)} < 0) - \tau_k\} \\
&+ \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} I(\epsilon_i^{(k)} \leq x) - I(\epsilon_i^{(k)} \leq 0)dx \\
&:= I_1(\theta) + I_2(\theta),
\end{aligned}$$

where $\epsilon_i^{(k)}$ is defined in Lemma 2.10.8. First consider $I_1(\theta)$. Let $v_i^{(k)} = I(\epsilon_i^{(k)} < 0) - \tau_k$

and $\Theta = \{\theta \in R^{d_0} : \|\theta\|_2 = \Delta\}$. Then we have

$$\begin{aligned}
\mathbb{E}\big[\sup_{\theta\in\Theta} I_1^2(\theta)\big] &= \frac{d_0}{n}\mathbb{E}\left[\sup_{\|\theta\|_2\in\Theta}\left\{(\sum_k\sum_i z_i^{(k)}v_i^{(k)})^T\theta\right\}^2\right] \\
&= \frac{d_0}{n}\mathbb{E}\big[\sup_{\|\theta\|_2\in\Theta}\theta^T ZZ^T\theta\big] \\
&\le \frac{d_0}{n}\Delta^2\mathbb{E}[\lambda_{\max}(ZZ^T)], \qquad (2.40)
\end{aligned}$$

where $Z = \sum_k\sum_i\{z_i^{(k)}v_i^{(k)}\}$. Note that ZZ^T is a zero matrix or rank–one matrix, and Z^TZ is a eigenvalue of ZZ^T when ZZ^T is rank–one. Hence $\lambda_{\max}(ZZ^T) \le Z^TZ$, which implies with (2.40) that

$$\begin{aligned}
\mathbb{E}\big[\sup_{\theta\in\Theta} I_1^2(\theta)\big] &\le \frac{d_0}{n}\Delta^2\mathbb{E}[Z^TZ] \\
&= \frac{d_0}{n}\Delta^2\mathbb{E}[\sum_k\sum_{k'}\sum_i v_i^{(k)}v_i^{(k')}\{z_i^{(k)}\}^T z_i^{(k')}] \\
&= \frac{d_0}{n}\Delta^2\sum_k\sum_{k'}(\tau_k\wedge\tau_{k'} - \tau_k\tau_{k'})\sum_i\{z_i^{(k)}\}^T z_i^{(k')} \\
&\le \Delta^2K_n^2d_0^2.
\end{aligned}$$

Hence, by Markov inequality,

$$\mathbb{P}\left(\sup_{\|\theta\|_2\in\Theta}|I_1(\theta)| \ge \frac{\Delta K_n d_0}{\sqrt{\epsilon/2}}\right) \le \frac{\epsilon}{2}.$$

Hence, with probability at least $1-\epsilon/2$, we have $\sup_{\|\theta\|_2\in\Theta}|I_1(\theta)| \le \frac{\Delta K_n d_0}{\sqrt{\epsilon/2}}$.

Now consider $I_2(\theta)$. Then for any $\theta \in \Theta$, we have

$$\begin{aligned}
\mathbb{E}\left(I_2(\theta)\right) &= \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} \mathbb{P}\left(y_i \le {x_i}^T\beta(\tau_k) + x\right) - \mathbb{P}\left(y_i \le {x_i}^T\beta(\tau_k)\right) dx \\
&= \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} x f_i\left({x_i}^T\beta(\tau_k)\right) + \frac{x^2}{2} f_i'({x_i}^T\beta(\tau_k) + \widetilde{x}_i^{(k)}) dx \\
&\ge \sum_k \sum_i \frac{f_i\left({x_i}^T\beta(\tau_k)\right)}{2} \frac{d_0}{n} [\{z_i^{(k)}\}^T\theta]^2 - \frac{\bar{f}}{6} \sum_k \sum_i \left(\frac{d_0}{n}\right)^{1.5} [\{z_i^{(k)}\}^T\theta]^3 \\
&\ge \frac{d_0 \underline{f}}{2} \sum_k \sum_i \theta^T \frac{1}{n} z_i^{(k)} \{z_i^{(k)}\}^T\theta - \frac{\Delta M_n \bar{f} d_0^{1.5}\sqrt{s_0}}{6\sqrt{n}} \sum_k \sum_i \theta^T \frac{1}{n} z_i^{(k)} \{z_i^{(k)}\}^T\theta \\
&\ge \frac{d_0 \underline{f}}{4} \sum_k \sum_i \theta^T \frac{1}{n} z_i^{(k)} \{z_i^{(k)}\}^T\theta \\
&\ge \frac{K_n d_0 \underline{f}}{4} k^2(s_0, 0)\Delta^2
\end{aligned}$$

where $\widetilde{x}_i^{(k)} \in (0, x)$ which depends on i and k in the second line, and the first and the second inequality follow from Condition 2.3.1 and $|\{z_i^{(k)}\}^T\theta| \le \|z_i^{(k)}\|_2 \|\theta\|_2 \le M_n\sqrt{s_0}\Delta$. The third inequality holds due to $M_n^2 d_0 s_0 = o(n)$, and the last inequality follows from Condition 2.3.1 with the fact that $\sum_i z_i^{(k)}\{z_i^{(k)}\}^T/n$ is a $s_k \times s_k$–dimensional sub–matrix of $\sum_i x_i {x_i}^T/n$. By Lemma 2.10.8 and the conditions of Theorem 2.6.1, $I_2(\theta) \ge \frac{K_n d_0 \underline{f}}{4} k^2(s_0, 0)\Delta^2 - \Delta^{3/2} o_p(K_n d_0)$, where $o_p(1)$ is uniformly over $\theta \in \Theta$.

Hence for any $\epsilon > 0$, with probability at least $1 - \epsilon/2$,

$$\inf_{\theta \in \Theta} G_n(\theta) \ge \frac{K_n d_0 \underline{f}}{4} k^2(s_0, 0)\Delta^2 - \Delta^{3/2} o_p(K_n d_0) - \frac{\Delta K_n d_0}{\sqrt{\epsilon/2}} > 0$$

with a sufficiently large Δ, which completes the proof. □

Proofs of Theorem 2.6.2

Lemma 2.10.9. *Let A_n and B_n be the matrix stated in Theorem 2.6.2. Then we*

have

$$\bar{f}^{-2}\phi^{-2}(s_0)k^2(s_0,0)(\min_k \tau_k)(1-\max_k \tau_k) \le \lambda_{\min}({A_n}^{-1}B_n{A_n}^{-1}),$$

$$\lambda_{\max}({A_n}^{-1}B_n{A_n}^{-1}) \le L_0^{-2}\phi(s_0)k^{-4}(s_0,0).$$

Lemma 2.10.10. *Assume conditions of Theorem 2.6.2 hold. Then for any sequence of $\alpha_n \in R^{d_0}$ with $\|\alpha_n\|_2 = 1$, the following asymptotic normality holds:*

$$n^{-1/2}\alpha_n^T({A_n}^{-1}B_n{A_n}^{-1})^{-1/2}{A_n}^{-1}\sum_k\sum_i z_i^{(k)}\left(I(y_i - {x_i}^T\beta(\tau_k)<0)-\tau_k\right) \to N(0,1),$$

where A_n and B_n are $d_0 \times d_0$ matrices defined in Theorem 2.6.2.

Proof of Theorem 2.6.2. Recall that $T(\widehat{\mathcal{B}}^{po})$ is

$$T(\widehat{\mathcal{B}}^{po}) = \arg\min_{\beta\in R^{d_0}} \sum_k\sum_i \rho_{\tau_k}(y_i - \{z_i^{(k)}\}^T\beta) \tag{2.41}$$

By $\theta = \sqrt{n/d_0}\,(\beta - T(\mathcal{B}^o))$, $T(\widehat{\mathcal{B}}^{po}) = T(\mathcal{B}^o) + \sqrt{d_0/n}\widehat{\theta}$, where $\widehat{\theta}$ is

$$\widehat{\theta} = \arg\min_{\theta\in R^{d_0}} \sum_k\sum_i \rho_{\tau_k}\Big[y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o) - \sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta\Big]. \tag{2.42}$$

Then $\widehat{\theta}$ can be written as $\widehat{\theta} = G_n(\theta)$, where

$$\begin{aligned} G_n(\theta) = \text{argmin}_{\theta\in R^{d_0}} \quad & \sum_k\sum_i \rho_{\tau_k}\left(y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o) - \sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta\right) \\ - \quad & \sum_k\sum_i \rho_{\tau_k}\left(y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o)\right). \end{aligned}$$

We consider θ over the set $\Theta_n = \{\theta \in R^{d_0} \mid \|\theta\|_2 \le C\}$ with some positive constant C independent of n. Decompose G_n into two terms as similarly used in the proof of Theorem 2.6.1:

$$G_n(\theta) = I_1(\theta) + I_2(\theta),$$

where

$$I_1(\theta) = \sqrt{\frac{d_0}{n}}\sum_k\sum_i \{z_i^{(k)}\}^T\theta\{I(\epsilon_i^{(k)}<0) - \tau_k\},$$

$$I_2(\theta) = \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} I(\epsilon_i^{(k)} \leq x) - I(\epsilon_i^{(k)} \leq 0)dx.$$

Consider the term $I_2(\theta)$. From the proof of Theorem 2.6.1, we have

$$\begin{aligned}
& \Big|\mathbb{E}[I_2(\theta)] - \sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} f_i({x_i}^T\beta(\tau_k))xdx\Big| \\
\leq & \Big|\sum_k \sum_i \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} \frac{x^2}{2} f_i'(\widetilde{x}_i^{(k)})dx\Big| \\
\leq & \frac{\bar{f}}{6}\sum_k \sum_i \big(\frac{d_0}{n}\big)^{1.5}|\{z_i^{(k)}\}^T\theta|^3 \\
\leq & \frac{\bar{f}}{6}K\frac{d_0^{1.5}\sqrt{s_0}M_n}{\sqrt{n}}\|\theta\|_2^3\phi(s_0) \\
= & o(\|\theta\|_2 K),
\end{aligned}$$

where $\widetilde{x}_i^{(k)} \in ({x_i}^T\beta(\tau_k), {x_i}^T\beta(\tau_k) + x)$ depends on i and k in the first inequality, the second inequality follows from Condition 2.3.1, the third inequality holds due to Sparse(s_0) and $|\{z_i^{(k)}\}^T\theta| \leq \|z_i^{(k)}\|_2\|\theta\|_2 \leq M_n\|\theta\|_2\sqrt{s_0}$, and the last small o results follows from $M_n^2 d_0^3 s_0 = o(n)$. Moreover, Lemma 2.10.8 and the conditions of Theorem 2.6.2 imply

$$I_2(\theta) - \mathbb{E}[I_2(\theta)] = o_p(\|\theta\|_2 K_n),$$

where o_p is uniform over $\theta \in \Theta_n$. Hence, for all $\theta \in \Theta_n$,

$$I_2(\theta) = \sum_k \sum_i \frac{f_i({x_i}^T\beta(\tau_k))}{2}\frac{d_0}{n}[\{z_i^{(k)}\}^T\theta]^2 + o_p(\|\theta\|_2 K_n).$$

Thus, for all $\theta \in \Theta_n$, $G_n(\theta)$ can be written as

$$\begin{aligned}
G_n(\theta) = & \sqrt{\frac{d_0}{n}}\sum_k \sum_i \{z_i^{(k)}\}^T\theta\left(I(\epsilon_i^{(k)} < 0) - \tau_k\right) \\
+ & \sum_k \sum_i \frac{f_i({x_i}^T\beta(\tau_k))}{2}\frac{d_0}{n}[\{z_i^{(k)}\}^T\theta]^2 + o_p(\|\theta\|_2 K_n).
\end{aligned}$$

By matrix calculus,

$$\begin{aligned}
\widehat{\theta} &= \sqrt{\frac{n}{d_0}}\Big\{\sum_k\sum_i f_i({x_i}^T\beta(\tau_k))z_i^{(k)}\{z_i^{(k)}\}^T\Big\}^{-1}\sum_k\sum_i z_i^{(k)}\{I(\epsilon_i^{(k)}<0)-\tau_k\} \\
&+ \left(\sum_k\sum_i \frac{f_i({x_i}^T\beta(\tau_k))}{2}\frac{d_0}{n}z_i^{(k)}\{z_i^{(k)}\}^T\right)^{-1} K_n o_p(1) \\
&= (nd_0)^{-0.5}{A_n}^{-1}\sum_k\sum_i z_i^{(k)}\{I(\epsilon_i^{(k)}<0)-\tau_k\} + 2A_n^{-1}\frac{K}{d_0}o_p(1) \\
&= d_0^{-0.5}({A_n}^{-1}B_n{A_n}^{-1})^{\frac{1}{2}}\Big[n^{-0.5}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}\sum_k\sum_i z_i^{(k)}\{I(\epsilon_i^{(k)}<0)-\tau_k\}\Big] \\
&+ \frac{1}{d_0}o_p(1),
\end{aligned}$$

where $o_p(1)$ represents any d_0 dimensional vector whose ℓ_2 norm is $o_p(1)$.

For any $\alpha_n \in R^{d_0}$ with $\|\alpha_n\|_2 = 1$, Lemma 2.10.10 implies

$$\alpha_n^T\Big[n^{-0.5}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}\sum_k\sum_i z_i^{(k)}\{I(\epsilon_i^{(k)}<0)-\tau_k\}\Big] \to N(0,1).$$

Hence

$$\begin{aligned}
\|\widehat{\theta}\|_2 &\le d_0^{-0.5}\lambda_{\max}\{({A_n}^{-1}B_n{A_n}^{-1})^{\frac{1}{2}}\}O_p\{\sqrt{d_0}\} + o_p(1) \\
&\le L_0^{-1}\sqrt{\phi(s_0)}k^{-2}(s_0,0)O_p(1),
\end{aligned}$$

due to Lemma 2.10.9. Since C can be chosen to be much larger than $L_0^{-1}\sqrt{\phi(s_0)}k^{-2}(s_0,0)$, $\widehat{\theta}$ is included in Θ_n. Hence by Lemma 2.10.10 ,

$$\alpha_n^T\sqrt{n}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\sqrt{\frac{d_0}{n}}\widehat{\theta} \to N(0,1).$$

Thus,

$$\alpha_n^T\sqrt{n}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\{T(\widehat{\mathcal{B}}^{po}) - T(\mathcal{B}^o)\} \to N(0,1),$$

which completes the proof. □

Proof of Lemmas

Proof of Lemma 2.10.1,2.10.2 and 2.10.3. The proofs essentially follow from the proofs of Lemmas 4 and 5 in Belloni and Chernozhukov (2011) by using $K_n < p$. □

Proof of Lemma 2.10.4. Suppose $\mathbb{E}_\eta$ holds. Then $\beta(\tau_k) \in \mathbb{R}^{(k)}(r_k)$ $(k = 1, \ldots, K_n)$, where $\mathbb{R}^{(k)}(r_k)$ is defined in (3.1) which implies that

$$\mathcal{B}^{(k)} = [\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(k-1)}, \beta(\tau_k), \widehat{\beta}^{(k+1)}, \cdots, \widehat{\beta}^{(K)}]$$

is feasible for all k. We fix any k. Since $\widehat{\mathcal{B}}$ is a global minimizer of (3.1), we have $G(\widehat{\mathcal{B}}) \leq G\{\mathcal{B}^{(k)}\}$, where $G(\cdot)$ is defined in (2.25), which implies

$$\begin{aligned} & \sum_{j=1}^{p} w_j^{(k)} |\widehat{\beta}_j^{(k)}| + \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j=1}^{p} v_j^{(k)} |\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k-1)}| + \frac{\lambda}{|\tau_{k+1} - \tau_k|} \sum_{j=1}^{p} v_j^{(k+1)} |\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k+1)}| \\ \leq & \sum_{j=1}^{p} w_j^{(k)} |\beta_j(\tau_k)| + \frac{\lambda}{|\tau_k - \tau_{k-1}|} \sum_{j=1}^{p} v_j^{(k)} |\beta_j(\tau_k) - \widehat{\beta}_j^{(k-1)}| + \frac{\lambda}{|\tau_{k+1} - \tau_k|} \sum_{j=1}^{p} v_j^{(k+1)} |\beta_j(\tau_k) - \widehat{\beta}_j^{(k+1)}|. \end{aligned}$$

By applying the triangle inequality and the definition of $d_{\min}$, it reduces to

$$\sum_{j \in \{T^{(k)}\}^c} w_j^{(k)} |\widehat{\beta}_j^{(k)}| \leq \sum_{j \in T^{(k)}} w_j^{(k)} (|\beta_j(\tau_k)| - |\widehat{\beta}_j^{(k)}|) + \frac{\lambda}{d_{\min}} \sum_{j=1}^{p} (v_j^{(k)} + v_j^{(k+1)}) |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)|.$$

Rearranging the terms yields

$$\sum_{j \in \{T^{(k)}\}^c} [w_j^{(k)} - \frac{\lambda}{d_{\min}} \{v_j^{(k)} + v_j^{(k+1)}\}] |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)| \leq \sum_{j \in T^{(k)}} [w_j^{(k)} + \frac{\lambda}{d_{\min}} \{v_j^{(k)} + v_j^{(k+1)}\}] |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)|.$$

By the definition of W_2, W_1 and W as stated in Subsection 2.3.2, we have

$$\sum_{j \in \{T^{(k)}\}^c} \left(W_2 - \frac{2\lambda}{d_{\min}} W \right) |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)| \leq \sum_{j \in T^{(k)}} \left(W_1 + \frac{2\lambda}{d_{\min}} W \right) |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)|.$$

Condition 2.3.2 implies $W_2 - \frac{2\lambda}{d_{\min}} (W_0 \vee W_1) > 0$, and we have for $k = 1, \ldots, K_n$,

$$\sum_{j \in \{T^{(k)}\}^c} |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)| \leq \frac{d_{\min} W_1 + 2\lambda (W_0 \vee W_1)}{d_{\min} W_2 - 2\lambda (W_0 \vee W_1)} \sum_{j \in T^{(k)}} |\widehat{\beta}_j^{(k)} - \beta_j(\tau_k)|,$$

which completes the proof. □

Proof of Lemma 2.10.5. Lemma 1.5 in Ledoux and Talagrand (1991) implies that for any independent mean zero random variables $Z_1, \cdots, Z_n$, which satisfy $|Z_i| \leq c_i$ $(i = 1, \ldots, n)$, where c_is are some constants, we have that for any $t > 0$,

$$\mathbb{P}\left(|\sum_{i=1}^{n} Z_i| > t \right) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2} \right). \tag{2.43}$$

Fix j, k and any $t > 0$. Let $Z_i = x_{ij}[\tau_k - I\{y_i \leq {x_i}^T \beta(\tau_k)\}]/n$, where x_{ij} is the jth component of x_i. It follows from (2.43) that

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} x_{ij}[\tau_k - I\{y_i \leq {x_i}^T \beta(\tau_k)\}]/n \right| \geq t \right) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} x_{ij}^2/n^2} \right) = 2 \exp\left(-\frac{nt^2}{2} \right),$$

where we set $c_i = x_{ij}/n$. By the union bound,

$$\mathbb{P}\left(\max_k \max_j |\sum_{i=1}^{n} x_{ij}[\tau_k - I\{y_i \leq {x_i}^T \beta(\tau_k)\}]/n| \geq t \right) \leq 2Kp \exp(-\frac{nt^2}{2}).$$

Putting $t = 3\sqrt{\log p/n}$ and using $p > n \vee K_n$ yields

$$\mathbb{P}\left(\max_k \max_j |\sum_{i=1}^{n} x_{ij}[\tau_k - I\{y_i \leq {x_i}^T \beta(\tau_k)\}]/n| \geq 3\sqrt{\log p/n} \right) \leq \frac{1}{n},$$

which completes the proof. □

Proof of Lemma 2.10.6. Suppose E_1 holds. Then we have for all $k = 1, \ldots, K_n$,

$$\|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_1 \leq \sqrt{\|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_0} \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \leq \sqrt{(C_4 + 1)s_0} C_3 \sqrt{\frac{s_0 \log p}{n}}. \tag{2.44}$$

Note that (2.44) uniformly holds for all k, with probability at least $1 - q_n$.

Note that r_k in the Dantzig-type joint quantile regression setting stated in Section 2.4 is $r_k = \mathbb{Q}_n^{(k)}\{\widetilde{\beta}^{(k)}\} + \Lambda_k \widetilde{s} \log p/n$, where $\widetilde{s} = \max_k \|\widetilde{\beta}^{(k)}\|_0$. Then the event $\mathbb{E}_\eta$, which is defined in (3.3), is equivalent to

$$\mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} \leq \mathbb{Q}_n^{(k)}(\widetilde{\beta}^{(k)}) + \Lambda_k \frac{\widetilde{s} \log p}{n} \leq \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + \eta \quad (k = 1, \ldots, K_n). \tag{2.45}$$

To prove (2.45), we use the fact that $\mathbb{Q}_n^{(k)}$ is a convex function and $-\sum_{i=1}^n x_i[\tau_k - I\{y_i \le {x_i}^T\beta(\tau_k)\}]/n$ is the subgradient of $\mathbb{Q}_n^{(k)}$ at $\beta(\tau_k)$. Then we have

$$\begin{aligned}
\mathbb{Q}_n^{(k)}\{\widetilde{\beta}^{(k)}\} - \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} &\ge \left(-\frac{1}{n}\sum_{i=1}^n x_i[\tau_k - I\{y_i \le {x_i}^T\beta(\tau_k)\}]\right)^T \{\widetilde{\beta}^{(k)} - \beta(\tau_k)\} \\
&\ge -\|\frac{1}{n}\sum_{i=1}^n (x_i[\tau_k - I\{y_i \le {x_i}^T\beta(\tau_k)\}])\|_\infty \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_1. \qquad (2.46)
\end{aligned}$$

Let E_4 be the event

$$E_4 = \left\{ \|\frac{1}{n}\sum_{i=1}^n (x_i[\tau_k - I\{y_i \le {x_i}^T\beta(\tau_k)\}])\|_\infty \le 3\sqrt{\frac{\log p}{n}} \right\}. \qquad (2.47)$$

By Lemma 2.10.5, $P(E_4) \ge 1 - 1/n$. Combining (2.44), (2.46) and (2.47), on event E_4,

$$\begin{aligned}
\mathbb{Q}_n^{(k)}\{\widetilde{\beta}^{(k)}\} - \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} &\ge -3\sqrt{C_4+1}C_3\frac{s_0 \log p}{n}, \\
&\ge -\Lambda_k \frac{\widetilde{s}\log p}{n}, \qquad (2.48)
\end{aligned}$$

where the last inequality uses the condition 4. Hence the first inequality of (2.45) holds for all k.

