IMPUTATION OF AUTOMATIC CONTROL ALGORITHMS AND ESTIMATION IN HIGH-DIMENSIONAL LINEAR REGRESSION

BY FEI YE

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Professor Cun-Hui Zhang

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2010

ABSTRACT OF THE DISSERTATION

Imputation of Automatic Control Algorithms and Estimation in High-Dimensional Linear Regression

by FEI YE

Dissertation Director: Professor Cun-Hui Zhang

This thesis contains two parts. In the first part, we study a semiparametric imputation method to simulate a time series of blood glucose level under certain closed-loop control algorithm of a diabetic patient equipped with a continuous glucose monitor and an insulin pump, from the "frozen" measurements under self-adjusted open-loop control. The Star One data set provided by Medtronic Inc illustrates the feasibility of a simple PID algorithm, as an example of automatic control algorithms, in controlling blood glucose levels from the perspective of reducing the A1c level and controlling hypoglycemia risk.

In the second part, we consider ℓ_1-penalized selection of variables and estimation of regression coefficients in a high-dimensional linear model. Under an ℓ_0 sparsity condition on the regression coefficients, we sharpen an upper bound of Candes and Tao [4] for the ℓ_2 loss of the Dantzig selector and extend it to the ℓ_q loss and the Lasso. By allowing $q = \infty$, our bound implies the variable selection

consistency of threshold Dantzig selectors. For the estimation of regression coefficients in ℓ_r balls, we provide minimax lower bounds for the ℓ_q risk and the tail quantiles of the ℓ_q loss as well as sufficient conditions on the design matrix and penalty level for the Dantzig and Lasso estimators to attain these minimax rates.

Acknowledgements

Although I typed out this dissertation, many others have greatly contributed to the preparation of my thesis. I am deeply grateful to my advisor, Professor Cun-Hui Zhang for his invaluable guidance, extensive support and constant encouragement. Not only does he gives me hands-on guidance on the topics of this thesis, but he also teaches me how to learn effectively and think deeply. His devotion to mathematical and statistical sciences will always guide and inspire me.

I would like to thank graduate director John Kolassa, the other faculty and staff at Department of Statistics and Biostatistics of Rutgers University for supporting me in various ways and at various occasions during my study. My thanks also go to my colleagues Jerry Cheng, Wenhua Jiang, Jue Wang and Jane Zhang, who provide many suggestion and help.

Last but not least, I want to thank the other members of my dissertation committee, Professor Wanpracha Chaovalitwongse, Professor Lawrence Shepp, and Professor Tong Zhang for their helpful comments on the manuscript. I am also grateful to Medtronic Inc for providing the Star One data set.

Dedication

To My Parents and My Wife

Table of Contents

Abstract ii

Acknowledgements iv

Dedication v

List of Tables viii

List of Figures ix

1. Introduction 1

1.1. Imputation of Blood Glucose Levels and HbA1c's for Automatic Control Algorithms 1

1.2. Selection and Estimation with the ℓ_1 Penalty 4

2. Imputation of Blood Glucose Levels and HbA1c's for Automatic Control Algorithms 9

2.1. Introduction 9

2.2. Research Design and Methods 12

2.2.1. Semiparametric Imputation Method 12

2.2.2. PID Algorithm 16

2.2.3. Mathematical Formula for A1c 17

2.3. Data Screening and Interpolation of Missing Values 18

2.4. Numerical Experiments 19

2.5. Hypoglycemia Measures 24

2.6. Discussion 27

3. Selection and Estimation with the ℓ_1 Penalty 29
3.1. Introduction 29
3.2. Error bounds and variable selection under ℓ_0 sparsity 33
3.3. Estimation with ℓ^q loss in ℓ^r balls 38
3.3.1. Lower bounds for the estimation risk and loss 38
3.3.2. Upper bounds for the Dantzig and Lasso estimation risk . 39
3.3.3. Upper bounds for the Lasso estimation loss 42
3.3.4. Oracle inequalities 43
3.4. Proofs 45
3.4.1. Proofs of lower bounds 46
3.4.2. Proofs of oracle inequalities 47
3.4.3. Proofs of upper bounds 51
3.5. Discussion 56
References 58
Vita 60

List of Tables

2.1. Scores of data quality, 3 types of A1c's of 21 subjects: PID-imputed A1c's, estimated A1c's based on "frozen" CGM's and laboratory measured A1c's. 20