Now, by using the fact that $\widetilde{\beta}^{(k)}$s and $\widetilde{\lambda}$ satisfy (2.34) on event E_1, we can show that the second inequality of (2.45) holds with $\eta = \eta_n^*$ as follows:

$$\begin{aligned}
\mathbb{Q}_n^{(k)}\{\widetilde{\beta}^{(k)}\} + \Lambda_k\frac{\widetilde{s}\log p}{n} &\le \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + \widetilde{\lambda}\{\|\beta(\tau_k)\|_1 - \|\widetilde{\beta}^{(k)}\|_1\} + \Lambda_k\frac{\widetilde{s}\log p}{n} \\
&\le \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + \widetilde{\lambda}\|\beta(\tau_k) - \widetilde{\beta}^{(k)}\|_1 + \Lambda_k\frac{\widetilde{s}\log p}{n} \\
&\le \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + C_2C_3\sqrt{C_4+1}\frac{s_0\log p}{n} + \Lambda_k\frac{\widetilde{s}\log p}{n} \\
&\le \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + \{C_2C_3\sqrt{C_4+1} + C_4\max_k\Lambda_k\}\frac{s_0\log p}{n} \\
&= \mathbb{Q}_n^{(k)}\{\beta(\tau_k)\} + \eta_n^*, \qquad (2.49)
\end{aligned}$$

where the first inequality follows from the definition of $\widetilde{\beta}^{(k)}$. Combining (2.48) and (2.49) implies that (2.45) holds with $\eta = \eta_n^*$, which completes the proof. □

Proof of Lemma 2.10.7. Suppose E_1 holds. Then we have

$$\begin{aligned}
\min_k \min_{j \in T^{(k)}} |\widetilde{\beta}_j^{(k)}| &\geq \min_k \min_{j \in T^{(k)}} |\beta_j(\tau_k)| - \max_k \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \\
&\geq (a\alpha + C_3)\sqrt{\frac{s_0 \log p}{n}} - C_3\sqrt{\frac{s_0 \log p}{n}} \\
&= a\alpha\sqrt{\frac{s_0 \log p}{n}} = a\zeta_n, \qquad (2.50)
\end{aligned}$$

where the second inequality follows from the beta-min condition as stated in Theorem 2.5.2. Similarly,

$$\begin{aligned}
\min_{k \geq 2} \min_{j \in B^{(k)}} |\widetilde{\beta}_j^{(k)} - \widetilde{\beta}_j^{(k-1)}| &\geq \min_{k \geq 2} \min_{j \in B^{(k)}} |\beta_j(\tau_k) - \beta_j(\tau_{k-1})| - 2\max_k \|\widetilde{\beta}^{(k)} - \beta(\tau_k)\|_2 \\
&\geq (a\alpha + 2C_3)\sqrt{\frac{s_0 \log p}{n}} - 2C_3\sqrt{\frac{s_0 \log p}{n}} \\
&= a\alpha\sqrt{\frac{s_0 \log p}{n}} \geq a\zeta_n. \qquad (2.51)
\end{aligned}$$

By (2.50) and (2.51), we have $W_1 = 0$, which completes the proof. □

Proof of Lemma 2.10.8. We fix $k \in \{1, \cdots, K_n\}$, and let

$$I_2^{(k)}(\theta) := \sum_i \frac{1}{\|\theta\|_2} \int_0^{\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta} I(\epsilon_i^{(k)} \leq x) - I(\epsilon_i^{(k)} \leq 0)dx := \sum_i I_{2,i}^{(k)}(\theta).$$

For the simplicity, let $D := \{\theta \in \mathbb{R}^{d_0} \mid \|\theta\|_2 \leq \frac{1}{nd_0M_n\sqrt{n}}\}$. First consider the case when $\|\theta\|_2 \in D$. Then

$$|\sqrt{\frac{d_0}{n}}\{z_i^{(k)}\}^T\theta| \leq \sqrt{\frac{d_0 s_0}{n}}\frac{M_n}{nd_0M_n\sqrt{n}} \leq \frac{1}{n^2}.$$

Define the events B and C as follows:

$$B = \left\{|\epsilon_i^{(k)}| > \frac{1}{n^2}, \quad \text{for all } i.\right\}$$

$$C = \left\{ \sup_{\theta:\ \theta \in D} I_2^{(k)}(\theta) = 0 \right\}$$

Then $\mathbb{P}(B) \geq 1 - n\frac{2\bar{f}}{n^2} = 1 - \frac{2\bar{f}}{n}$, which implies $\mathbb{P}(C) \geq 1 - \frac{2\bar{f}}{n}$. Moreover,

$$\sup_{\theta \in D} \left| \mathbb{E}[I_2^{(k)}(\theta)] \right| \leq \frac{2\bar{f}}{n} M_n \sqrt{n d_0 s_0} = \frac{2\bar{f} M_n \sqrt{d_0 s_0}}{\sqrt{n}}.$$

Hence with probability at least $1 - 2\bar{f}/n$,

$$\sup_{\theta \in D} \left| I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)] \right| \leq \frac{2\bar{f} M_n \sqrt{d_0 s_0}}{\sqrt{n}}.$$

Now consider the case when $\|\theta\|_2 > 1/(n d_0 M_n \sqrt{n})$. We have for any $\lambda > 0$,

$$\mathbb{P}\left(|I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)]| \geq t \right) \leq \exp\left(-\lambda t - \lambda \mathbb{E}[I_2^{(k)}(\theta)] \right) \mathbb{E}\left[\exp(\lambda I_2^{(k)}(\theta)) \right].$$

We have

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda I_2^{(k)}(\theta)) \right] &= \prod_i \mathbb{E}\left[\exp(\lambda I_{2,i}^{(k)}(\theta)) \right] \\
&= \prod_i \mathbb{E}\left[1 + \lambda I_{2,i}^{(k)}(\theta) + \lambda^2 (I_{2,i}^{(k)}(\theta))^2 O(1) \right] \\
&= \prod_i \left(1 + \lambda \mathbb{E}[I_{2,i}^{(k)}(\theta)] + \lambda^2 O(\mathbb{E}[(I_{2,i}^{(k)}(\theta))^2]) \right) \\
&\leq \exp\left(\lambda \sum_i \mathbb{E}[I_{2,i}^{(k)}(\theta)] + \lambda^2 \sum_i O(\mathbb{E}[(I_{2,i}^{(k)}(\theta))^2]) \right),
\end{aligned}$$

where in the second equality $O(1)$ holds uniformly for all i and θ, provided that $\max_i |\lambda I_{2,i}^{(k)}(\theta)| \leq \lambda M_n \sqrt{\frac{d_0 s_0}{n}} = o(1)$. Hence

$$\begin{aligned}
& \mathbb{P}\left(|I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)]| \geq t \right) \\
\leq\ & \exp\left(-\lambda t - \lambda \mathbb{E}[I_2^{(k)}(\theta)] + \lambda \sum_i \mathbb{E}[I_{2,i}^{(k)}(\theta)] + \lambda^2 \sum_i O(\mathbb{E}[(I_{2,i}^{(k)}(\theta))^2]) \right) \\
=\ & \exp\left(-\lambda t + \lambda^2 \sum_i O(\mathbb{E}[(I_{2,i}^{(k)}(\theta))^2]) \right) \\
=\ & \exp\left(-\lambda t + \lambda^2 O\left(\Delta \frac{s_0^{3/2} (d_0)^{3/2}}{\sqrt{n}} \right) \right),
\end{aligned}$$

where we use

$$\sum_i \mathbb{E}[(I_{2,i}^{(k)}(\theta))^2] \le \frac{1}{\|\theta\|_2^2}\sqrt{\frac{d_0}{n}}\frac{\bar{f}d_0}{2n}\sum_i \|\{|z_i^{(k)}\}^T\theta\|_2^2 \max_i |\{z_i^{(k)}\}^T\theta| \le \frac{\Delta\bar{f}M_n^3}{2}\frac{s_0^{3/2}(d_0)^{3/2}}{\sqrt{n}}.$$

Choosing $\lambda = \frac{t\sqrt{n}}{2\Delta s_0^{3/2}(d_0)^{3/2}\log n}$ with the growth condition $\frac{tM_n}{\Delta s_0 d_0 \log n} = o(1)$, we have

$$\mathbb{P}(|I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)]| \ge t) \le \exp\left(-\frac{t^2\sqrt{n}}{4\Delta s_0^{3/2}(d_0)^{3/2}\log n}\right).$$

To apply the chaining argument, consider ϵ size balls that cover Θ_n. Then the number of balls is $(C/\epsilon)^{d_0}$. Let B be the set of centers of the balls. Then we have

$$\mathbb{P}(\sup_{\theta\in B}|I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)]| \ge t) \le \exp\left(d_0\log\frac{C}{\epsilon} - \frac{t^2\sqrt{n}}{4\Delta s_0^{3/2}d_0^{3/2}\log n}\right).$$

Moreover, if $\theta_1, \theta_2 \notin D$ and $|\theta_1 - \theta_2| \le \epsilon$, then

$$\begin{aligned} & \left|I_2^{(k)}(\theta_1) - \mathbb{E}[I_2^{(k)}(\theta_1)] - I_2^{(k)}(\theta_2) + \mathbb{E}[I_2^{(k)}(\theta_2)]\right| \\ \le\ & \left|I_2^{(k)}(\theta_1) - I_2^{(k)}(\theta_2)\right| + \left|\mathbb{E}[I_2^{(k)}(\theta_1)] - \mathbb{E}[I_2^{(k)}(\theta_2)]\right|. \end{aligned}$$

Note that

$$\begin{aligned} & |I_2^{(k)}(\theta_1) - I_2^{(k)}(\theta_2)| \le \frac{1}{\|\theta_1\|_2\|\theta_2\|_2}\left|\sum_i \|\theta_2\|_2\|\theta_1\|_2 I_{2i}^{(k)}(\theta_1) - \sum_i \|\theta_1\|_2\|\theta_2\|_2 I_{2i}^{(k)}(\theta_2)\right| \\ \le\ & n^3 d_0^2 M^2\left(\|\theta_2\|_2\left|\sum_i \|\theta_1\|_2 I_{2i}^{(k)}(\theta_1) - \sum_i \|\theta_2\|_2 I_{2i}^{(k)}(\theta_2)\right| + |\|\theta_2\|_2 - \|\theta_1\|_2|\left|\sum_i \|\theta_2\|_2 I_{2i}^{(k)}(\theta_2)\right|\right) \\ \le\ & n^3 d_0^2 M_n^2\left(\Delta n\sqrt{\frac{d_0}{n}}\sqrt{s_0}M_n\epsilon + \epsilon n\sqrt{\frac{d_0}{n}}\sqrt{s_0}M_n\Delta\right) \\ =\ & 2n^{3.5}d_0^{2.5}s_0^{0.5}M_n^3\epsilon\Delta. \end{aligned}$$

Similarly,

$$\begin{aligned} \left|\mathbb{E}[I_2^{(k)}(\theta_1)] - \mathbb{E}[I_2^{(k)}(\theta_2)]\right| &\le n^3 d_0^2 M_n^2\left(\bar{f}\frac{d_0}{n}ns_0M_n^2\epsilon^2 + \bar{f}\epsilon n\Delta^2\frac{d_0}{n}s_0M_n^2\right) \\ &\le 2n^3 d_0^3 s_0 M_n^2\epsilon\Delta^2. \end{aligned}$$

If we choose t such that $\epsilon n^{3.5} d_0^3 M_n^3 = o(t)$ and $\epsilon n^3 d_0^4 M_n^2 = o(t)$ with ϵ being small enough, then the above implies that

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n\setminus D} |I_2^{(k)}(\theta) - \mathbb{E}[I_2^{(k)}(\theta)]| \geq t/2\right) \leq \exp\left(d_0 \log\frac{C}{\epsilon} - \frac{t^2\sqrt{n}}{4\Delta s_0^{3/2} d_0^{3/2}\log n}\right).$$

Hence

$$\mathbb{P}\left(\sup_{\theta\in\Theta_n\setminus D} |I_2(\theta) - \mathbb{E}[I_2(\theta)]|/\|\theta\|_2 \geq tK/2\right) \leq \exp\left(\log K + d_0 \log\frac{C}{\epsilon} - \frac{t^2\sqrt{n}}{4\Delta s_0^{3/2} d_0^{3/2}\log n}\right).$$

Letting $t = 3n^{-1/4}\Delta^{1/2} s_0^{3/4} d_0^{5/4} (\log n)^{3/2}$ and $\epsilon = n^{-9}$ with the growth condition $d_0 M_n^4 \log^2 n = o(n)$ yields that

$$\mathbb{P}(\sup_{\theta\in\Theta_n\setminus D} |I_2(\theta) - \mathbb{E}[I_2(\theta)]|/\|\theta\|_2 \geq 1.5n^{-1/4} K_n \Delta^{1/2} s_0^{3/4} d_0^{5/4} (\log n)^{3/2}) \leq n^{-9d_0\log n}.$$

Note that we have shown with probability at least $1 - \frac{2\bar{f}K_n}{n}$ that

$$\sup_{\theta\in D} |I_2(\theta) - \mathbb{E}[I_2(\theta)]| \leq \frac{2\bar{f} M_n \sqrt{d_0 s_0} K_n}{\sqrt{n}}.$$

Therefore, we have with probability at least $1 - n^{-9d_0\log n} - \frac{2\bar{f}K_n}{n}$,

$$\sup_{\theta\in\Theta_n} |I_2(\theta) - \mathbb{E}[I_2(\theta)]| \leq 2n^{-1/4} K_n \Delta^{1/2} s_0^{3/4} d_0^{5/4} (\log n)^{3/2}.$$

□

Proof of Lemma 2.10.9. We can easily see that

$$K\underline{f} k^2(s_0, 0) \leq \lambda_{\min}(A_n) \leq \lambda_{\max}(A_n) \leq K_n \bar{f}\phi(s_0),$$

$$\big(\sum_{k,k'} \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}\big) k^2(s_0, 0) \leq \lambda_{\min}(B_n) \leq \lambda_{\max}(B_n) \leq \big(\sum_{k,k'} \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}\big)\phi(s_0).$$

Hence

$$\begin{aligned}
\lambda_{\min}({A_n}^{-1} B_n {A_n}^{-1}) &\geq \lambda_{\min}^2({A_n}^{-1})\lambda_{\min}(B_n) \\
&= \lambda_{\max}^{-2}(A_n)\lambda_{\min}(B_n) \\
&\geq \bar{f}^{-2}\phi^{-2}(s_0) k^2(s_0, 0) \frac{\sum_{k,k'} \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}}{K_n^2} \\
&\geq \bar{f}^{-2}\phi^{-2}(s_0) k^2(s_0, 0)(\min_k \tau_k)(1 - \max_k \tau_k).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\lambda_{\max}({A_n}^{-1}B_n{A_n}^{-1}) &\leq \lambda^2_{\max}({A_n}^{-1})\lambda_{\max}(B_n) \\
&= \lambda^{-2}_{\min}(A_n)\lambda_{\max}(B_n) \\
&\leq L_0^{-2}\phi(s_0)k^{-4}(s_0,0)\frac{\sum_{k,k'}\tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}}{K_n^2} \\
&\leq L_0^{-2}\phi(s_0)k^{-4}(s_0,0),
\end{aligned}$$

which completes the proof. □

Proof of Lemma 2.10.10. Recall $\epsilon_i^{(k)} = y_i - {x_i}^T\beta(\tau_k) = y_i - \{z_i^{(k)}\}^T T(\mathcal{B}^o)$. Now define D_n as follows:

$$D_n = \alpha_n^T({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}n^{-0.5}\sum_i\sum_k z_i^{(k)}\{I(\epsilon_i^{(k)} < 0) - \tau_k\} := \sum_i Z_{ni},$$

where $Z_{ni} = (n^{-0.5})\Big[\alpha_n^T({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}\sum_k z_i^{(k)}\{I(\epsilon_i^{(k)} < 0) - \tau_k\}\Big]$. Then $\mathbb{E}[Z_{ni}] = 0$ and

$$\begin{aligned}
&\sum_i \mathrm{Var}(Z_{ni}) \\
= &\sum_i \alpha_n^T({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}\sum_{k,k'}\frac{1}{n}z_i^{(k)}\{z_i^{(k)}\}^T\{\min(\tau_k,\tau_{k'}) - \tau_k\tau_{k'}\}{A_n}^{-1}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n \\
= &\ \alpha_n^T({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}{A_n}^{-1}B_n{A_n}^{-1}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n \\
= &\ 1.
\end{aligned}$$

Consider an upper bound of Z_{ni} for all $i = 1, \ldots, n$:

$$|Z_{ni}| \leq n^{-0.5}\|\sum_k z_i^{(k)}\{I(\epsilon_i^{(k)} < 0) - \tau_k\}\|_2\|{A_n}^{-1}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n\|_2. \tag{2.52}$$

Since $\sum_k z_i^{(k)}\{I(\epsilon_i^{(k)} < 0) - \tau_k\}$ is a d_0-dimensional vector and the absolute value of each components is upper bounded by K_nM_n,

$$\|\sum_k z_i^{(k)}\{I(\epsilon_i^{(k)} < 0) - \tau_k\}\|_2 \leq \sqrt{d_0}KM_n. \tag{2.53}$$

Since $\|\alpha_n\|_2 = 1$, we have

$$\begin{aligned}
\|({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n\|_2 &\leq \lambda_{\max}\{({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\} \\
&= \{\lambda_{\min}({A_n}^{-1}B_n{A_n}^{-1})\}^{-0.5} \\
&\leq \bar{f}\phi(s_0)k^{-1}(s_0,0)(\min_k \tau_k)^{-0.5}(1-\max_k \tau_k)^{-0.5},
\end{aligned}$$

where the second inequality uses Lemma 2.10.9. Similarly,

$$\begin{aligned}
&\|{A_n}^{-1}({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n\|_2 \leq \lambda_{\max}({A_n}^{-1})\|({A_n}^{-1}B_n{A_n}^{-1})^{-\frac{1}{2}}\alpha_n\|_2 \\
\leq\ & K^{-1}L_0^{-1}k^{-3}(s_0,0)\bar{f}\phi(s_0)(\min_k \tau_k)^{-0.5}(1-\max_k \tau_k)^{-0.5}. \qquad (2.54)
\end{aligned}$$

Combing (2.53) and (2.54) with (2.52) yields

$$\max_i |Z_{ni}| \leq \sqrt{d_0/n}M_nL_0^{-1}k^{-3}(s_0,0)\bar{f}\phi(s_0)(\min_k \tau_k)^{-0.5}(1-\max_k \tau_k)^{-0.5}.$$

Hence

$$\begin{aligned}
\sum_i \mathbb{E}(|Z_{ni}|^3) &\leq \sum_i E(|Z_{ni}|^2)\sqrt{\frac{d_0}{n}}M_nL_0^{-1}k^{-3}(s_0,0)\bar{f}\phi(s_0)(\min_k \tau_k)^{-0.5}(1-\max_k \tau_k)^{-0.5} \\
&= \sqrt{d_0/n}M_nL_0^{-1}k^{-3}(s_0,0)\bar{f}\phi(s_0)(\min_k \tau_k)^{-0.5}(1-\max_k \tau_k)^{-0.5} \\
&\to 0.
\end{aligned}$$

Thus $\{Z_{ni}\}_{i=1}^n$ for all n are triangular array satisfying Lyapunov Condition. By applying central limit theorem for triangular arrays,

$$\sum_i Z_{ni} \to N(0,1),$$

which completes the proof. □

CHAPTER 3

Errors-in-Variables Regression

3.1 Introduction

There are classical and recent results which involve measurement error models. Rudelson and Zhou (2015) consider errors-in-variables regression with high dimensional covariates by allowing measurement errors that are possibly dependent across subjects. Liang and Li (2009) study variable selection for partially linear models when the covariates are measured with additive errors. Søresen et al. (2014) consider measurement error on linear regression with the Lasso penalty. They propose the method of correction for measurement error in the Lasso, and establish model selection consistency. In this chapter, we review the Kronecker sum covariance model and errors-in-variables regression problem as in Rudelson and Zhou (2015). We aim to compare Lasso-type and Conic-type estimators used in Rudelson and Zhou (2015) via simulations in terms of convergence rates. This chapter helps to find an appropriate errors-in-variables regression method which will be used to estimate the covariance matrices in Chapter 4.

3.2 The Model

In this section, we will first review the errors-in-variables regression model with dependent measurements studied in Rudelson and Zhou (2015). The Kronecker sum covariance model has been extensively studied in this context. Suppose that we observe $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times m}$ in the model

$$y = X_0\beta^* + \varepsilon, \tag{3.1a}$$

$$X = X_0 + W, \tag{3.1b}$$

where $\beta^* \in \mathbb{R}^m$, X_0 is a $n \times m$ matrix with independent rows, $\varepsilon \in \mathbb{R}^n$ is a noise vector and W is a mean zero $n \times m$ random noise matrix, independent of X_0 and ε, with independent column vectors $\omega^1, \ldots, \omega^m$. For the details of the model, see Model 3.2.1. For a scalar random variable V, recall the norm

$$\|V\|_{\psi_2} := \inf\{\mathrm{t} > 0 : \ \mathbb{E}(\exp(\mathrm{V}^2/\mathrm{t}^2)) \leq 2\}.$$

In Model 3.2.1, we define a specific description of the model (3.1a) and (3.1b).

Model 3.2.1. *(Rudelson and Zhou (2015)) Let Z be an $n \times m$ random matrix with independent entries Z_{ij} satisfying*

$$\mathbb{E}Z_{ij} = 0, \quad 1 = \mathbb{E}Z_{ij}^2 \leq \|Z_{ij}\|_{\psi_2} \leq K.$$

Let Z_1, Z_2 be independent copies of Z. Consider (3.1a)*, and let*

$$X \sim \mathcal{M}_{n,m}(0, A \oplus B), \quad \textit{where} \ \ A \oplus B := A \otimes I_n + I_m \otimes B \tag{3.2}$$

the Kronecker sum of positive definite $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$. This model means that

(1) *one covariance component* $A \otimes I_n$ *is used to describe the covariance of the signal* $\mathrm{vec}\{\,\mathrm{X}_0\,\}$, *where* $X_0 = Z_1 A^{1/2}$ *is an* $n \times m$ *random design matrix with independent subgaussian row vectors,*

(2) *and the other component* $I_m \otimes B$ *is used to describe that of the noise* $\mathrm{vec}\{\,\mathrm{W}\,\}$, *where* $W = B^{1/2} Z_2$ *is a noise matrix with independent subgaussian column vectors* $w^1, \ldots, w^m$, *independent of* X_0*;*

(3) *the error vector* $\varepsilon \in \mathbb{R}^n$ *is independent of* W *or* X_0, *with independent entries* ε_j *satisfying* $\mathbb{E}\varepsilon_j = 0$ *and* $\|\varepsilon_j\|_{\psi_2} \leq M_\varepsilon$.

3.3 The Lasso-type and Conic Programming Estimators

In this section, we will review the Lasso and Conic estimators in the errors-in-variables regression model studied by Rudelson and Zhou (2015), where these estimators can be used in nodewise regression method to estimate the inverse covariance matrices $\Theta = A^{-1}$ and $\Omega = B^{-1}$ in Chapter 4.

Rudelson and Zhou (2015) focus on deriving the statistical properties of two estimators for estimating β^* in (3.1a) and (3.1b) despite the presence of the additive error W in the observation matrix X. In the present work, we use the concentration of measure results derived in Rudelson and Zhou (2015) to derive the theoretical properties of the nodewise estimates.

Suppose that $\widehat{\mathrm{tr}}(B)$ is an estimator for $\mathrm{tr}(B)$; for example, if we know a $\mathrm{tr}(A)$, we can construct an estimator for $\mathrm{tr}(B)$ (Rudelson and Zhou, 2015):

$$\widehat{\mathrm{tr}}(B) \;=\; \tfrac{1}{m}\big(\,\|X\|_F^2 - n\mathrm{tr}(A)\big)_+, \qquad \widehat{\tau}(B) := \tfrac{1}{n}\widehat{\mathrm{tr}}(B) \geq 0, \tag{3.3}$$

where $(a)_+ = \max(a, 0)$. Let

$$\widehat{\Gamma} \;=\; \frac{1}{n}X^T X - \widehat{\tau}(B) I_m \quad \text{and} \;\; \widehat{\gamma} \;=\; \frac{1}{n}X^T y. \tag{3.4}$$

For chosen penalization parameters $\lambda, b_0 > 0$, we consider the following variant of the Lasso estimator, which provides regularized estimation with the ℓ_1-norm penalty,

$$\widehat{\beta} \;=\; \arg \min_{\beta : \|\beta\|_1 \le b_0} \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \langle\, \widehat{\gamma}, \beta \,\rangle \; + \lambda \|\beta\|_1, \tag{3.5}$$

which is a variation of the Lasso Tibshirani (1996) or the Basis Pursuit Chen et al. (1998) estimator.

For chosen penalization parameters $\lambda, \mu, \tau > 0$, we consider the following estimator:

$$\begin{aligned} \widehat{\beta} \;&=\; \arg\min \left\{ \|\beta\|_1 + \lambda t \;:\; (\beta, t) \in \Upsilon \right\} \text{ where} \\ \Upsilon \;&=\; \left\{ (\beta, t) \;:\; \beta \in \mathbb{R}^m, \; \left\| \widehat{\gamma} - \widehat{\Gamma}\beta \right\|_\infty \le \mu t + \tau, \; \|\beta\|_2 \le t \right\}, \end{aligned} \tag{3.6}$$

where $\widehat{\gamma}$ and $\widehat{\Gamma}$ are as defined in (3.4) with $\mu \sim \sqrt{\frac{\log m}{n}}$, $\tau \sim \sqrt{\frac{\log m}{n}}$ when $\mathrm{tr}(B)/n = \Omega(1)$. We refer to this estimator as the Conic programming estimator. Recently, Belloni et al. (2014) discuss the conic programming compensated matrix uncertainly (MU) selector, which is a variant of the Dantzig selector Candes and Tao (2007), Rosenbaum and Tsybakov (2010), and Rosenbaum and Tsybakov (2013).

3.4 Simulations

In this section, we use simulation studies to compare the two estimators when A and B follows AR(1) and Random graph.

1. AR(1) model : For $\rho_a \in (0,1)$, the covariance matrix A is of the form

$$A = \begin{bmatrix} 1 & \rho_a & \rho_a^2 & \cdots & \rho_a^{m-1} \\ \rho_a & 1 & \rho_a & \cdots & \rho_a^{m-2} \\ \rho_a^2 & \rho_a & 1 & \cdots & \rho_a^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_a^{m-1} & \rho_a^{m-2} & \rho_a^{m-3} & \cdots & 1 \end{bmatrix}.$$

2. Random graph: The graph is generated according to a Erdos-Renyi random graph model. Initially, we set $\Omega = I_{n\times n}$. Then we randomly select $s \in \{n, 2n\}$ edges and update Ω as follows: for each new edge (i, j), a weight $w > 0$ is chosen uniformly at random from $[0.1, 0.3]$; we subtract w from ω_{ij} and ω_{ji}, and increase ω_{ii} and ω_{jj} by w. And we multiply the constant $c > 0$ to the Ω such that $\mathrm{tr}(\mathrm{c}^{-1}\Omega^{-1}) = \mathrm{n}$, and $B := c^{-1}\Omega^{-1}$, which makes the trace of B equal to n.

For each experiment, we calculate the Signal-to-noise ratio used by Rudelson and Zhou (2015):

$$S/M := \frac{K^2\|\beta^*\|_2^2}{\tau_B^+ K^2\|\beta^*\|_2^2 + M_\epsilon^2}, \quad \tau_B^+ = \left(\sqrt{\tau_B} + \frac{2(\|A\|_2^{1/2} + \|B\|_2^{1/2})}{\sqrt{m}}\right)^2.$$

Figures 3.1 and 3.2 display the performances of the Lasso estimator defined in (3.5) using the constraint $\|\beta\|_1 \le R\|\beta^*\|_1$, where $R \in \{1, 2, 3, 7, 10\}$. We consider the case where $m = 400$, $n = 100$, and the coefficient $\beta^* = [0.9, \cdots, 0.9, 0, \cdots, 0]^T$, which has a sparsity level 10. Three metrics are included: relative errors in ℓ_2 norm, the probability of success for exact recovery of the sparsity pattern (success rate), FPs and FNs; see Table 3.1 for details of these metrics, where we use T and $\widehat{T}$ for

the support sets of $\widehat{\beta}$ and β^*, respectively. The estimates are not sensitive to the choice of R in the sense that it provides similar relative errors, success rate and FNs.

Figure 3.3 displays the results using the same settings in Figure 3.2 except that $\beta^* = [0.9, \cdots, 0.9, 0, \cdots, 0]^T$ has a sparsity level 20. Figure 3.4 displays the success rate of Lasso and thresholded Lasso estimator with thresholding level $\tau = \sigma\sqrt{\frac{\log m}{n}}$ when the β^* has moderate nonzero signal. This thresholding level works well in this numerical example. We can see that thresholding helps decrease false positives which yields a higher success rate than Lasso.

Table 3.1: Metrics

Metric	Definition
Relative errors in ℓ_2 norm	$\|\widehat{\beta} - \beta^*\|_2 / \|\beta^*\|_2$
Success rate	Probability of success of $\widehat{T} = T$
False positives (FPs)	$\|\widehat{T} \setminus T\|$
False negatives (FNs)	$\|T \setminus \widehat{T}\|$
True positives (TPs)	$\|\widehat{T} \cap T\|$
True negatives (TNs)	$\|\widehat{T}^c \cap T^c\|$
Recall	TPs/(TPs + FNs)
Precision	TPs/ (FPs+TPs)

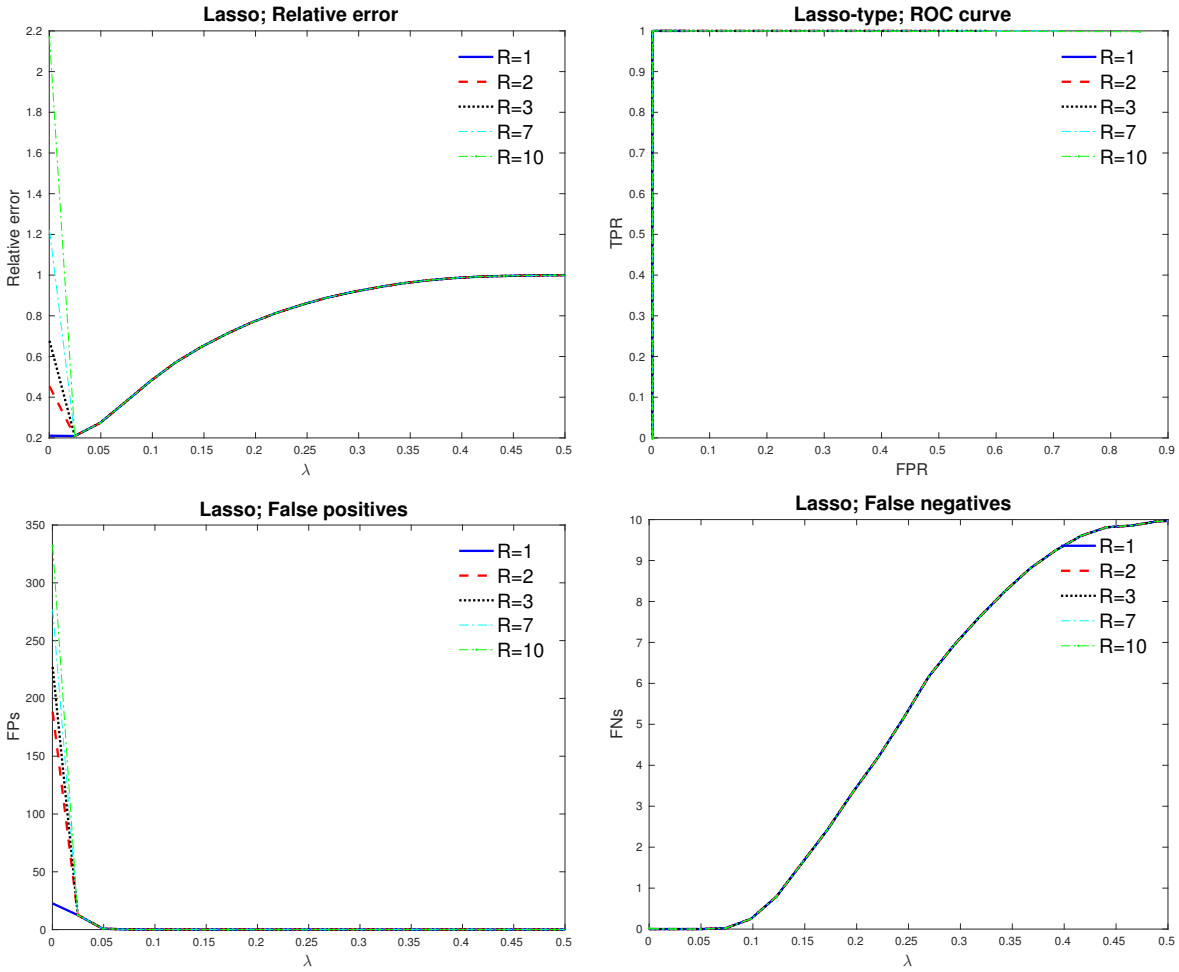

Figure 3.1: Plots for the Lasso estimator with a constraint $\|\beta\|_1 \leq R\|\beta^*\|_1$, where $\beta^* = [0.5, \cdots, 0.5, 0, \cdots, 0]^T$, where $d = 10 \asymp 0.6n/log(m)$. Step size $\eta = 2\|A\|_2$ are chosen. Five values are used for R and λ change from 0 to 0.5, when $m = 400$ and $n = 100$. A is generated using AR(1) model with parameter $\rho_A = 0.5$, and $B = 0.1B^*$, where B^* follows AR(1) model with parameter 0.8. The standard deviation of noise is $\sigma = 1$. The Signal-to-noise ratio S/M is 1.35.

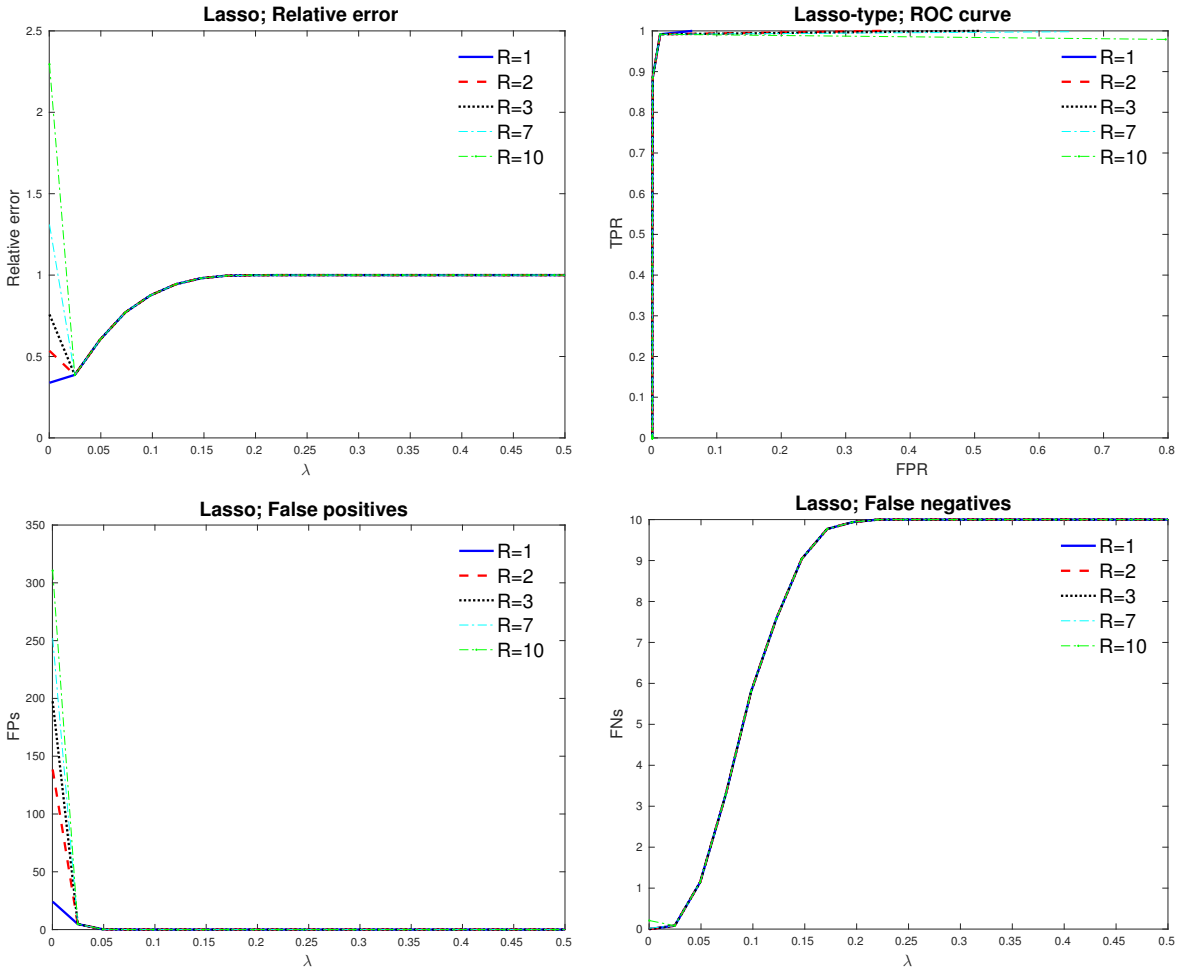

Figure 3.2: Plots for the Lasso estimator under the same settings used in Figure 3.1 except that $B = 0.7B^*$. The Signal-to-noise ratio S/M is 0.50.

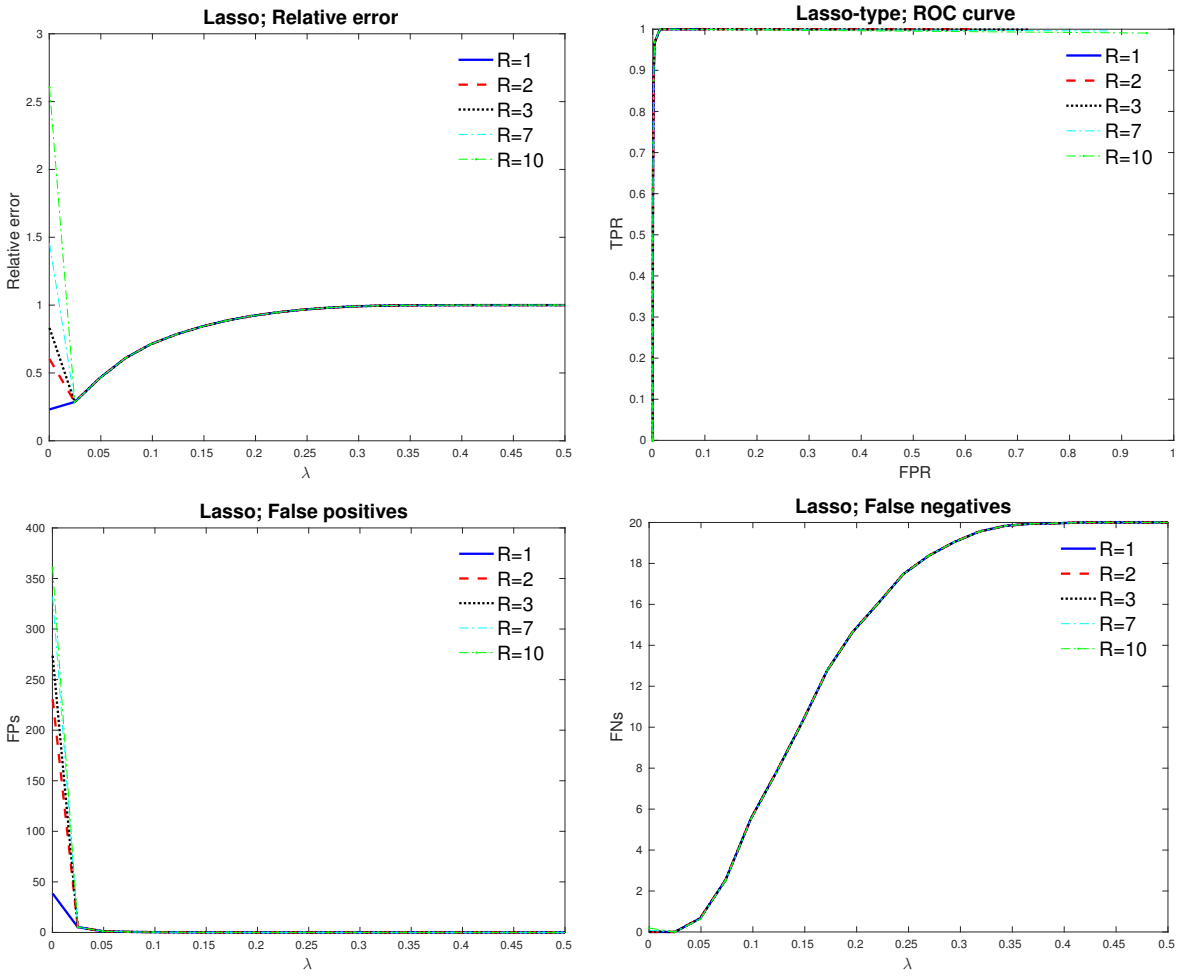

Figure 3.3: Plots for the Lasso estimator with a constraint $\|\beta\|_1 \leq R\|\beta^*\|_1$, where $\beta^* = [0.9, \cdots, 0.9, 0, \cdots, 0]^T$, where $d = 20 \asymp 1.2n/log(m)$. Step size $\eta = 2\|A\|_2$ are chosen. The other settings are exactly same as the one used in Figure 3.2. The Signal-to-noise ratio S/M is 0.60.

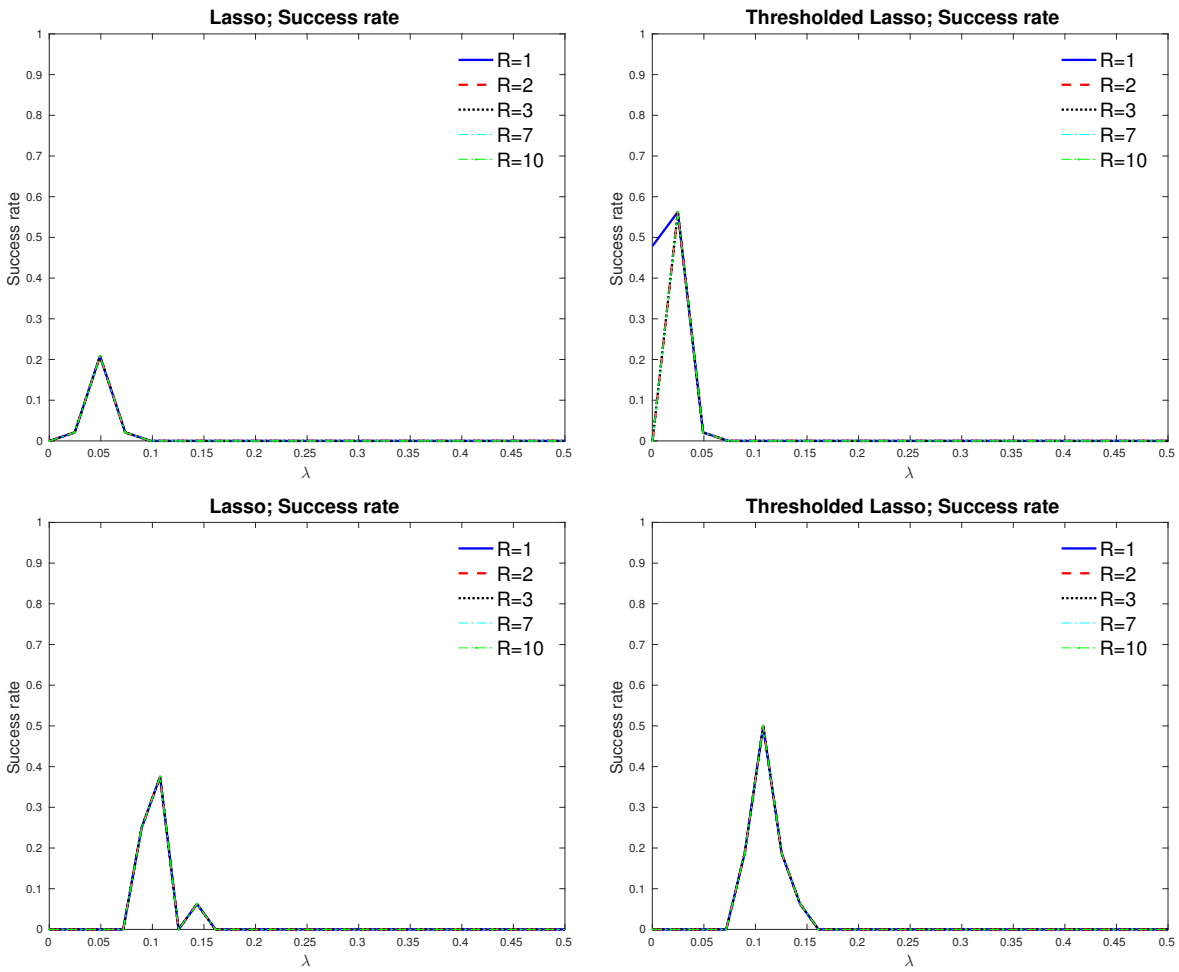

Figure 3.4: Plots for the Lasso estimator. The top plots are when $m = 400$, $n = 100$ and $\beta^* = [0.5, \cdots, 0.5, 0, \cdots, 0]^T$, where $d = 10$. The below plots are when $m = 600$, $n = 200$ and $\beta^* = [0.9, \cdots, 0.9, 0, \cdots, 0]^T$, where $d = 20$. A is generated using AR(1) model with parameter $\rho_A = 0.5$, and B follows random model. The standard deviation of noise is $\sigma = 1$. The Signal-to-noise ratio S/M for top and below are 0.50 and 0.65, respectively.

In Figure 3.5, we plot the relative errors in ℓ_2 norm for the Lasso (top) using $R = 2$ and the Conic (middle), with dimension $m \in \{128, 256, 512\}$, sample size $n \in [50, 2700]$, and the sparsity level $d = [\sqrt{m}]$. For the below plots, we use Conic with sparsity level $d = [m^{1/3}]$. The coefficient β^* has nonzero values between -1 and 1. The error versus the rescaled sample size $n/(d \log m)$ is also shown, where

the curves roughly align for different values of m for the Lasso estimates. With smaller sparsity level, Conic also show curves roughly aligned for different values of m. This shows that the Conic requires smaller sparsity level to obtain the theoretical convergence rate $\sqrt{(d \log m)/n}$, which is analyzed by Rudelson and Zhou (2015).