2.2. Some descriptive statistics of "frozen" CGM's and PID-imputed CGM's for 21 subjects: sample mean and sample standard deviation. 23

2.3. Some hypoglycemia measures of "frozen" CGM's and PID-imputed CGM's for 21 subjects: number of hypoglycemia eposides, median of durations (in hours), median of average areas underneath the threshold of glucose level 60 mg/dL. 27

List of Figures

2.1. The insulin concentration after 1 unit subcutaneous delivery of insulin. 14

2.2. The illustrative example of cubic spline interpolation of missing values of glucose measurements of patient 14 on 10/22/2005. . . 21

2.3. Left: the life-style component of patient 14; right: average blood glucose levels of patient 14 from 8/8/2005 to 11/3/2005. 21

2.4. Left: PID-imputed A1c's and measured A1c's; right: PID-imputed A1c's and estimated A1c's. 22

2.5. Left: glucose levels, insulin deliveries and insulin concentration of patient 14 on 8/9/2005; right: imputed glucose levels, calculated insulin deliveries and insulin concentration of patient 14 on 8/9/2005. 24

2.6. Left: glucose levels, insulin deliveries and insulin concentration of patient 79 on 1/4/2006; right: imputed glucose levels, calculated insulin deliveries and insulin concentration of patient 79 on 1/4/2006. 26

Chapter 1
Introduction

1.1 Imputation of Blood Glucose Levels and HbA1c's for Automatic Control Algorithms

This thesis concerns two problems. The first problem considers imputation of blood glucose levels and HbA1c's for automatic control algorithms from the "frozen" measurements under self-adjusted open-loop control. The Star One data set of 137 diabetic patients contains continuous blood glucose measurements $G(t)$, subcutaneous insulin deliveries $u(t)$, and A1c measurements for one year. Each patient is equipped with a continuous glucose monitor and an insulin pump. The monitor reads the blood glucose level $G(t)$ every 5 minutes. The pump delivers the insulin according to a prespecified basal rate and self-controlled bolus rate for meals. The patients receive no other source of insulin except through their insulin pumps. Our goal is to simulate a time series of blood glucose level had certain closed-loop control algorithm been applied to the diabetic patient.

Inspired by the novel nonparametric statistical imputation method of Mastrotaro *et al* [15], we study a semiparametric imputation method which utilizes an insulin absorption model and an insulin action model as parametric components and a nonparametric life-style component. We describe the insulin absorption using a two-compartment plasma insulin concentration model of Hovorka *et al* [14] and the insulin action using a revised minimal insulin action model of Steil *et al* [21]. For simplicity we denote the insulin absorption system as $H(\cdot)$ such

that $I(t) = H(u(s), s \leq t)$, where $I(t)$ is the plasma insulin concentration at time t and $\{u(s), s \leq t\}$ are discrete subcutaneous insulin deliveries up to time t. The revised minimal insulin action model can be written as

$$G(t + \Delta_t) - G(t) = \beta_0(t) + \beta_1 G(t) + \beta_2 G(t) \sum_{s \leq t} I(s) e^{-w(t-s)}, \tag{1.1}$$

where $\beta_0(t)$ is the nonparametric life-style component of the patient, β_1, β_2 are unknown parameters and w is an unknown weight of the exponential moving average of $\{I(s), s \leq t\}$.

Consider an individual patient in the Star One data set. Suppose measurements are taken in days $d = 1, \ldots, m$ at time points t_k, $k = 1, \ldots, n$ with fixed $\Delta_t = t_{k+1} - t_k$ (5 minutes). Let $u_d(t_k)$ be the actual amounts of subcutaneous insulin delivery and $G_d(t_k)$ be the actual blood glucose measurements. From $u_d(t_k)$, the plasma insulin concentration $I_d(t_k)$ are computed using the insulin absorption model $H(\cdot)$. We describe the semiparametric imputation method as follows:

Step 1: Estimate β_1, β_2, w and the life-style component $\beta_0(\cdot)$ using (1.1) and all time/day points. In this step, $\beta_0(\cdot)$ is approximated by a cubic spline function $\widehat{\beta}_0(\cdot)$.