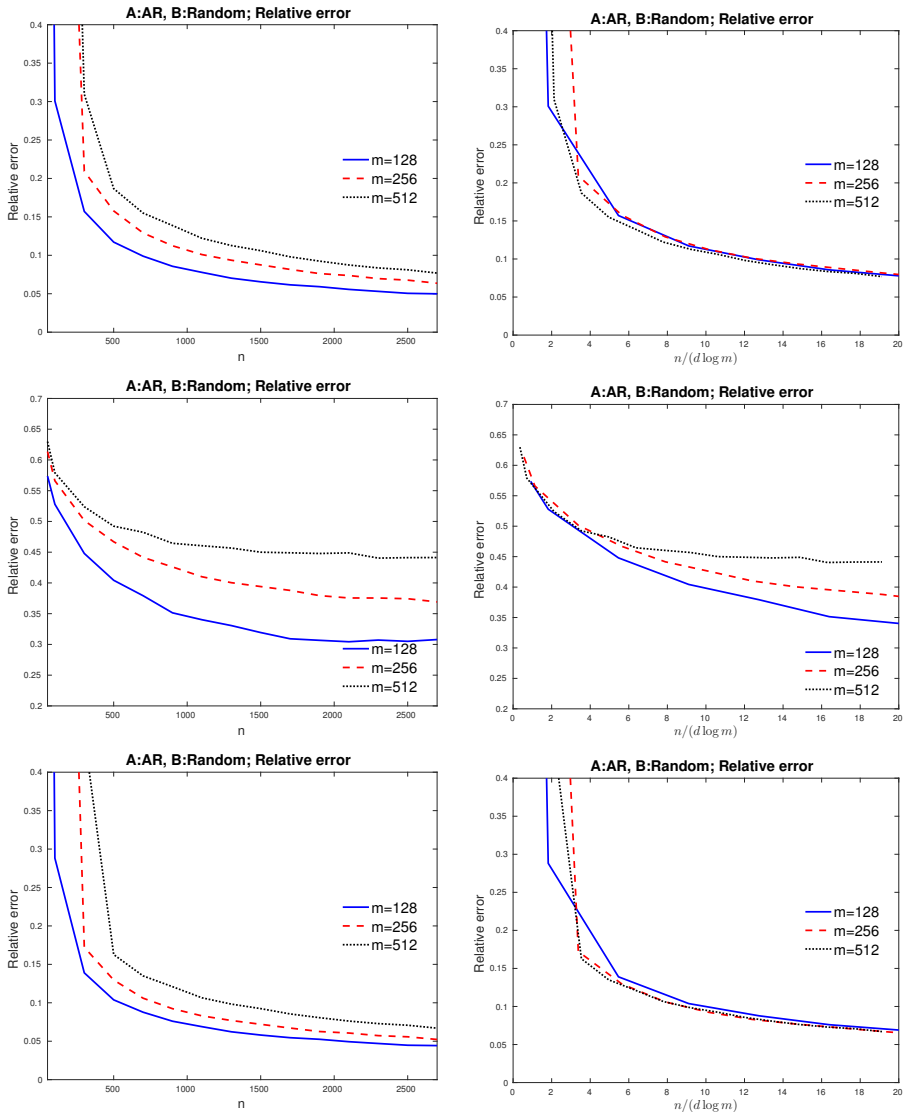

Figure 3.5: Plots for the relative error $\|\widehat{\beta} - \beta^*\|_2 / \|\beta^*\|_2$ of the Lasso estimator (top) and Conic estimator (middle) with the sparsity level $d = [\sqrt{m}]$. The below plot is for Conic with $d = [m^{1/3}]$. The left plot is an error plot with $m \in \{128, 256, 512\}$ and n changes from 50 to 2700. A is generated using AR(1) model with parameter $\rho_A = 0.5$, and B follows random model. The standard deviation of noise is $\sigma = 0.5$.

3.5 Optimization Error

In this section, we analyze simple approximate algorithms for solving the Lasso regression in (3.5). We use the composite gradient descent algorithm as studied in Agarwal et al. (2012) and Loh and Wainwright (2012). Let $L(\beta) := \frac{1}{2}\beta^T\widehat{\Gamma}\beta - < \widehat{r}, \beta >$, where $\widehat{\Gamma}$ and $\widehat{r}$ are defined in (3.4). The gradient of this loss function is $\nabla L(\beta) = \widehat{\Gamma}\beta - \widehat{r}$. The composite projected gradient descent algorithm produces a sequence of iterates $\{\beta^t,\ t = 0, 1, 2, \cdots, \}$ by

$$\beta^{t+1} = \underset{\|\beta\|_1 \leq b_0}{\arg\min}\, L(\beta^t) + < \nabla L(\beta^t), \beta - \beta^t > + \frac{\eta}{2}\|\beta - \beta^t\|_2^2 + \lambda\|\beta\|_1$$

with the stepsize parameter $\eta > 0$. This can be updated by two operations: Step 1 is soft thresholding the vector $\beta^t - \frac{1}{\eta}\nabla L(\beta^t)$ at a level λ. Step 2 is projecting the thresholded vector on to the ℓ_1 ball $\{\beta :\ \|\beta\|_1 \leq b_0\}$ by minimizing the Euclidean distance if the thresholded vector has ℓ_1-norm greater than b_0; see Agarwal et al. (2012) for details of the steps.

The following theorem shows that the composite gradient descent algorithm provides the solution near the global optimum $\widehat{\beta}$.

Theorem 3.5.1. *(Loh and Wainwright (2012)) Let ϕ denote the objective function of (3.5). Let β^t $(t = 0, 1, \cdots,)$ be the t^{th} iterate for composite projected gradient descent algorithm. For any optimum $\widehat{\beta}$ of the Lasso estimator in (3.5), there are absolute positive constants c_1 and c_2 and a contraction coefficient $r \in (0, 1)$, independent of m and n such that the iterates satisfy*

$$\|\beta^t - \widehat{\beta}\|_2^2 \leq c_1\|\widehat{\beta} - \beta^o\|_2^2, \quad \text{for all } t \geq T \asymp \log\frac{\phi(\beta^0) - \phi(\widehat{\beta})}{\|\widehat{\beta} - \beta^*\|_2^2}.$$

Figure 3.6 demonstrates the statistical error and the optimization error of the

above algorithm. We consider the errors-in-variables regression model in (3.1a) and (3.1b) when A and B are generated using AR(1) model with parameter $\rho_A = 1$ and the random graph model, respectively. The error vector ε has i.i.d. components from $N(0, \sigma^2)$, where $\sigma \in \{0.5, 1\}$. The iterates $\{\beta^t\}$ geometrically converge to the fixed point while the statistical error does not geometrically converge to zero. For fixed $\tau > 0$, it is seen that the iterates $\{\beta^t\}$ for $\sigma = 0.5$ requires larger T to achieve $\|\beta^t - \widehat{\beta}\|_2 \leq \tau$ for t $\geq$ T, compared to the iterates for $\sigma = 2$. This can be explained by the lower bound of t in Theorem 3.5.1, i.e. T tends to be increasing as $\|\beta^* - \widehat{\beta}\|_2$ decreases, which is the case when $\sigma = 0.5$ than when $\sigma = 2$.

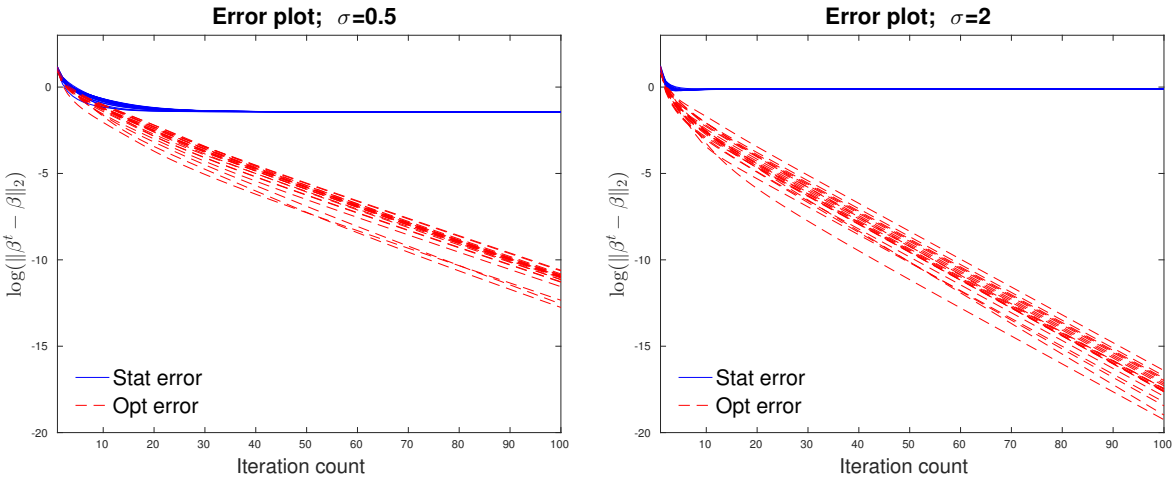

Figure 3.6: Plot for the optimization error $\log(\|\beta^t - \widehat{\beta}\|_2)$ and statistical error $\log(\|\beta^t - \beta^*\|_2)$ for each tth iterate. The blue lines and the red lines correspond to the statistical error and the optimization error, respectively. Each plot shows the solution path using 20 different starting points. We fix $m = 500$, $n = 200$ and $\beta^* = [1, 0.9, \cdots, 0.1, 0, \cdots, 0]^T$, where the first 10 components are non-zero. A and B are generated using AR(1) model with parameter $\rho_A = 1$ and the random graph model, respectively.

CHAPTER 4

Analysis of Kronecker Sum Model

4.1 Introduction

This chapter will demonstrate applications and methods for estimating the inverse covariance matrices $\Theta := A^{-1}$ and $\Omega := B^{-1}$ in the Kronecker sum covariance model,

$$\Sigma = A \otimes I_n + I_m \otimes B. \tag{4.1}$$

This is motivated by the data generating scheme considered in Model 3.2.1.

High-dimensional covariance estimation has been studied under the sparsity assumptions on the covariance matrix (Bickel and Levina, 2008; Lam and Fan, 2009; Rigollet and Tsybakov, 2012), and the precision matrix (Meinshausen and Bühlmann, 2006; Friedman et al., 2007; Rothman et al., 2008; Yuan, 2010; Zhou et al., 2011; Cai et al., 2011; Loh and Wainwright, 2012; Zhou, 2014). We rely on the sparsity assumption imposed on the precision matrix. The nodewise regression method (Meinshausen and Bühlmann, 2006; Yuan, 2010; Loh and Wainwright, 2012) requires the local sparsity restriction, i.e. the the number of nonzero entries on each column in the precision matrix is bounded. The sparsity assumption of the graphical Lasso method (Friedman et al., 2007; Rothman et al., 2008; Zhou, 2014) is imposed on

the entire precision matrix, i.e. the number of nonzero entries in the precision matrix is bounded. We consider the Kronecker sum model in (3.2), but now assume normality, which is needed only for the graphical model interpretation as shown in Definition 4.1.1.

Definition 4.1.1. *Let $Z = (Z_1, ..., Z_m)^T \sim \mathcal{N}_m(0, \Sigma)$ be a mean zero random Gaussian vector with covariance Σ. The corresponding undirected graph $G = (\mathcal{V}, F)$ is defined as follows: The vertex set $\mathcal{V} := \{1, \ldots, m\}$ has one vertex for each component of the vector Z. The edge set F consists of pairs (j, k) joined by an edge. If Z_j is independent of Z_k given the other variables, then $(j, k) \notin F$.*

Denote the precision matrix by $\Theta = (\theta_{ij}) := \Sigma^{-1} \in \mathbb{R}^{m \times m}$. Consider following m regressions, where we regress one variable against all others:

$$\begin{aligned} Z_i \quad &= \quad \sum_{j \neq i} \zeta_j^i Z_j + V_i \quad \text{where} \quad V_i \sim \mathcal{N}(0, \sigma_{V_i}^2) \text{ independent of } \{Z_j; j \neq i\} \\ \{i, j\} \in F \iff \theta_{ij} \neq 0 \iff &\zeta_i^j \neq 0 \text{ and } \zeta_j^i \neq 0 \text{ assuming that } \mathrm{Var}(V_i), \mathrm{Var}(V_j) > 0, \\ &\text{where } \zeta_j^i = -\theta_{ij}/\theta_{ii}, \quad \text{and} \quad \mathrm{Var}(V_i) := \sigma_{V_i}^2 = 1/\theta_{ii} \ (i, j = 1, \ldots, m). \end{aligned}$$

This method was proposed and studied in Meinshausen and Bühlmann (2006). Yuan (2010), Zhou et al. (2011), and Loh and Wainwright (2012) have also used the nodewise regression to estimate the covariance matrix and its inverse. For any matrix Z, let Z_j and Z_{-j} be the jth column of Z and the sub-matrix of Z without the jth column, respectively. In the current setting, nodewise regression can be written as following regression equation:

$$X_{0,j} = X_{0,-j}\beta^j + \varepsilon^j, \quad \text{where} \quad \beta^{\mathrm{j}} = \mathrm{A}_{-\mathrm{j},-\mathrm{j}}^{-1}\mathrm{A}_{-\mathrm{j},\mathrm{j}}, \tag{4.2}$$

and $\varepsilon_i^j \sim N(0, A_{jj} - A_{j,-j}A_{-j,-j}^{-1}A_{-j,j})$ is independent of $X_{0,-j}$ for $i = 1, \cdots, n$ and

$j = 1, \cdots, m$. By the inverse formula for block matrices, the inverse matrix $\Theta = A^{-1}$ satisfies

$$\Theta_{j,j} = (A_{jj} - A_{j,-j}\beta^j)^{-1}, \quad \Theta_{-j,j} = -(A_{jj} - A_{j,-j}\beta^j)^{-1}\beta^j. \tag{4.3}$$

In the above procedure, we can not observe the matrix X_0 due to the W. Therefore, we adapt the regression equation in (4.2) to our setting:

$$X_j = X_{0,-j}\beta^j + \varepsilon_j + W_j, \quad X_{-j} = X_{0,-j} + W_{-j}. \tag{4.4}$$

Here, we only observe X_j and X_{-j}, and the components in $\varepsilon_j + W_j$ are dependent due to W_j. To estimate β^j, we regress X_j on X_{-j} by using the Lasso-type errors-in-variables regression estimate described in (3.5) as we can interpret W_{-j} as measurement errors.

In Section 4.2, we describe details of the nodewise regression procedure to estimate Θ and Ω. Section 4.3 describes an alternative method (the graphical Lasso estimation) as a comparison with the nodewise regression method. Section 4.4 includes simulation study. In Section 4.5, we apply the Kronecker sum covariance model to analysis of hawkmoth neural encoding data studied in Sponberg et al. (2015). We demonstrate how the Kronecker sum model and measurement error regression techniques developed earlier are informative in this setting.

Before we proceed, we fix notations. For sequences $\{a_n\}$ and $\{\zeta_n\}$, we write $a_n = O(\zeta_n)$ to mean that $a_n \leq C\zeta_n$ for a universal constant $C > 0$. Similarly, $a_n = \Omega(\zeta_n)$ when $a_n \geq C'\zeta_n$ for some universal constant $C' > 0$. We use $a_n = o(\zeta_n)$ to mean that for every positive constant ϵ, there exists a constant N such that $|a_n| \leq \epsilon|b_n|$ for all $n \geq N$. Similarly $s_n = \omega(\zeta_n)$ when for every positive constant ϵ, there exists a constant N such that $|a_n| \geq \epsilon|b_n|$ for all $n \geq N$. In Table 4.1, we

define other notations used in this chapter.

Table 4.1: The Notations

Parameters	Definitions
X_i	The ith column of a matrix X
X_{-i}	The sub-matrix of X without the ith column
$\text{diag}(\Sigma)$	The diagonal matrix of a square matrix Σ
$\lambda_{\max}(\Sigma)$	The maximum eigenvalue of a square matrix Σ
$\lambda_{\min}(\Sigma)$	The minimum eigenvalue of a square matrix Σ
$\kappa(\Sigma)$	The condition number of a square matrix Σ
$\tau(\Sigma)$	$\text{tr}(\Sigma)/\text{p}$ for a square matrix $\Sigma \in \mathbb{R}^{p\times p}$
$\|\Sigma\|_{0,\text{off}}$	The number of nonzero non-diagonal entries in a square matrix Σ
$\|\Sigma\|_{1,\text{off}}$	$\sum_{i\neq j} \|\Sigma_{ij}\|$ for a square matrix Σ
$\|\Sigma\|_1$	$\max_j \sum_i \|\Sigma_{ij}\|$ for a matrix Σ
$\mathcal{S}^m$	The set of symmetric $m \times m$ matrices
$\mathcal{S}_+^m$	The set of positive symmetric $m \times m$ matrices
$\text{tr}(\Sigma)$	The trace of a square matrix Σ
Θ	A^{-1}
Ω	B^{-1}

4.2 Nodewise Regression Procedure

In this section, we describe the nodewise regression procedure to estimate the covariance matrix A and its inverse following Loh and Wainwright (2012).

Step 1: Nodewise regression: To construct an estimator for $\Theta = A^{-1}$ with $X = X_0 + W$ as in (3.1b), we obtain m vectors of $\widehat{\beta}^i$ for $i = 1, \ldots, m$ by solving (3.5) with

$$\widehat{\Gamma} = \tfrac{1}{n}(X_{-i})^T X_{-i} - \widehat{\tau}(B) I_{m-1} \text{ and } \widehat{\gamma} = \tfrac{1}{n}(X_{-i})^T X_i, \tag{4.5}$$

where $\widehat{\tau}(B)$ is defined in (3.3).

Step 2: Intermediate step: To exploit (4.3), define an initial estimate of A:

$$\widetilde{A} = \frac{1}{n} X^T X - \widehat{\tau}(B) I_m. \tag{4.6}$$

Denote $\widehat{a}_j = -(\widetilde{A}_{jj} - \widetilde{A}_{j,-j}\widehat{\beta}^j)^{-1}$. Based on the equation (4.3), define $\widetilde{\Theta}$:

$$\widetilde{\Theta}_{j,-j} = \widehat{a}_j \widehat{\beta}^j, \qquad \widetilde{\Theta}_{jj} = -\widehat{a}_j. \tag{4.7}$$

Step 3: Symmetrization: To obtain a symmetric matrix for the estimate, consider

$$\widehat{\Theta} = \arg\min_{\Sigma \in \mathcal{S}^m} \|\Sigma - \widetilde{\Theta}\|_1. \tag{4.8}$$

The matrices $\widehat{\Theta}$ is an estimate of Θ.

Algorithm 4.2.1. *(Nodewise procedure for estimating Θ)*

1. *Perform m Lasso-type errors-in-variables regressions in* (3.5) *with*

$$\widehat{\Gamma} = \frac{1}{n} X_{-j}^T X_{-j} - \widehat{\tau}(B) I_m, \quad \widehat{\gamma} = \frac{1}{n} X_{-j}^T X_j,$$

   where $\widehat{\tau}(B)$ is defined in (3.3)*. Let $\widehat{\beta}^j$ be the estimates for $j = 1, \cdots, m$.*

2. *Denote $\widehat{a}_j = -(\widetilde{A}_{jj} - \widetilde{A}_{j,-j}\widehat{\beta}^j)^{-1}$, where $\widetilde{A} = \frac{1}{n} X^T X - \widehat{\tau}(B) I_m$.*

   Form $\widetilde{\Theta}$ with $\widetilde{\Theta}_{j,-j} = \widehat{a}_j \widehat{\beta}^j$ and $\widetilde{\Theta}_{jj} = -\widehat{a}_j$.

3. *Set $\widehat{\Theta} = arg\,min_{\Sigma \in S^m} \|\Sigma - \widetilde{\Theta}\|_1$.*

Similarly, we can obtain the estimates of $\Omega := B^{-1}$. The theoretical and further methodological development is an on-going joint work with Shedden and Zhou. This nodewise regression based covariance matrix estimation does not guarantee positiveness of the estimates (Meinshausen and Bühlmann, 2006; Yuan, 2010; Cai et al., 2011; Loh and Wainwright, 2012).

4.3 Projected Graphical Lasso Method

In this section, we consider the projected Graphical Lasso method (projected GLasso) as an alternative to the nodewise regression method. Recall the initial estimate of A defined in (4.6):

$$\widetilde{A} := \frac{1}{n} X^T X - \widehat{\tau}_B I_m. \tag{4.9}$$

Consider the following projection:

$$\widetilde{A}_+ = \underset{C \in \mathcal{S}_+^m}{\arg\min} \left\{ \left\| \widetilde{A} - C \right\| \right\}, \tag{4.10}$$

where $\|\cdot\|$ can take the operator norm, the Frobenius norm or the matrix $\|\cdot\|_\infty$ norm which bounds the maximum absolute entry-wise distance. Similarly, let $\widetilde{B}_+$ be the projected matrix of $\widetilde{B}$. We consider the maximum absolute entry-wise norm in (4.10). Algorithms for solving such estimators (4.10) have been intensively discussed Schwertmanand and Allen (1979), Higham (2002), Malick (2004), Gao and Sun (2010), and Henrion and Malick (2012), where one can generally replace $\mathcal{S}_+^m$ with more constraints when more about the covariance structure is known. The literature recognizes the optimization problem as the constrained nearest correlation matrix problem (Gao and Sun, 2010).

Consider the correlation matrix of $\widetilde{A}_+$,

$$\widetilde{A}_1 := \mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2} \widetilde{\mathrm{A}}_+ \mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2}. \tag{4.11}$$

Let $\widehat{A}_1$ and $\widehat{B}_1$ be the graphical lasso estimates:

$$\widehat{A}_1 \quad = \quad \underset{\Sigma \succ 0}{\arg\min} \left(\mathrm{tr}(\widetilde{A}_1 \Sigma^{-1}) + \log|\Sigma| + \lambda_A |\Sigma^{-1}|_{1,\mathrm{off}} \right), \tag{4.12}$$

where λ_A is a regularization parameter. The estimates of A is

$$\widehat{A} := \mathrm{diag}(\widetilde{\mathrm{A}}_+)^{1/2}\widehat{\mathrm{A}}_1\mathrm{diag}(\widetilde{\mathrm{A}}_+)^{1/2}. \tag{4.13}$$

Algorithm 4.3.1. *(projected GLasso procedure for estimating Θ)*

1. *Consider the projection $\widetilde{A}_+ = arg\,min_{C\in\mathcal{S}_+^m}\left\{\left\|\widetilde{A}-C\right\|\right\}$, where $\widetilde{A}$ is the initial estimate of A as described in* (4.9).

2. *Let* $\widetilde{A}_1 = \mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2}\widetilde{\mathrm{A}}_+\mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2}$.

3. *Solve the Graphical Lasso program:*

$$\widehat{\Theta}_1 = \underset{\Sigma\in\mathcal{S}_+^m}{arg\,min}\{\mathrm{tr}(\widetilde{A}_1\Sigma^{-1}) + \log|\Sigma| + \lambda_A|\Sigma^{-1}|_{1,\mathrm{off}}\}.$$

4. *The estimate of Θ is* $\widehat{\Theta} := \mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2}\widehat{\Theta}_1\mathrm{diag}(\widetilde{\mathrm{A}}_+)^{-1/2}$.

Similarly, we can obtain the estimates of Ω.

4.4 Simulations

In this section, we perform simulations to investigate the performances of the proposed covariance matrix estimators. For the topologies of A and B, see Section 3.4. We compare the nodewise regression and the projected GLasso methods for the estimation of A and B. We fix $m = 400$ and $n = 100$. We repeat 200 times and record the average of the performances in the simulation study. For the projected GLasso, we obtain the estimates by using λ_A and λ_B from the interval $(0, 1)$. For the nodewise regression, we use the constraint $\|\beta\|_1 \leq 2\|\beta^*\|_1$ and the penalty parameter $\lambda \in (0, 0.31)$.

In Figure 4.1, we compare the relative Frobenius error of the nodewise regression and the projected GLasso estimates by changing the regularization parameters. It is seen that the performance of nodewise regression is better for the estimation of the covariance matrix A when $\mathrm{tr}(A)$ is large. However, in terms of estimating B, projected GLasso is better when $\tau(A)$ is less than $\tau(B)$. Overall, nodewise regression method seems to work well when estimating the covariance matrix A, which has relatively larger dimension than B. Figure 4.2 displays of ROC curve for the two estimates when $\mathrm{tr}(A) = 1.5m$. Nodewse regression provides more accurate edge selection than projected GLasso. In Figures 4.3, we show Recall and Precision of nodewise regression estimates for A and B, respectively. It is observed that the estimator has accurate edge selection when the covariance matrix has larger trace. Figure 4.4 displays relative L2 error curves aligned against a rescaled sample size for $m \in \{128, 256, 512\}$ cases. We see the agreement for the three cases, which demonstrate the L2 error bound rate $d\sqrt{\log m}/n$.

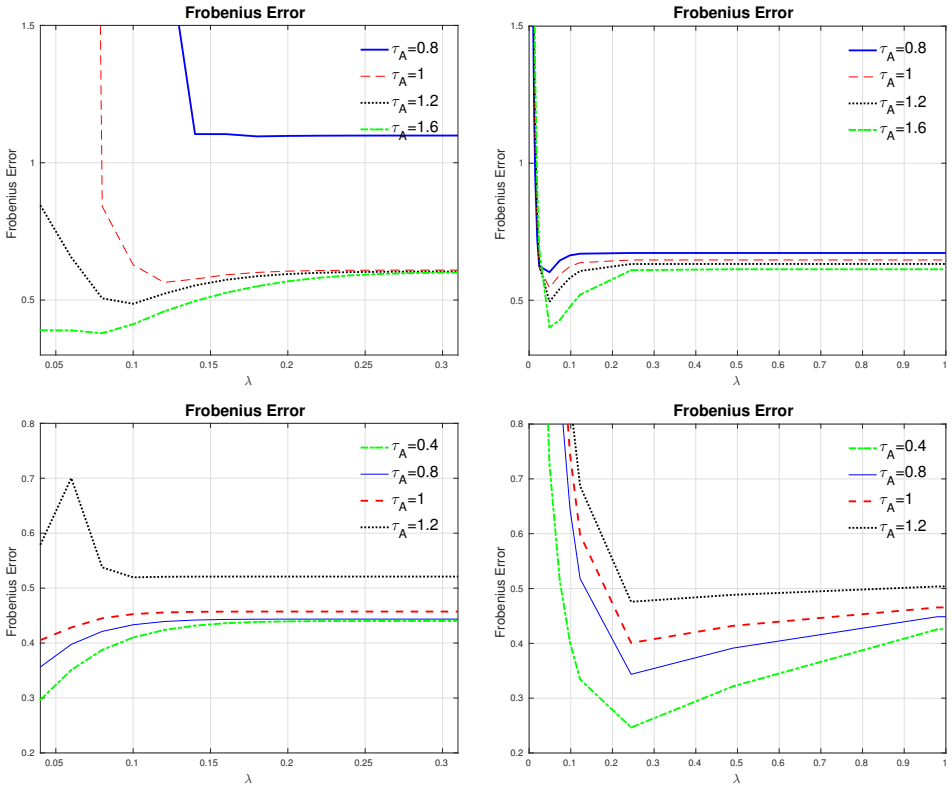

Figure 4.1: The relative Frobenius error of the estimates $\widehat{A} = \widehat{\Theta}^{-1}$ (top) and $\widehat{B} = \widehat{\Omega}^{-1}$ (below) when $m = 400$, $n = 100$ and the covariance matrix A is AR(1). The left figure and the right figure show the relative Frobenius error of nodewise regression estimate and projected GLasso estimate, respectively.

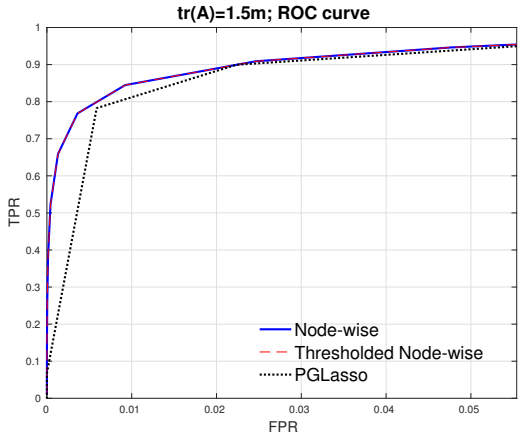

Figure 4.2: The ROC curve of the estimates $\widehat{\Theta}$ when $m = 400$, $n = 100$. The left figures and the right figures are when $\tau(A) = 1.5$.

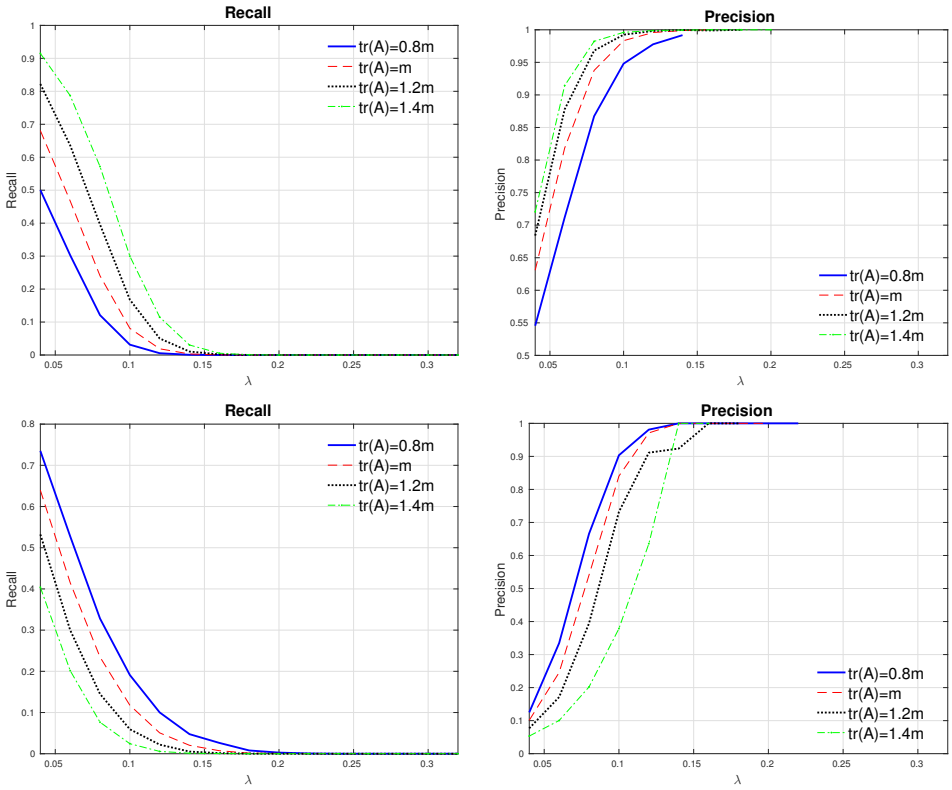

Figure 4.3: The recall and precision curves of the nodewise regression estimate of Θ (top) and Ω (below), respectively, when $m = 400$, $n = 100$ and the covariance matrix A is AR(1).

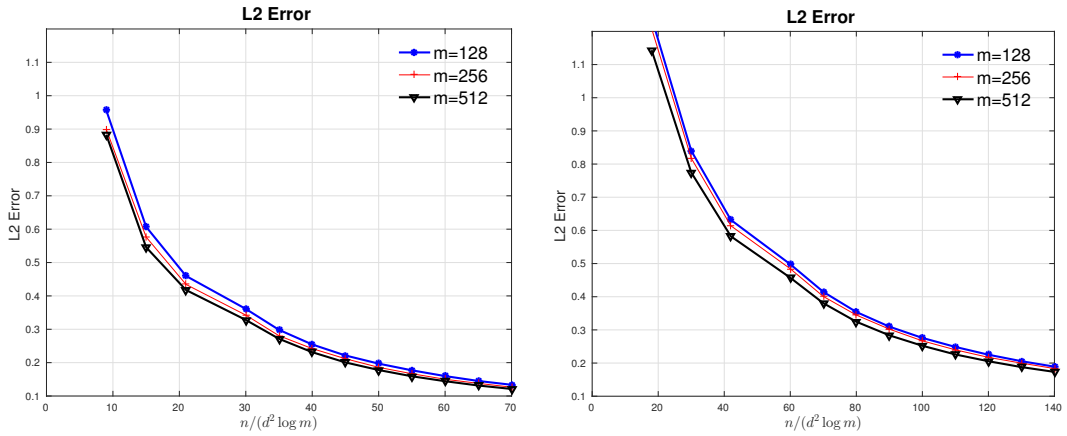

Figure 4.4: The L2 error for the nodewise regression estimate $\widehat{A}$ when A is AR(1) and $B = 0.1B^*$ (left) or $B = 0.5B^*$, where B^* follows Random graph. The error versus the rescaled sample size $n/(d^2 \log m)$ are plotted for three different m cases.

To understand the relative performances of the two estimates with different trace values of A, we record the performances of the estimates when $\tau(A)$ has a value from 0.1 to 1.9. We generate the Kronecker sum covariance matrix $\Sigma = cA^* \oplus (2-c)B^*$, where A^* and B^* follow AR(1) and random graph model, respectively. Here the matrices $A := cA^*$ and $B := (2-c)B^*$ are the parameters to be estimated. In Figure 4.5, the relative Frobenius error and L2 error are recorded for the two estimates. This shows that the nodewise regression seems to be favorable when $\mathrm{tr}(A)/m$ is larger than 1. When $\mathrm{tr}(A)/m$ is small, projected GLasso has lower error rates. Overall, for the estimation of the larger dimensional covariance matrix A, nodewise regression works well than projected GLasso when $\mathrm{tr}(A)/m \geq \mathrm{tr}(B)/n$ for this example.

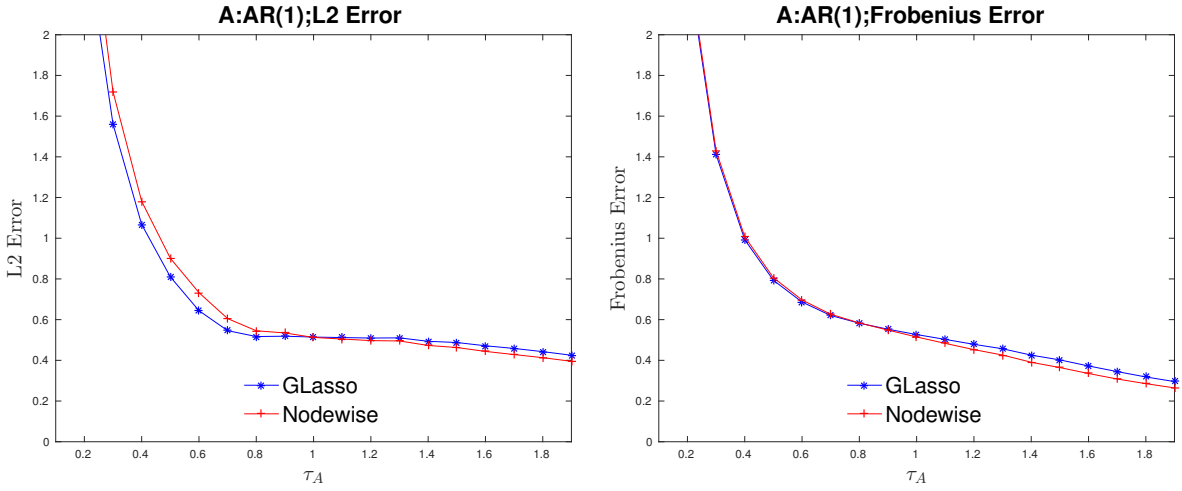

Figure 4.5: The performance results of the estimates $\widehat{A}$ when $\tau(A)$ increases from 0.1 to 1.9; the dimensions are fixed at $m = 200$ and $n = 200$; the two plots show the L2 error and Frobenius error, respectively.

4.5 Analysis of Hawkmoth Neural Encoding Data

In this section, we fit the Kronecker sum covariance model to the data studied in Sponberg et al. (2015). The basic idea is to study how the turning behavior

of hawkmoths in flight is related to the firing of neurons in their left and right dorsolongitudinal muscles (DLMs). The difference in peak firing times between the left and right DLM controls turning while the moth is in flight. The measured difference in peak DLM firing time can be correlated with measured torque values. As a baseline hypothesis, we anticipate a positive correlation between neural firing and mean torque within each wingstroke.

In the experiment, both torque and neural firing were measured for up to 1020 wingstrokes of moths while in flight. Data were collected for seven moths. The neural firing was measured using electrical probes, and the torque was measured via a torque meter using an optical sensor. Separate left and right neural firing rates were captured, and the time of peak firing prior to each wingstroke was obtained. Thus, for each wingstroke we have single readings of peak neural firing time in the left and right DLM, and a timecourse of 500 sampled torque values spanning the duration of the wingstroke.

In this section, we demonstrate how the Kronecker sum model and measurement error regression techniques developed earlier are informative in this setting. The rationale for considering the Kronecker sum model for the torque time series is that these measurements are difficult to obtain and are known to be quite noisy. Therefore we consider whether the Kronecker sum model might be capable of separating the underlying motion signal from measurement error. We then were able to calculate what the relationship between neural firing and flight torque might have been based on more accurate torque measurements.

The outline of this section is as follows. In Subsection 4.5.1, we propose a method for assessing the goodness of fit for Kronecker sum and product models to a given

dataset, and argue that the Kronecker sum model provides a better fit for the moth torque data than the Kronecker product model. In Subsection 4.5.2, we introduce a method to determine the trace of covariance matrix for Kronecker sum model. In Subsection 4.5.3, we consider the graphical structures for the temporally dependent signal component and for the signal component capturing dependence among trials. We show that the temporally dependent signal component has an approximately stationary time series structure, and the trial-by-trial correlations reflect similarities between trials that are weakly connected to the mean torque of the trial. In Subsection 4.5.4, we further consider the relationship between the mean structure of the data and the residual covariances among trials, viewing this as a form of mean-covariance relationship. In Subsection 4.5.5, we use measurement error regression techniques from chapter 3, and argue that the relationship between neural firing and torque is stronger than apparent through simpler analyses. To the extent that the coefficient patterns found when using measurement error regression are interpreted as being constant, this is consistent with the simple 1-dimensional neuro-encoding hypothesis, and suggests that any evidence for complementary neural-encoding pathways may be an artifact resulting from not fully accounting for the presence of measurement errors.

4.5.1 Fit of the Additive Covariance Model

We propose a method for assessing the goodness of fit for Kronecker sum and product models to a given dataset. If X follows the Kronecker sum covariance model

with covariance matrix $\Sigma = A \oplus B$, then

$$\Sigma = \begin{bmatrix} \Sigma(1,1) & \Sigma(1,2) & \cdots & \Sigma(1,m) \\ \Sigma(2,1) & \Sigma(2,2) & \cdots & \Sigma(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(m,1) & \Sigma(m,2) & \dots & \Sigma(m,m) \end{bmatrix} = \begin{bmatrix} a_{11}I_n + B & a_{12}I_n & \cdots & a_{1m}I_n \\ a_{21}I_n & a_{22}I_n + B & \cdots & a_{2m}I_n \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}I_n & a_{m2}I_n & \dots & a_{mm}I_n + B \end{bmatrix}.$$

The sample covariance matrix $\widehat{S} := \mathrm{vec(X)vec(X)^T} \in \mathrm{R^{mn \times mn}}$ is an unbiased but noisy estimate of the covariance matrix, for $\widehat{S}(i,j) = X_{\cdot,i}X_{\cdot,j}^T \in \mathbb{R}^{n\times n},\ 1 \le i,j \le m$,

$$\widehat{S} = \begin{bmatrix} \widehat{S}(1,1) & \widehat{S}(1,2) & \cdots & \widehat{S}(1,m) \\ \widehat{S}(2,1) & \widehat{S}(2,2) & \cdots & \widehat{S}(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{S}(m,1) & \widehat{S}(m,2) & \dots & \widehat{S}(m,m) \end{bmatrix}_{(mn)\times(mn)}$$

To assess the goodness of fit of the Kronecker sum model, we use the fact that if the covariance matrix has the form of $\Sigma = A \oplus B$, then $\Sigma(i,j)_{k,\ell} = 0$ for all $1 \le k \ne \ell \le n$ and $1 \le i \ne j \le m$. Define the statistic S_{off}

$$S_{\text{off}} := \sum_{1\le i\ne j\le m} \sum_{1\le k\ne \ell\le n} \widehat{S}(i,j)_{k,l}/(\sqrt{mn}\|X\|_F^2),$$

If the true covariance matrix Σ is near the Kronecker sum space, then S_{off} should be close to zero. The statistic S_{off} is also scale-invariant in X. This statistic can be efficiently calculated as follows: Let $X_{k\cdot}$ and $X_{\cdot l}$ be the sum of the entries on the kth row and the lth column of X, respectively. The statistic S_{off} can be efficiently calculated using

$$\sqrt{mn}\|X\|_F^2 S_{\text{off}} = (\sum_{i,j} X_{ij})^2 - \sum_{k=1}^{m}(X_{k\cdot})^2 - \sum_{l=1}^{n}(X_{\cdot l})^2 + \|X\|_F^2.$$

Remark 4.5.1. If a model follows Kronecker product, i.e. $\Sigma = A \otimes B$, then

$$\mathbb{E}\left[\sum_{1\le i\neq j\le m}\ \sum_{1\le k\neq \ell\le n} \widehat{S}(i,j)_{k,l}\right] = \left(\sum_{1\le i\neq j\le m} A_{ij}\right)\left(\sum_{1\le k\neq \ell\le n} B_{k\ell}\right).$$

If A follows AR(1) model with a parameter ρ, then we can show

$$\sum_{1\le i\neq j\le m} A_{ij} = 2\frac{\rho-(m+1)\rho^{m+1}+m\rho^{m+2}}{(1-\rho)^2} \asymp 2m\frac{\rho}{1-\rho},$$

provided that m is large enough.

Figure 4.6 displays the simulated S_{off} for 500 samples generated from Kronecker sum or product when A is Star-Block (left) and AR(1) (right), respectively. It is seen that the simulated statistic for the Kronecker sum model is more concentrated around zero than those for the Kronecker product model.

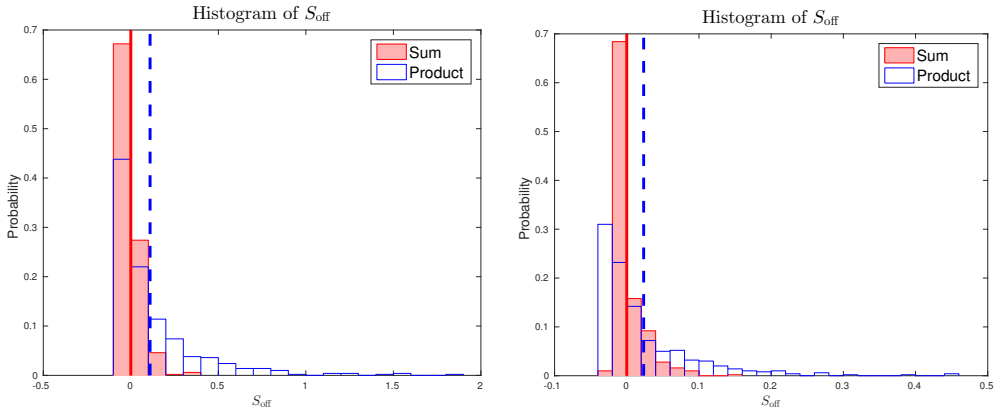

Figure 4.6: Histogram of the statistic S_{off} for 500 samples generated from Kronecker sum or product when A is Star-Block (left) and AR(1) (right), respectively. The blue vertical line is the expected value of the statistic for the Kronecker product model. For the Star-Block model, the (mean, standard deviation) from the sum and product model are $(0.0016, 0.055)$ and $(0.1184, 0.2629)$, respectively. For the AR(1) model, the sum and product model have $(-0.0005, 0.0223)$ and $(0.0206, 0.0671)$, respectively.

If we only have one data realization X, then we can simulate the reference distribution of S_{off} using the parametric Bootstrap. Let $\widehat{A} \oplus \widehat{B}$ and $\widetilde{A} \otimes \widetilde{B}$ be the

Kronecker sum and product estimates, respectively. We generate random samples $X^*(t)$ $(t = 1, \cdots, N)$ from $N(0, \widehat{A} \oplus \widehat{B})$ and $Z^*(t)$ $(t = 1, \cdots, N)$ from $N(0, \widetilde{A} \otimes \widetilde{B})$, respectively. For each $X^*(t)$, let $\widehat{S}^* = \mathrm{vec(X^*(t))vec(X^*(t))^T} \in \mathrm{R^{mn \times mn}}$. We calculate

$$S^*_{\text{off}} := \sum_{1 \le i \ne j \le m} \sum_{1 \le k \ne \ell \le n} \widehat{S}^*(i,j)_{k,l} / (\sqrt{mn} \|X^*\|_F^2),$$

Similarly, we calculate the statistic for the samples $Z^*(t)$ $(t = 1, \cdots, N)$. We then compare the simulated statistic S^*_{off} from the samples $X^*(t)$ $(t = 1, \cdots, N)$ and $Z^*(t)$ $(t = 1, \cdots, N)$ with the observed statistic S_{off} from the original data X, and choose the model which provides closer simulated statistics to the observed one. Figure 4.7 displays the histogram of the simulated statistic S^*_{off} for 500 bootstrap samples generated from Kronecker sum and product estimates. This shows that the observed statistic S_{off} is closer to the simulated S^*_{off} when the samples are generated from the estimates for their true covariance structure, i.e. Kronecker sum or product.

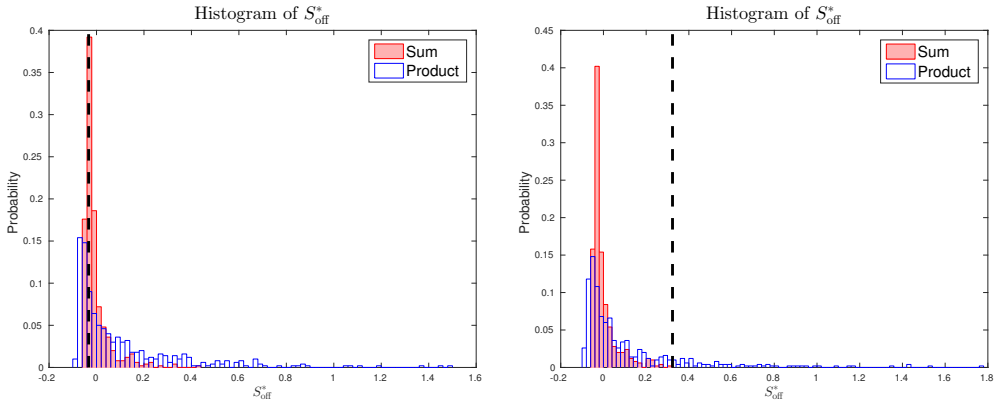

Figure 4.7: Histogram of the statistic S^*_{off} for 500 bootstrap samples generated from Kronecker sum and product estimates. The left and right figures are when the observed data X follows Kronecker sum and product models, respectively. The black vertical line indicates one observed value of S_{off} from a Kronecker sum (left) product model (right).

Next we apply the goodness of fit test to the hawkmoth torque data. In Table 4.2, we record the observed statistics S_{off} for each moth. We also simulate the reference

distribution of S_{off} using the parametric Bootstrap. The table shows that the mean of the simulated S^*_{off} for the Kronecker sum estimate has a closer value to the observed statistic than that of the Kronecker product estimate. This demonstrates that the Kronecker sum model may be more appropriately explain the movement data than the Kronecker product model.

Table 4.2: The simulated statistic S^*_{off}

Moth	Sum		Product		Observed
	Mean($\times 10^{-3}$)	SD ($\times 10^{-3}$)	Mean ($\times 10^{-3}$)	SD ($\times 10^{-3}$)	
J	0.5	2.3	14.1	24.0	2.1
K	0.3	2.4	11.7	18.3	2.7
L	0.2	2.5	16.3	25.8	2.6
M	0.1	2.8	15.0	21.7	2.6
N	0.3	2.4	18.3	25.3	2.4
P	0.1	2.5	16.8	27.1	2.3
Q	0.1	2.9	11.7	19.0	3.1

In Figure 4.8, we draw histograms of S^*_{off} obtained from 200 random samples generated from the Kronecker sum and product fits for moth J and moth L. As the figure demonstrates, the simulated statistics from the Kronecker sum estimate tend to concentrate more around the observed statistic compared to the Kronecker product estimate.

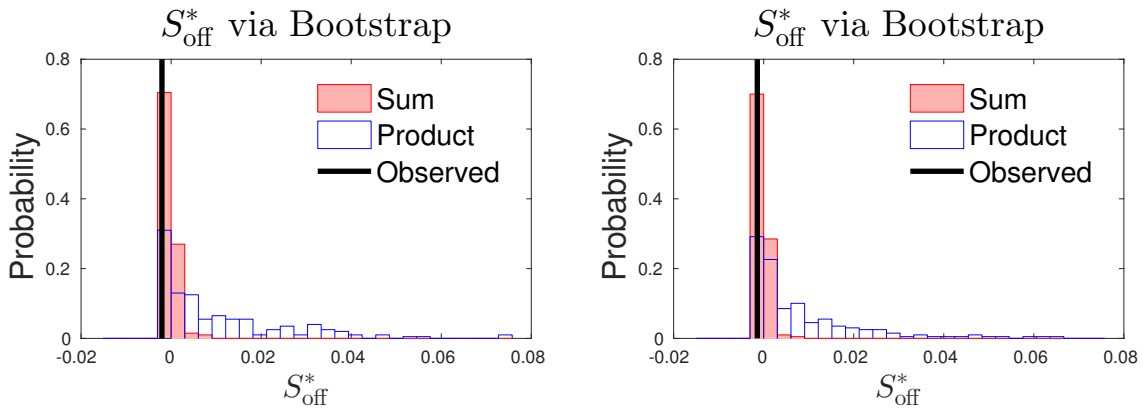

Figure 4.8: The histogram of S^*_{off} calculated from 200 generated random samples using estimates of A and B from the Kronecker sum and product models. The blue bar indicates the observed statistic from the moth torque data X, and right bars and blue bars are from sum and product based samples. The two histograms are from moth J and moth L, respectively.

4.5.2 Estimating the Trace Parameter

In this subsection, we propose a method to estimate tr(A) by comparing the estimates $\widehat{A} = \widehat{\Theta}^{-1}$ and $\widehat{B} = \widehat{\Omega}^{-1}$ obtained by using the different values of $\widehat{\text{tr}}(A)$. We note that $\text{tr}(A)/m$ reflects the overall contribution of A to the data variance, and hence is very important in understanding the two-way dependence of X. While the decomposition $\Sigma = A \oplus B$ is not unique, in general if A and B are not diagonal $\Sigma = (A + cI_m) \oplus (B - cI_n) = \widetilde{A} \oplus \widetilde{B}$ will not has the property that $\widetilde{A}^{-1}$ and $\widetilde{B}^{-1}$ are sparse. Therefore, the decomposition is identifiable via its sparsity.

To estimate Θ and Ω for the Kronecker sum covariance model using the procedure described in Section 4.2, the trace of A or B must be fixed. We propose to estimate tr(A) by comparing the Kronecker sum estimates $\widehat{A}$ and $\widehat{B}$ obtained by using the different values of $\widehat{\text{tr}}(A)$ in the estimation. Without loss of generality, we normalize the data X such that $\|X\|_F^2 = 2mn$, which in turn we can assume that $\tau_A + \tau_B = 2$.

Now let $\widehat{A}(C)$ and $\widehat{B}(C)$ be the estimates using $\tau_A = C \in (0, 2)$. Let

$$C^* = \text{argmin}_{C\in(0,2)} \|\text{vec}(X)\text{vec}(X)^T - \widehat{A}(C) \oplus \widehat{B}(C)\|_F, \tag{4.14}$$

which minimizes the Frobenius distance between the rank-one sample covariance matrix $\text{vec}(X)\text{vec}(X)^T$ and the Kronecker sum covariance estimate $\widehat{A}(C) \oplus \widehat{B}(C)$. We estimate A and B using $\tau_A = \widehat{\tau}_A := C^*$. The intuitive idea of the minimizer C^* defined in (4.14) is that $\widehat{A}(C)$ and $\widehat{B}(C)$ estimate $\widetilde{A}$ and $\widetilde{B}$ such that $\widetilde{A} \oplus \widetilde{B} = \Sigma$ and $\tau_{\widetilde{A}} = C$. Since $\text{vec}(X)\text{vec}(X)^T$ is an unbiased estimate of Σ, the minimizing problem (4.14) may find the C such that $\widehat{A}(C)$ and $\widehat{B}(C)$ more accurately estimate corresponding pair A and B satisfying $A \oplus B = \Sigma$ and $\text{tr}(A) = C$ than that of other $C \in (0, 2)$. We note that the estimates $\widehat{A}(C)$ and $\widehat{B}(C)$ may be closer to the corresponding A and B when A^{-1} and B^{-1} are sparse. This roughly implies that $\widehat{A}(C^*)$ and $\widehat{B}(C^*)$ may estimate A and B satisfying $A \oplus B = \Sigma$ and A^{-1} and B^{-1} are the most sparsest pair among other pairs A and B.

Figure 4.9 displays simulation results for accuracy of $\widehat{\tau}_A$ for the three cases when $\tau_A \in \{0.4, 1, 1.8\}$. It is seen that $\widehat{\tau}_A$ tends not to concentrate around the true τ_A, although the mode seems to be close to the truth (with 0.1 distance).

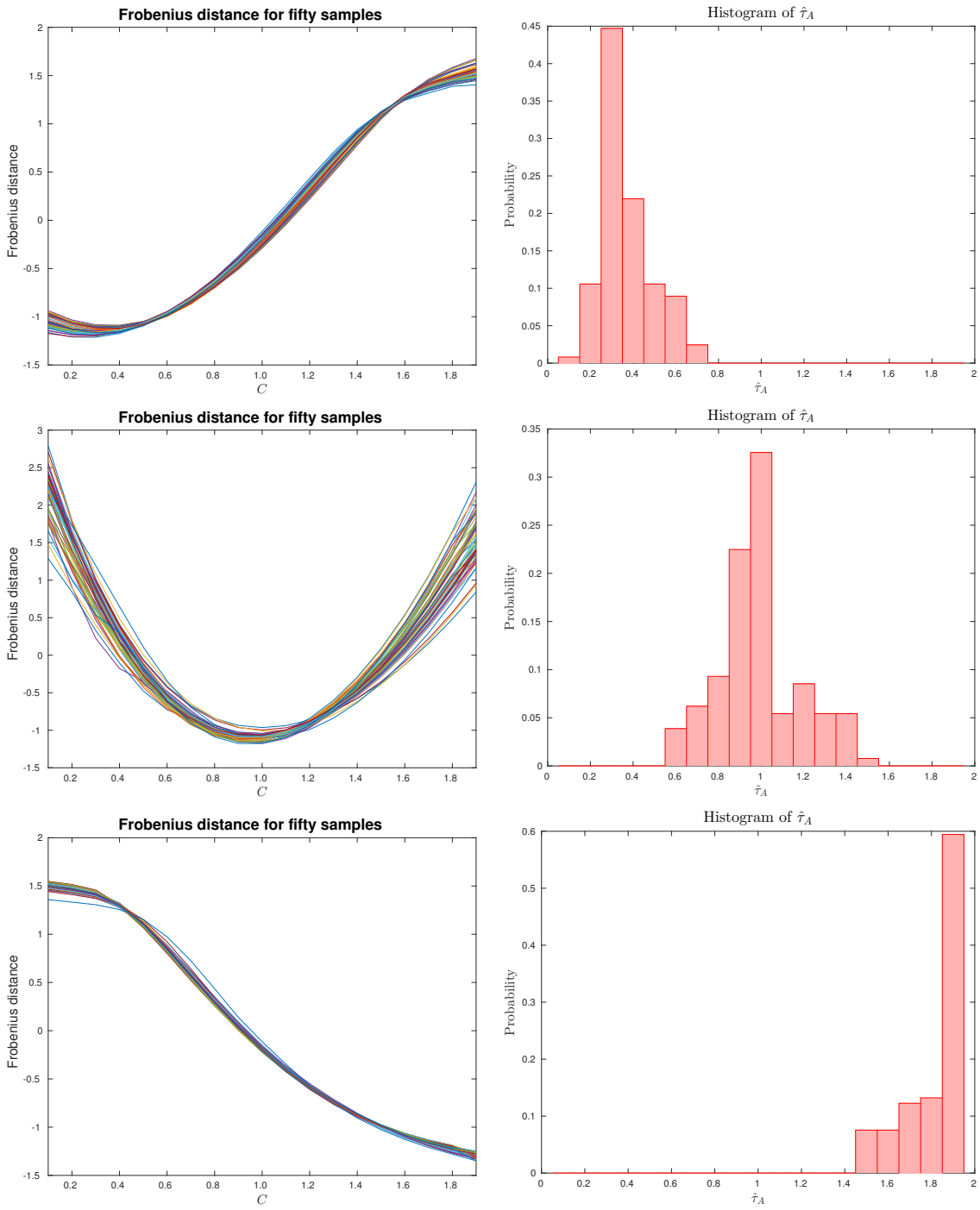

Figure 4.9: Normalized error paths for 50 samples (left) and the histogram of $\mathrm{tr}(A)/m$ (right) for 200 samples when $\tau_A = 0.4$ (top), $\tau_A = 1$ (middle) and $\tau_A = 1.8$ (below). The dimension $(m, n) = (400, 100)$. For the top plots, we use $A = 0.4A^*$, where A^* follows AR(1) model, and $B = 1.6B^*$, where B^* follows random model. For the middle and below plots, we use $A = A^*$ and $B = B^*$, and $A = 1.8A^*$ and $B = 0.2B^*$, respectively.

Figure 4.10 displays the average of error paths for 200 samples when A and B are diagonal matrices. We also consider the situation in which A and B are diagonal, in which case $\mathrm{tr}(A)$ is not identifiable, even through sparsity. The proposed procedure for estimating $\mathrm{tr}(A)$ is seen to converge to a degenerate Kronecker sum with only one term, which is sufficient for capturing the structure of Σ.

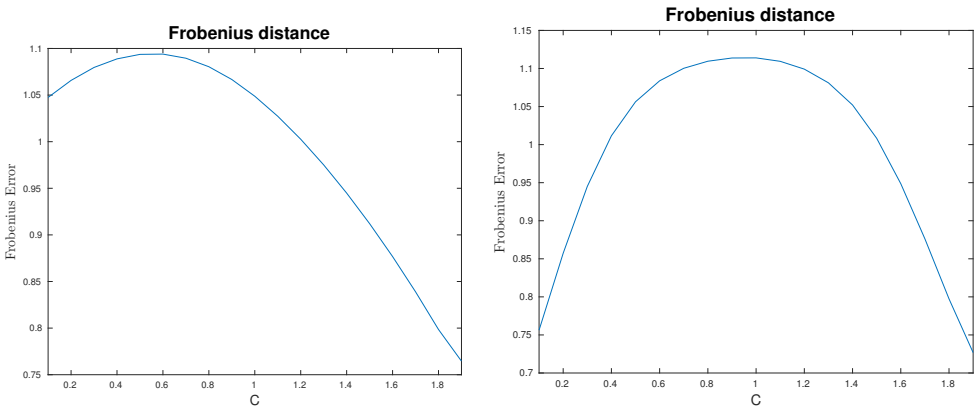

Figure 4.10: Average of normalized error paths for 200 samples when A and B are diagonal matrices. The left and right plots are when $\tau_A = 1$ and $\tau_A = 1.5$, respectively.

We apply this method to estimate the trace parameter for moth data. Figure 4.11 shows that $\widehat{\tau}_A = 1$ and $\widehat{\tau}_A = 1.3$ are the optimal values for the Moth J and L, respectively.

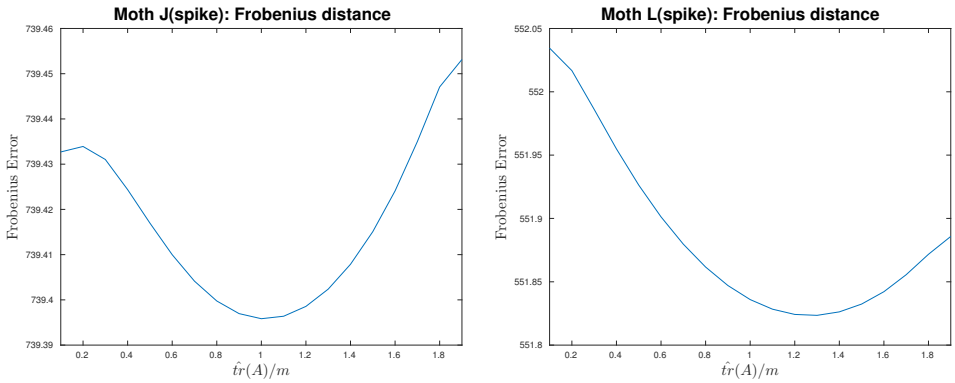

Figure 4.11: Moth J and L(spike): Frobenius distance of the Kronekcer sum covariance estimate obtained by using $\widehat{\tau}_A = d$, relative to rank-one sample covariance matrix.

4.5.3 Graphical Structures

In this subsection, we estimate the graphical structures among the time points and among the wingstrokes, respectively. Based on the analysis of the statistic S_{off} in Subsection 4.5.1, the Kronecker sum covariance model seems to be more appropriate than the Kronecker product model for explaining the two-way dependency in the movement data X.

Figure 4.12 displays the graphical structure of the estimates $\widehat{\Theta}$ for the optimal trace of A from the Kronecker sum model. Here $\widehat{A}$ is the estimated covariance matrix between time points. It is observed that all the connections in the graphical structure of $\widehat{\Theta}$ concentrate near the diagonal, which is expected as it is a temporal autocovariance matrix.

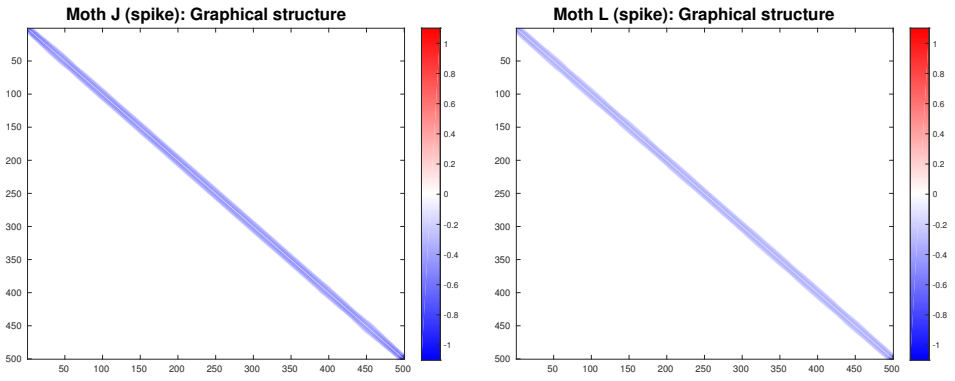

Figure 4.12: The graphical structure of $\widehat{\Theta}$ from Kronecker sum model using nodewise regression method

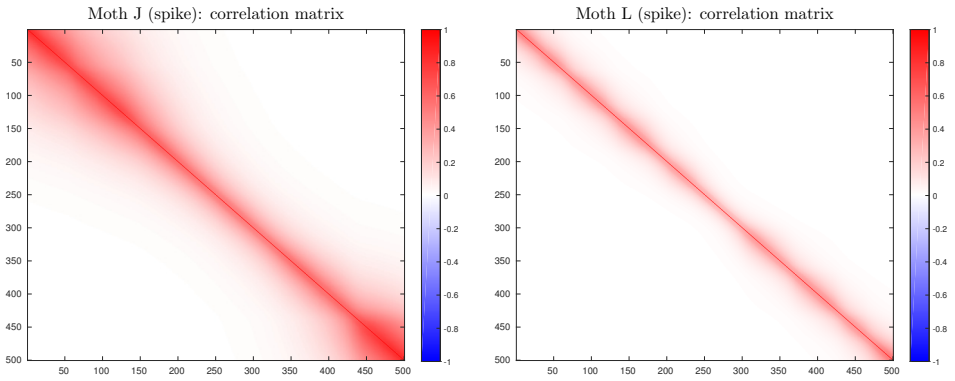

Figure 4.13: The estimated correlation matrix calculated from the Kronecker sum estimate $\widehat{A} = \widehat{\Theta}^{-1}$. The nodewise method is used for the estimates. The left and the right plots correspond to moth J and L, respectively.

Figure 4.13 displays the estimated correlation matrix plots for moths J and L. It is observed that the correlation structure have the clear periodic patterns, where the signals are decreasing as their time distances are increasing.

Figure 4.14 shows the component plots of the estimated inverse covariance matrix $\widehat{\Theta}$. The diagonal components dominate the off-diagonal components, and the off-diagonal components corresponding to the times whose distance are greater than 7 have the value zero. We can see that the diagonal and the off-diagonal components are stable except a first and last few components. This is consistent with the rows of X following an approximately stationary process with short memory.

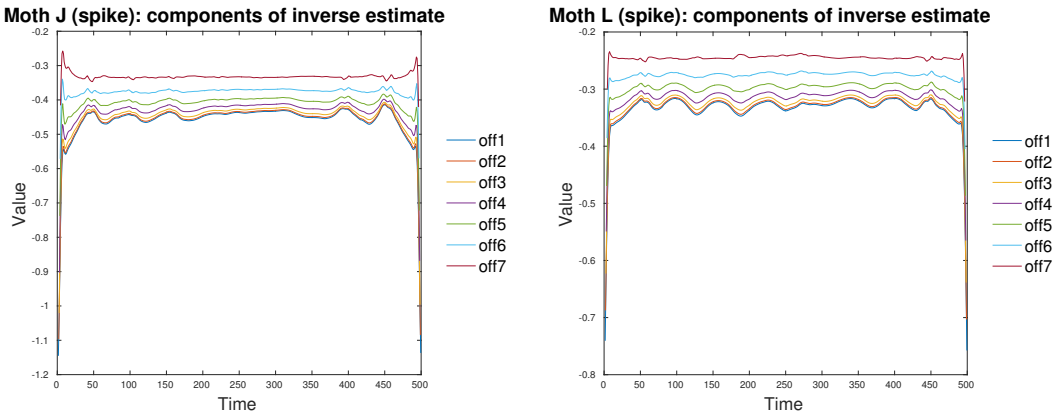

Figure 4.14: The diagonal components and the off-diagonal components of $\widehat{\Theta}$ from Kronecker sum model. Here off(k) records the off-diagonal components having the form $(\widehat{\Theta})_{i,i-k}$ for $i = 1, \cdots, 500$.

Figure 4.15 displays the estimated graphical structure of $\widehat{\Omega}$ for the Kronecker sum model. The number of wingstrokes varies by moth having values between 300 and 1020. For the moth J, let cluster 1 be the cluster that consists of five wingstrokes (wingstroke with numbers 235,411,436,453, and 503), cluster 2 be the cluster that consists of seven wingstrokes (wingstroke with numbers 215,253,257,260,339,420, and 486), and cluster 3 be the cluster that consists of eighteen wingstrokes (wingstroke with numbers 17,29,36,53,55,70,75,88,132, etc). Figure 4.16 shows torque ensembles for the three clusters. The average of pairwise correlations within the cluster are greater than those between clusters, which implies that wingstrokes within cluster have more similar pattern than that of wingstrokes from other clusters. Moreover, the average of the torque mean within clusters are obviously different; the average and standard deviation of the torque mean within clusters are (0.3288, 0.2481), (0.1318, 0.2394) and (-0.3678, 0.2680), respectively. This implies that the cluster1 and cluster3 may consist of wingstrokes with right and left turn, respectively. Cluster 2, however, seems to consist of wingstrokes with straight flights.

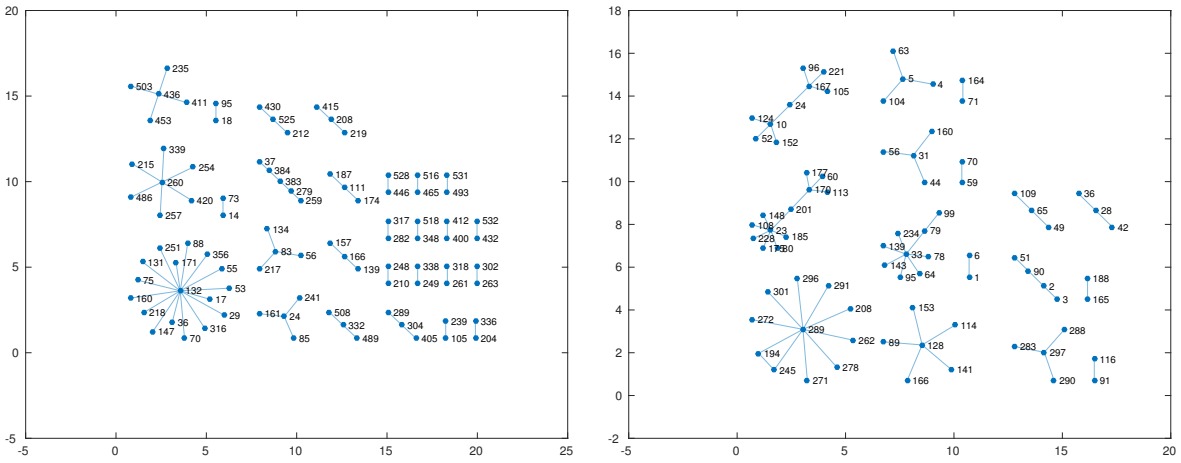

Figure 4.15: The estimated graphical structure of Ω from Kronecker sum model. The left and the right plots for Moth J and L, respectively.

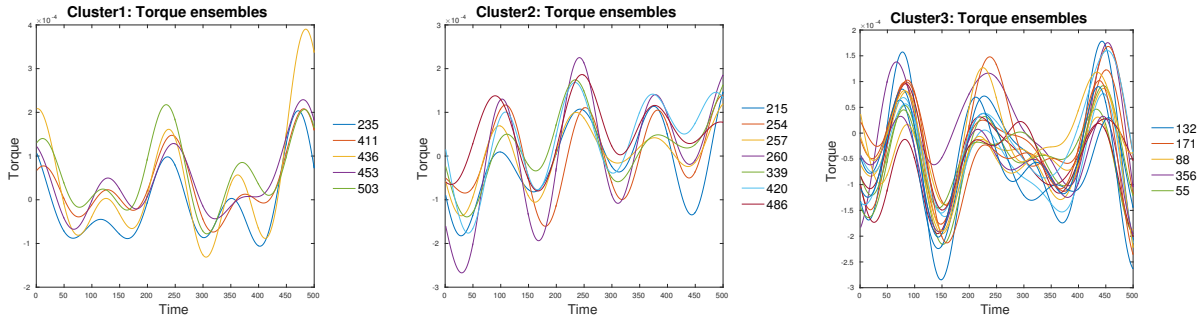

Figure 4.16: Torque ensembles for three clusters of Moth J. The average of pairwise correlations within the cluster 1, cluster 2 and cluster 3 are 0.6876, 0.7120, and 0.7322, respectively. The average of pairwise correlations between clusters 1&2, 1&3, and 2&3 are 0.3572, 0.1144, and 0.2125, respectively. The average and standard deviation of the torque mean within clusters are (0.3288, 0.2481), (0.1318, 0.2394) and (-0.3678, 0.2680), respectively.

Figure 4.17 displays a few wingstrokes plots whose estimated correlations are very high or low. For example, as can be seen in the left plot (moth J), the three wingstrokes are displayed, where the estimated correlation between w_{498} and w_{379} calculated from $\widehat{B}$ is 0.82. These two wingstroke show very similar patterns over the 500 time points. The estimated correlation between wingstrokes w_{498} and w_{432} is

-0.35, and they show quite different pattern over the times.

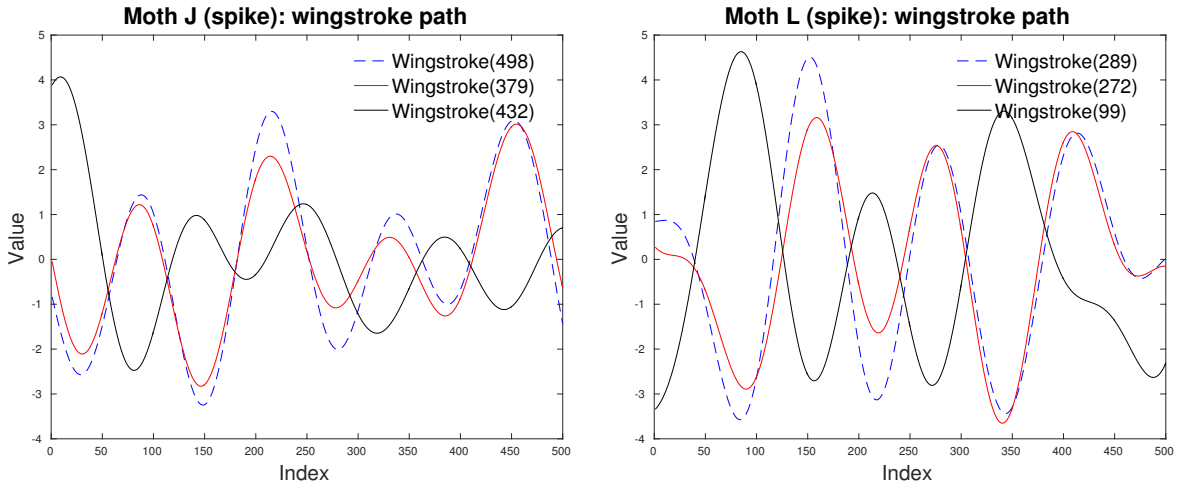

Figure 4.17: The left plot shows three wingstroke paths for moth J. For the wingstrokes w_{498} and w_{379}, the correlation obtained from the data and the correlation calculated from $\widehat{B}$ are 0.972 and 0.82, respectively. Those values for the wingstrokes w_{498} and w_{432} are -0.47 and -0.35, respectively. The right plot includes the three wingstrokes paths for moth L. For the wingstrokes w_{289} and w_{272}, the correlation from the data and the correlation calculated from $\widehat{B}$ are 0.946 and 0.85, respectively. The values for the wingstrokes w_{289} and w_{99} are -0.79 and -0.71, respectively.

4.5.4 Mean-Variance Analysis

In this subsection, we analyze the relationship between the mean torque of a wingstroke and its turning direction.

Figure 4.18 shows the scatter plots for the mean differences of wingstrokes and the corresponding entries in $\widehat{\Omega}$ and $\widehat{B}$. The red line shows the mean value of the entries. Wingstroke pairs are grouped by subsets of size 100 based on their mean differences, and the mean is calculated for each group. It is shown that the entries of $\widehat{\Omega}$ tend to have values near zero as the mean difference of the corresponding wingstrokes grows. This demonstrates that wingstrokes with different turning directions seem to be conditional independent, as the mean torque is known to be related to the turing

direction of wingstroke (Sponberg et al., 2015). Based on the estimated covariance, for moth J, the wingstrokes from the same turning direction seems to be more positively correlated than those of different turning direction. For moth L, as turning directions of wingstrokes are different, they tends to be negatively correlated. This may be explained by a phase shift of the wingstrokes for moth L, as pointed out by Sponberg et al. (2015).

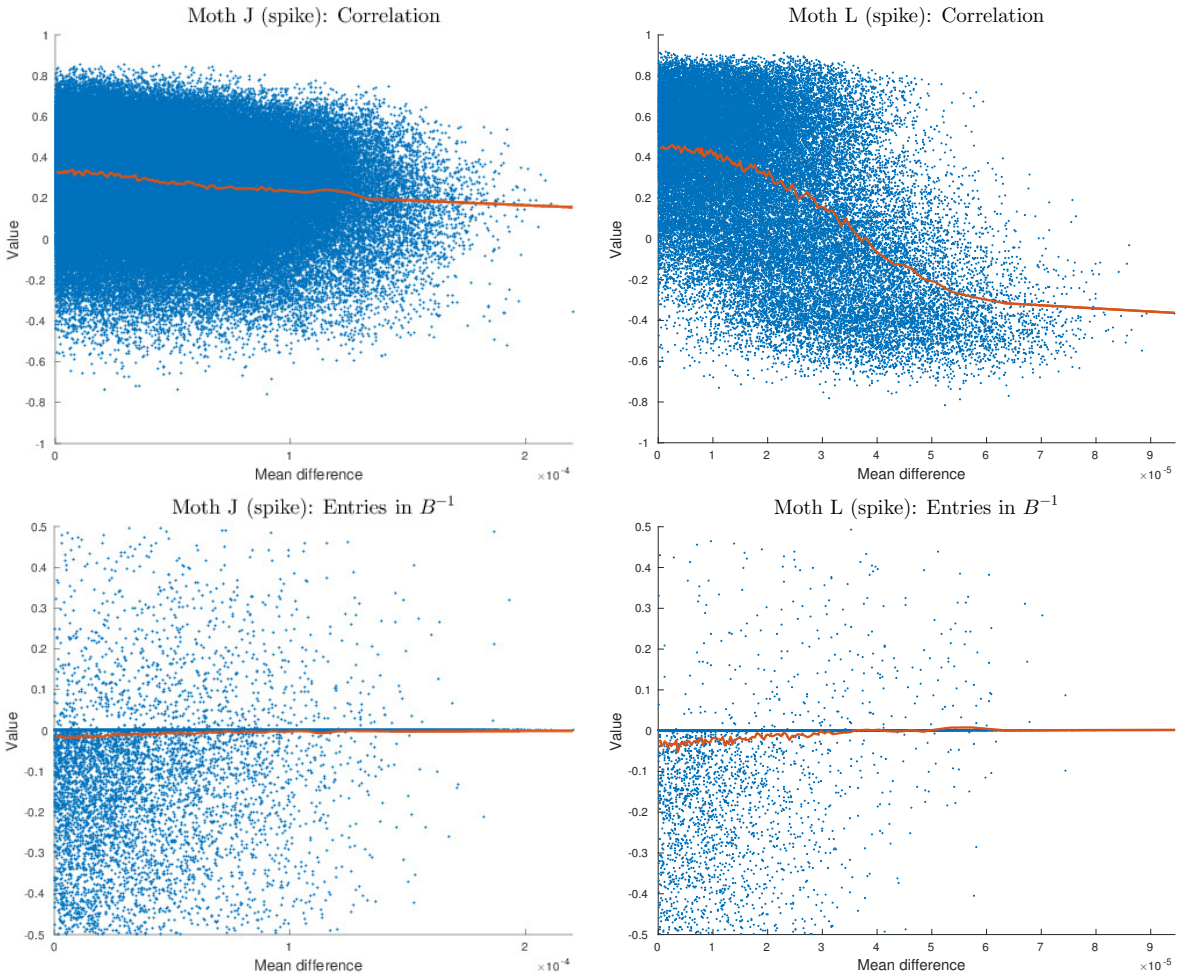

Figure 4.18: Scatter plots for the mean torque differences of wingstrokes and the corresponding entries in $\widehat{B}$ (top) and $\widehat{\Omega}$ (below).

4.5.5 Regression Analysis

In this subsection, we analyze a relationship of neural firing and torque by using linear regression and errors-in-variables regression. Note that a baseline assumption in our errors-in-variables regression is that the movement data (design matrix in the regression) follows a Kronecker sum covariance model. We will apply regular regression and errors-in-variables regression to the hawkmoth neural encoding data and compare the results.

Note that the motor signals data $Y \in \mathbb{R}^{n\times 2}$ consists of the two spike timing variables (t_L, t_R), where t_L and t_R is the left and right DLM spike times of wingstroke, respectively, and $n = N_i$ and $m = 500$. The number of wingstrokes N_i depends on the moth, and has the value between 300 and 1020 for each moth i. Sponberg et al. (2015) performed partial least squares (PLS) using these two data sets X and Y to find the relevant features encoded in the movement data X that are related to variations of the motor signal data Y. Let Y_1 and Y_2 be the first and the second columns of Y.

We perform the regression of $Y_c = Y_1 - Y_2$ on the torque measurement data X. We first extract the effect of mean torque from Y_c by using the simple regression of Y_c on the mean of X, and let $\widetilde{Y}_c$ be the residual vector after the regression. Then we apply both errors-in-variables regression (Rudelson and Zhou, 2015) and regular penalized regression with ℓ_2/ℓ_1 penalties such as ridge regression (Hoerl and Kennard, 1970) and Scaled Lasso (Sun and Zhang, 2012) to compute each of the coefficients

as follows:

$$\begin{aligned}\widetilde{Y}_c &= X\beta + \epsilon, \quad X \text{ and } \widetilde{\mathrm{Y}}_{\mathrm{c}} \text{ are observable,} \\ \widetilde{Y}_c &= X_0\beta + \epsilon,\ X = X_0 + W, \quad X \text{ and } \widetilde{\mathrm{Y}}_{\mathrm{c}} \text{ are observable.}\end{aligned}$$

Scaled Lasso (Sun and Zhang, 2012) involves an alternating minimization algorithm for the penalized joint loss function:

$$[\widehat{\beta}, \widehat{\sigma}] = \underset{\beta \in \mathbb{R}^m, \sigma > 0}{\arg\min} \ \frac{|\widetilde{Y}_c - X\beta|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 |\beta|_1.$$

Ridge regression (Hoerl and Kennard, 1970) estimates β such that

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^m}{\arg\min} \ \frac{|\widetilde{Y}_c - X\beta|_2^2}{2n\sigma} + \lambda \|\beta\|_2^2.$$

Let $\widehat{\beta}_R$ and $\widehat{\beta}$ be the coefficients obtained from the regular regression and the errors-in-variables regression, respectively. For the estimate $\widehat{\beta}_R$, we use the linear regression with Lasso penalty. We first calculate the explanatory power of $X = X_0 + W$ for $\widetilde{Y}_c$ using the proportion of explained variance as follows: $R_X^2 = \mathrm{corr}\left(X\widehat{\beta}_R, \widetilde{Y}_c\right)^2$. Similarly, we estimate the explanatory power of X_0 for $\widetilde{Y}_c$ using the estimates $\widehat{\beta}$ and the $\widehat{A}$: $R_{X_0}^2 = \mathrm{cor}\left(X_0\beta^o, \widetilde{Y}_c\right)^2$, where we substitute $X_0^T X_0/n$ with $\widehat{A}$.

Table 4.3: Explanatory power (R-squared)

Moth (Phase data)	$R_X^2(R)$	$R_X^2(S)$	$R_{X_0}^2$	Moth (Spike data)	$R_X^2(R)$	$R_X^2(S)$	$R_{X_0}^2$
J	0.10	0.11	0.21	J	0.21	0.31	0.38
K	0.09	0.14	0.24	K	0.53	0.53	0.57
L	0.34	0.32	0.40	L	0.51	0.51	0.59
M	0.06	0.01	0.24	M	0.06	0.01	0.30
N	0.26	0.27	0.37	N	0.41	0.43	0.55
P	0.38	0.41	0.50	P	0.55	0.52	0.69
Q	0.35	0.30	0.46	Q	0.24	0.53	0.67

Table 4.3 shows the explanatory power of X and the estimated lower bound of the explanatory power of X_0 in terms of R-squared. The $R_X^2(R)$, $R_X^2(S)$, and $R_{X_0}^2$

are the R^2 estimate of using Ridge regression, Scaled Lasso and errors-in-variables regression, respectively. It is shown that the explanatory power of the X is mostly less than that of X_0, which implies that movement features relevant to the difference of motor signal vectors Y_c are mostly encoded in X_0. This shows that the Kronecker sum decomposition provides two random parts, where one part is mainly related to the difference of the motor signals. In particular, there seems to be significant improvements in R^2 for moth M. To see if this improvement is due to the upward bias in the estimation of the R^2, we consider the errors-in-variables model with $\beta = 0$ and the same covariance matrices estimated for moth M. Figure 4.19 displays the estimate of R-squared, which shows that the improvement on R^2 for moth M is not due to upward bias.

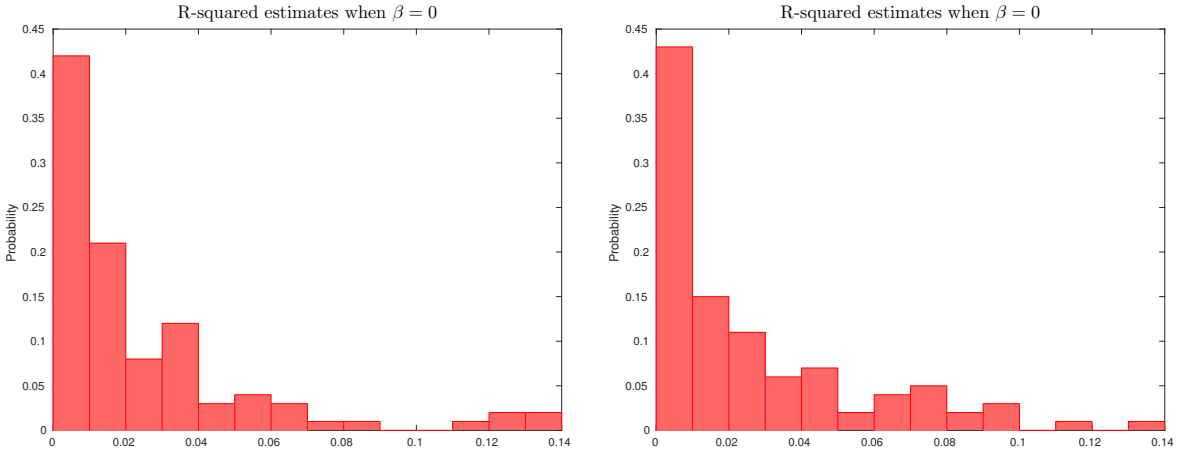

Figure 4.19: Histogram of R-squared estimates when $\beta = 0$ and X follows Kronecker sum covariance model, where the estimated covariance matrices for moth M (phase) and moth M (spike) are used for left and right plots, respectively.

Figures 4.20 and 4.21 display the estimated regression coefficients for the regular regression and the errors-in-variable regression (EIV). It is shown that the errors-in-variables regression coefficients have more stable and stronger components than

regular regression and have positive components over the all time points. To the extent that the coefficient patterns found when using measurement error regression are interpreted as being constant, this is consistent with the simple 1-dimensional neuro-encoding hypothesis, and suggests that any evidence for complementary neural-encoding pathways may be an artifact resulting from not fully accounting for the presence of measurement errors. Figures 4.22 and 4.23 display more sparse estimates whose sparsity is less than $n/\log m$) for errors-in-variables regression (EIV). Although coefficients are not stable, they have positive components over the all time points.

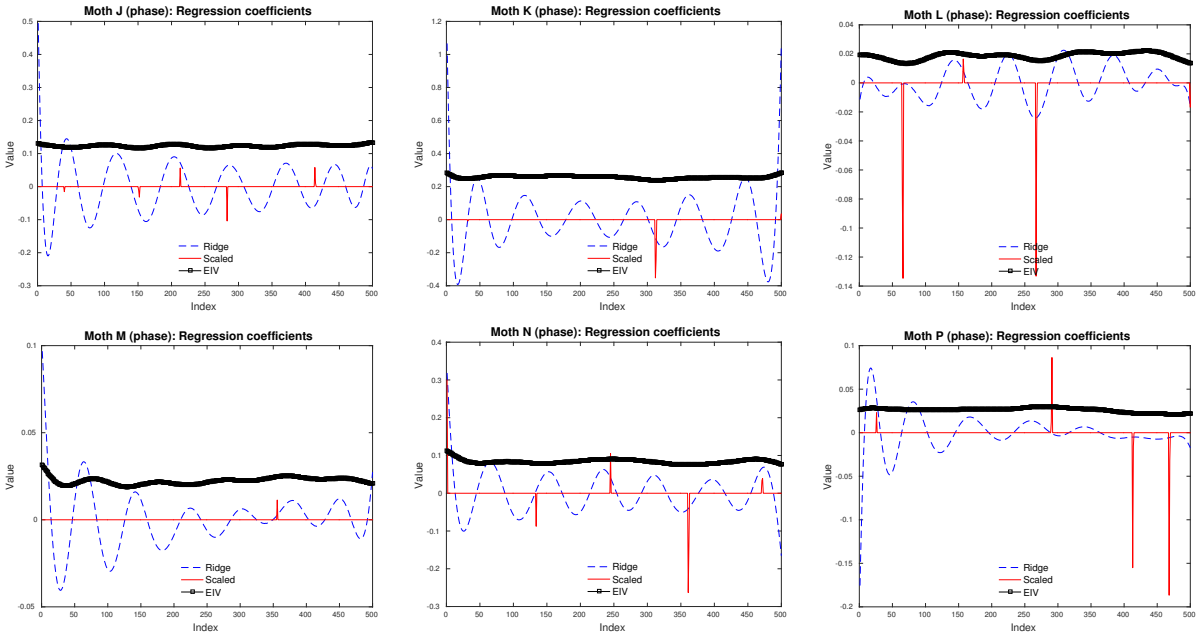

Figure 4.20: The estimated regression coefficients from the regular regression (ridge regression and Scaled Lasso) and the errors-in-variables regression (EIV) for phase data set.

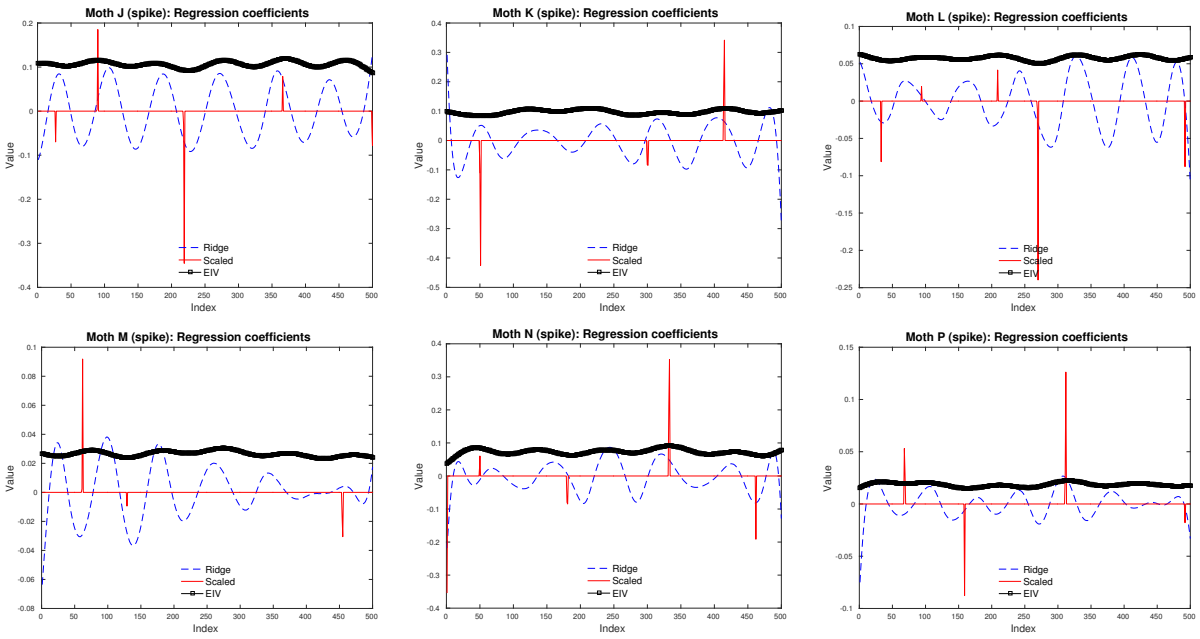

Figure 4.21: The estimated regression coefficients from the regular regression (ridge regression and Scaled Lasso) and the errors-in-variables regression (EIV) for spike data set.

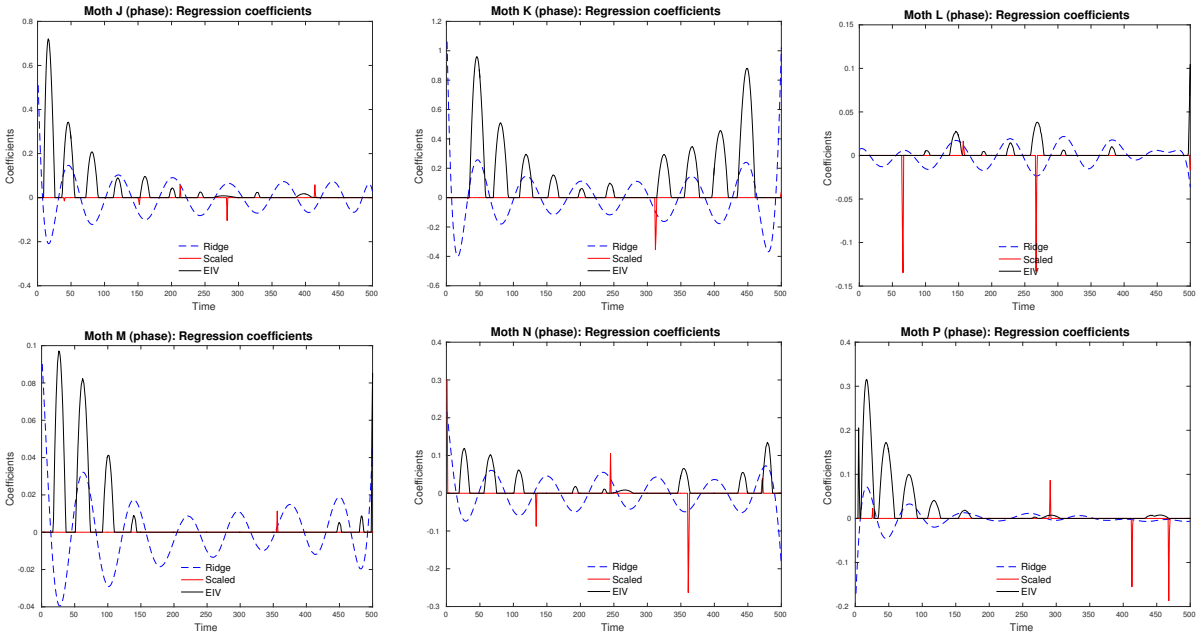

Figure 4.22: The estimated regression coefficients from the regular regression (ridge regression and Scaled Lasso) and the errors-in-variables regression (EIV) for phase data set.

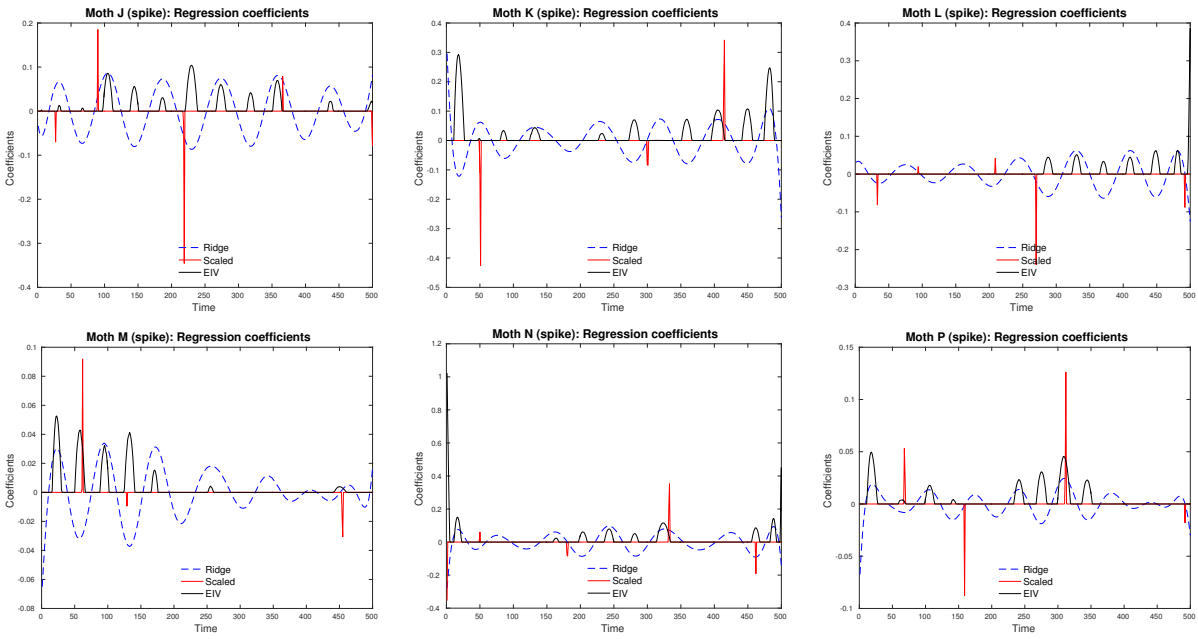

Figure 4.23: The estimated regression coefficients from the regular regression (ridge regression and Scaled Lasso) and the errors-in-variables regression (EIV) for spike data set.

CHAPTER 5

Future Work

5.1 Hypothesis Testing for Multiple Quantiles

In future research, I will extend my study of multiple quantile regression estimation in two ways. First, I will study the model selection consistency without imposing a strong beta-min condition on the quantile coefficients. Our current analysis relies on this beta-min condition, which is restrictive and non-checkable. Second, I will study hypothesis testing for multiple quantiles with high-dimensional covariates. We test the impact of a given covariate on the conditional quantile functions across a quantile interval Δ as follows: $H_0 : \beta_j(\tau) = 0_q$ for all $\tau \in \Delta$, $\quad H_1 : \beta_j(\tau) \neq 0_q$ for some $\tau \in \Delta$. The interval Δ may be chosen as $[0.45, 0.55]$ if it is desirable to test variables that impact the center of the conditional distributions, or $[0.8, 09]$ if one is interested in the upper tails. To approximate the quantile functions over quantile levels in the interval, we may use B-splines and consider composite quantile regression to estimate the B-spline coefficients. By using the estimated coefficients, the score-type test statistic (Gutenbrunner and Jurečková, 1992; Gutenbrunner et al., 1993) is constructed based on the asymptotic normality of the statistic under the null hypothesis H_0. In numerical examples, we have observed that the proposed

score-type test for multiple quantiles provides higher power than other tests based on a single quantile level. This indicates that this multiple quantile test is beneficial under certain cases.

5.2 Theory and Methods for EIV Regression

In ongoing research, we are developing theory and methods for errors-in-variables regression and graphical model selection using a single copy of the data as well as replicated data, extending the work in Chapters 3 and 4. Replicated data are available in many modern application areas. For example, in neuroscience studies, data often involve multiple trials and subjects. With the replicated data, the strong assumption, tr(A) is known, is not necessary.

Bibliography

Agarwal, A., Negahban, S., and Wainwright, M. (2012). Fast Global Concvergence of Gradient Methods for High-Dimensional Statistical Recovery. *Annals of Statistics*, 40(5), 2452–2482.

Allen, G. and Tibshirani, R. (2010). Transposable Regularized Covariance Models with an Application to Missing Data Imputation. *Annals of Applied Statistics*, 4(2), 764-790.

Bang, S. and Jhun, M. (2012). Simultaneous Estimation and Factor Selection in Quantile Regression via Adaptive Sup-norm Regularization. *Computational Statistics and Data Analysis*, 56, 813–826.

Belloni, A. and Chernozhukov, V. (2011). l_1-penalized Quantile Regression in High-dimensional Sparse Models. *Annals of Statistics*, 39, 82–130.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform Post-selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems. *Biometrika*, 102, 77–94.

Belloni, A., Rosenbaum, M., and Tsybakov, A. (2014). Linear and Conic Programming Estimators in High-Dimensional Errors-in-variables Models. *arXiv:1408.0241.*

Bickel, P. and Levina, E. (2008). Covariance Regularization by Thresholding. *Annals of Statistics*, 36, 2577–2604.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37, 1705–1732.

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Non-crossing Quantile Regression Curve Estimation. *Biometrika*, 97, 825–838.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3, 1-122.

Bruno B. (1967). The Distribution of a Quadratic Form of Normal Random Variables. *Ann. Math. Statist.*, 6, 1700-1704.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High–dimensional Data: Methods, Theory and Applications. *New York: Springer*.

Cai, T., Liu, W., and Luo, X. (2011). A Constrained $\ell 1$ Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106:594–607.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical Estimation When p is Much Larger Than n. *Annals of Statistics*, 35, 2313–2351.

Carroll, R. J., Gallo, P. P., and Gleser, L. J. (1985). Comparison of Least Squares and Errors-in-variables Regression with Special Reference to Randomized Analysis of Covariance. *Journal of American Statistical Association*, 80:929–932.

Carroll, R., Gail, M. H., and Lubin, J. H. (1993). Case-control Studies with Errors in Predictors. *Journal of American Statistical Association*, 88:177–191.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models (Second Edition). *Chapman & Hall*, 2006.

Carroll, R. and Wand, M. (1991). Semiparametric Estimation in Logistic Measurement Error Models. *J. R. Statist. Soc. B*, 53:573-585.

Chen, S., Donoho, D., and Saunders,M. (1998). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific and Statistical Computing*, 20, 33–61.

Constantinou, P. and Kokoszka, P. (2015). Testing Separability of Space Time Functional Processes. *arxiv: http://arxiv.org/pdf/1509.07017v1.pdf*.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, 89, 1314–1328.

Cressie, N. and Wikle, C. (2011). Statistics for Spatio-Temporal Data. *Wiley*, 2011.

Datta, A. and Zou, H. (2015). CoCoLasso for High-dimensional Error-in-variables Regression. *http://arxiv.org/pdf/1510.07123v1.pdf*.

Dawid, A. P. (1981). Some Matrix-variate Distribution Theory: Notational Considerations and a Bayesian Application. *Biometrika*, 68:265–274.

Dette, H. and Volgushev, S. (2008). Non-crossing Non-parametric Estimates of Quantile Curves. *Journal of the Royal Statistical Society, Series B*, 70, 609–627.

Dutilleul, P. (1999). The MLE algorithm for the Matrix Normal Distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123,

Efron, B. (2009). Are a Set of Microarrays Independent of Each Other? *Ann. App. Statist.*, 3, 922–942.

Emre, Telatar. (1999). Capacity of Multi-antenna Gaussian Channels. *European Trans. on Telecommunications*, 10(6), 585–595.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive Robust Variable Selection. *Annals of Statistics*, 42, 324–351.

Fan, J. and Truong, Y. K. (1993). Nonparametric Regression with Errors in Variables. *The Annals of Statistics*, 21, 1900–1925.

Fan, J., Xue, L., and Zou, H. (2014). Strong Oracle Optimality of Folded Concave Penalized Estimation. *Annals of Statistics*, 42, 819–849.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9, 432–441.

Fuller, W. A. (1987). Measurement Error Models. *John Wiley and Sons*, 1987

Gao, Y. and Sun, D. (2010). A Majorized Penalty Approach for Calibrating Rank Constrained Correlation Matrix Problems. *National University of Singapore*, 2010.

Greenewald, K. and Hero, O. (2015a). Kronecker PCA Based Robust SAR STAP. *arXiv preprint arXiv:1501.07481*.

Greenewald, K. and Hero, O. (2015b). Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation. *Signal Processing, IEEE Transactions*, 23, 6368-6378.

Gupta, A. and Varga, T. (1992). Characterization of Matrix Variate Normal Distributions. *Journal of Multivariate Analysis*, 41:80–88.

Gutenbrunner, C. and J.Jurečková (1992). Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics. *Annals of Statistics*, 20, 305–330.

Gutenbrunner, C., J.Jurečková, R.Koenker, and S.Portnoy (1993). Test of Linear Hypotheses Based on regression Rank Scores. *Journal of Nonparametric Statistics*, 2, 307–333.

Hall, P. and Ma, Y. (2007). Semiparametric Estimators of Functional Measurement Error Models with Unknown Error. *Journal of the Royal Statistical Society B*, 69:429–446.

He, X. (1997). Quantile Curves without Crossing. *The American Statistician*, 51, 186–192.

Henrion, D. and Malick, J. (2012). Projection Methods in Conic Optimization. *Handbook on Semidefinite, Conic and Polynomial Optimization*, 2012:565–600.

Higham, N. (2002). Computing the Nearest Symmetric Correlation Matrix-a Problem from Finance. *IMA J. Numer. Anal.*, 22:329–343.

Hoff, P. D. (2011). Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data. *Bayesian Analysis*, 6:179–196.

Hoff, P. D. (2011). Hierarchical Multilinear Models for Multiway Data. *Computational Statistics & Data Analysis*, 55:530–543.

Hoerl, A.E. and R.W. Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1): 55–67.

Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive Lasso for Sparse High-dimensional Regression Models. *Statist. Sinica*, 18, 1603–1618.

Hwang, J.T. (1986). Multiplicative Errors-in-variables Models with Applications to Recent Data Released by the U.S. Department of Energy. *Journal of American Statistical Association*, 81:680–688.

Iturria, S. J., Carroll, R. J., and Firth, D. (1999). Polynomial Regression and Estimating Functions in the Presence of Multiplicative Measurement Error. *Journal of the Royal Statistical Society, Series B, Methodological*, 61, 547-561.

Jang, W. and Wang, H. (2015). A Semiparametric Bayesian Approach for Joint-quantile Regression with Clustered Data. *Journal of Computational and Graphical statistics*, 84, 99–115.

Jiang, L., Wang, H., and Bondell, H. (2013). Interquantile Shrinkage in Regression Models. *Journal of Computational and Graphical statistics*, 69, 208–219.

Kalaitzis, A., Lafferty,J., Lawrence,N., and Zhou, S. (2013). The Bigraphical Lasso. *Proceedings of The 30th International Conference on Machine Learning ICML-13*, 1229-1237.

Kim, Y., Choi, H., and Oh, H. S. (2008). Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association*, 103, 1665–1673.

Knight, K. (1998). Limiting Distributions for l_1 Regression Estimators under General Conditions. *Annals of Statistics*, 26, 755–770.

Koenker, R. and Basset, G. (1978). Regression Quantiles. *Econometrica*, 46, 33–50.

Koenker, R. (2005). *Quantile Regression. Cambridge Univ. Press, Cambridge.*

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37, 4254.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes. Ergebnisse der Mathematik und ihrer Grenzgebiete, 23, Berlin: Springer.*

Leeb, H. and Pötscher B.M. (2003). *The Finite-sample Distribution of Post-model-selection Estimators and Uniform versus Nonuniform Approximations. Economic Theory*, 19, 100–142.

Leeb, H. and Pötscher B.M. (2005). *Model Selection and Inference: Facts and Fiction. Economic Theory*, 21, 21–59.

Leng, C. and Tang, C.Y. (2012). *Sparse Matrix Graphical Models*. *Journal of American Statistical Association*, 107, 1187–1200.

Li, Y. and Zhu, J. (2008). l_1-norm Quantile Regression. *Journal of Computational and Graphical statistics*, 17, 163–185.

Liang, H. and Härdle, W. and Carroll, R. J. (1999). Estimation in a Semiparametric Partially Linear Errors-in-variables Model. *Ann. Statist.*, 27, 519-1535.

Liang, H. and Li, R. (2009). Variable Selection for Partially Linear Models with Measurement Errors. *Journal of the American Statistical Association*, 104, 234–248.

Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition). *John Wiley and Sons*.

Loh, P. and Wainwright, M. (2012). High-dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity. *Annals of Statistics*, 40(3), 1637–1664.

Lu, N. and Zimmerman, D. (2005). The Likelihood Ratio Test for a Separable Covariance Matrix. *Statist. Probab. Lett.*, 73, 449–457.

Malick, J. (2004). A Dual Approach to Semidefinite Least-squares Problems. *SIAM Journal on Matrix Analysis and Applications*, 26, 272–284.

Marzetta, T. L. and Hochwald, B. M. (1999). Capacity of a Mobile Multiple-antenna Communication Link in Rayleigh Flat Fading. *IEEE Trans. Info. Theory*, 45(1):139–157.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional Graphs and Variable Selection with the Lasso. *Annals of Statistics*, 34, 1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type Recovery of Sparse Representations for High Dimensional Data. *Annals of Statistics*, 37, 246–270.

Peng, L., Xu, J., and Kutner, N. (2014). Shrinkage Estimation of Varying Covariate Effects Based on Quantile Regression. *Statistics and computing*, 24, 853–869.

Rigollet, P. and Tsybakov, A. (2012). Estimation of covariance matrices under sparsity constraints. *arXiv preprint arXiv:1205.1210*.

Rosenbaum, M., and Tsybakov, A. (2010). Sparse Recovery under Matrix Uncertainty. *The Annals of Statistics*, 38(5):2620–2651.

Rosenbaum, M., and Tsybakov, A. (2013). Improved Matrix Uncertainty Selector. *IMS Collections*, 9:276–290.

Rothman, A.J., Bickel, P, J., and Levina, E. (2008). Sparse Permutation Invariant Covariance Estimation. *JElectronic Journal of Statistics*, 2, 494–515.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright Inequality and Sub-gaussian Concentration. *Electron. Commun. Probab*, 82, 1-9.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59, 3434–3447.

Rudelson, M. and Zhou, S. (2015). High Dimensional Errors-in-variables Models with Dependent Measurements. *arxiv: http://arxiv.org/pdf/1502.02355v1.pdf*.

Rudin, L., Osher, S., and Fatemi , E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*, 60, 259-268 .

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of Gene Expression in the Mammalian Eye and its Relevance to Eye Disease. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 14429–14434.

Schwertman, N. and Allen, D. (1979). Smoothing an Indefinite Variance-Covariance Matrix. *Journal of Statistical Computation and Simulation*, 9, 183–194.

Smith, R., Kolenikov, S., and Cox, L. (2003). Spatiotemporal Modeling of $PM_{2.5}$ Data with Missing Values. *Journal of Geophysical Research*, 108.

Søresen and Frigenssi, A., and Thoresen, M. (2014). Measurement Error in Lasso: Impact and Likelihood Bias Correction. *Statistical Sinica Preprint*.

Sponberg, S, Daniel, TL., and Fairhall, AL. (2015). Dual Dimensionality Reduction Reveals Independent Encoding of Motor Features in a Muscle Synergy for Insect Flight Control. *PLoS Comput Biol*, 11(4).

Städler, N., Stekhoven, D.J., and Bühlmann, P. (2014). Pattern Alternating Maximization Algorithm for Missing Data in High-dimensional Problems. *Journal of Machine Learning Research*, 15, 1903-1928.

Stefanski, L. A. (1985). The Effects of Measurement Error on Parameter Estimation. *Biometrika*, 73, 583–592.

Stefanski, L. A. (1990). Rates of Convergence of Some Estimators in a Class of Deconvolution Problems. *Statistics and Probability Letters*, 9, 229–235.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, 90, 1247–1256.

Sun, T. and Zhang, C.-H. (2012). Scaled Sparse Linear Regression. *Biometrika*, 102, 246–270.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, Series B*, 67, 91–108.

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance Estimation in High Dimensions Via Kronecker Product Expansions. *Annals of Statistics*, 61, 5347–5360.

Tsiligkaridis, T., Hero, A., and Zhou, S. (2013). On Convergence of Kronecker Graphical Lasso Algorithms. *IEEE Transactions on Signal Processing*, 61, 1743–1755.

van de Geer, S. and Bühlmann, P. (2009). On the Conditions Used to Prove Oracle Results for the Lasso. *Electronic Journal of Statistics*, 3, 1360–1392.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The Adaptive and the Thresholded Lasso for Potentially Misspecified Models (and a Lower Bound for the Lasso). *Electronic Journal of Statistics*, 5, 688–749.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. New York: Springer.*

Van Loan, C.F. and Pitsianis, N. (1993). Approximation with Kronecker Products. *Linear Algebra for Large Scale and Real Time Applications*, 293-314.

Volgushev, S., Wagener, J., and Dette, H. (2014). Censored Quantile Regression Processes under Dependence and Penalization. *Electronic Journal of Statistics*, 8, 2405–2447.

Wainwright, M. (2009). Sharp Thresholds for High-dimensional and Noisy Sparsity Recovery using ℓ_1–constrained Quadratic Programming. *IEEE Trans. Inform. Theory*, 55, 2183–2202.

Wang, L., Wu, Y., and Li, R. (2012). Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension. *Journal of the American Statistical Association*, 107, 214–222.

Wang, L. (2013). l_1 penalized LAD Estimator for High–dimensional Linear Regression. *Journal of Multivariate Analysis*, 120, 135–151.

Werner, K., Jansson, M., and Stoica, P. (2008). On Estimation of Covariance Matrices with Kronecker Product Structure. *IEEE Transactions on Signal Processing*, 56(2):478–491.

Wu, Y. and Liu, Y. (2009). Variable Selection in Quantile Regression. *Statist. Sinica*, 19, 801–817.

Yin, J. and Li, H. (2012). Model Selection and Estimation in the Matrix Normal Graphical Model.. *Journal of Multivariate Analysis*, 107:119–140.

Yuan, M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, 11, 2261–2286.

Yu, K., Lafferty, J., Zhu, S., and Gong, Y. (2009). Large-scale Collaborative Prediction Using a Nonparametric Random Effects Model. *Proceedings of the 26th International Conference on Machine Learning*.

Zhang, C. H. and Huang, J. (2008). The Sparsity and Bias of the LASSO Selection in High–dimensional Linear Regression. *Annals of Statistics*, 36, 1567–1594.

Zhang, Y. and Schneider, J. (2010). Learning Multiple Tasks with a Sparse Matrix–Normal Penalty. *In Advances in Neural Information Processing Systems 23 (NIPS 2010)*.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541–2567.

Zheng, Q., Gallagher, C., and Kulasekera, K. B. (2013). Adaptive Penalized Quantile Regression for High-dimensional Data. *Journal of Computational and Graphical Statistics*, 143, 1029–1038.

Zheng Q., Peng, L., and He, X. (2015). Globally Adaptive Quantile Regression with Ultra-high Dimensional Data. *Annals of Statistics*, 43, 2225–2258.

Zhou, S. (2009). Restricted Eigenvalue Conditions on Subgaussian Random Matrices. *Technical Report: http://arxiv.org/abs/0912.4045*.

Zhou, S., Rutimann, P., Xu, M., and Buhlmann, P. (2011). High-dimensional Covariance Estimation Based on Gaussian Graphical Models. *Journal of Machine Learning Research*, 12, 2975–3026.

Zhou, S. (2011). Thresholded Lasso for high dimensional variable selection. *Technical Report: http://arxiv.org/abs/1002.1583*.

Zhou, S. (2014). GEMINI: Graph Estimation with Matrix Variate Normal Instances. *Annals of Statistics*, 42(2), 532–562.

Zou, H. and Yuan, M. (2008a). Regularized Simultaneous Model Selection in Multiple Quantiles Regression. *Journal of Computational and Graphical Statistics*, 52, 5296–5304.

Zou, H. and Yuan, M. (2008b). Composite Quantile Regression and The Oracle Model Selection Theory. *Annals of Statistics*, 36, 1108–1126.