Step 2: Compute the adjustment $\epsilon_d(t_k)$ of life-style component for each day using the estimated parameters $\{\widehat{\beta}_1, \widehat{\beta}_2, \widehat{w}\}$, $\widehat{\beta}_0(\cdot)$ from Step 1 and the actual data for that day,

$$\epsilon_d(t_k) = G_d(t_{k+1}) - G_d(t_k) - \widehat{\beta}_0(t_k) - \widehat{\beta}_1 G(t_k) - \widehat{\beta}_2 G(t_k) \sum_{j \leq k} I_d(t_j) e^{-\widehat{w}(t_k - t_j)}.$$

Step 3: Dynamically update $\widetilde{u}_d(t_k)$, $\widetilde{I}_d(t_k)$ and impute the blood glucose level $\widetilde{G}_d(t_{k+1})$ for a given closed-loop control algorithm $A(\cdot)$,

$$\begin{aligned}
&\widetilde{u}_d(t_k) = A(\widetilde{G}_d(t_j), j \leq k) \\
&\widetilde{I}_d(t_k) = H(\widetilde{u}_d(t_j), j \leq k) \\
&\widetilde{G}_d(t_{k+1}) = \widetilde{G}_d(t_k) + \widehat{\beta}_0(t_k) + \widehat{\beta}_1 \widetilde{G}_d(t_k) + \widehat{\beta}_2 \widetilde{G}_d(t_k) \sum_{j \leq k} \widetilde{I}_d(t_j) e^{-\widehat{w}(t_k - t_j)} + \epsilon_d(t_k).
\end{aligned}$$

As an example we impute the glucose levels under a simple proportional-integral-derivative (PID) algorithm. PID algorithm could be used to control the blood glucose level toward a desired target by providing suitable dose of insulin delivery. In PID, the insulin dose is calculated using continuous glucose measurements $G(t)$ as a linear combination of three components: proportional (P), integral (I) and derivative (D). PID algorithm can be written as $PID(t) = Pcontrol(t) + Icontrol(t) + Dcontrol(t)$ where $Pcontrol(t) = \kappa_p(G(t) - G_{target})$, $Icontrol(t) = Icontrol(t-1) + \kappa_i(G(t) - G_{target})$ and $Dcontrol(t) = \kappa_d G'(t)$ with $G'(t)$ being the derivative of $G(t)$.

HbA1c, or A1c, is the percentage of glycated hemoglobin molecules in a red blood cell. The higher blood glucose level, the more glucose molecules would join hemoglobin, resulting in a higher A1c level. Hemoglobin A1c test is the standard measure of diabetic hyperglycemia. Palerm, Shepp, Cabrera and Zhang [20] proposed a mathematical formula to estimate A1c based on continuous blood glucose levels. The A1c's under PID are imputed by applying the mathematical formula for A1c to the imputed time series of blood glucose level. We compare PID-imputed A1c's with laboratory measured A1c's to evaluate the feasibility of the simple PID algorithm.

Besides hyperglycemia, the other risk of diabetic disorder is hypoglycemia, which could lead to ketoacidotic coma. Hypoglycemia happens when a diabetic patient injects too much insulin and has no matching carbohydrate intake to spend the insulin. An efficient closed-loop control algorithm should keep low both risks of hyperglycemia and hypoglycemia. To this end, we define a number of hypoglycemia measures and compare each of them between the PID algorithm and the open-loop control. Although the simple PID algorithm has good performance in controlling hyperglycemia, we find it difficult for the PID to further lower hypoglycemia risk, compared with the open-loop control. The reason is that the Star One data set has the survival bias. No diabetic patients die in the clinical

trial due to ultra low glucose levels. When they feel uncomfortable or hungry, they would simply eat some additional food to keep up their blood glucose levels. This component is not imputed.

1.2 Selection and Estimation with the ℓ_1 Penalty

The second problem considers estimation in high-dimensional linear regression. We assume a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is a design matrix and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of stochastic errors. Ordinary least squares estimator $\arg\min_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2$ is neither parsimonious nor unique thus it is difficult to interpret the selected model, in the case of $p > n$. In many statistical and engineering applications, the number p of design variables (features, covariates) can be larger or even of large order than the sample size n. For example, in signal processing (sparse recovery, compressed sensing), a sparse p-dimensional signal $\boldsymbol{\beta}$ is encoded through a linear transformation $\boldsymbol{X}$ to an n-dimensional vector $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with stochastic or deterministic noises $\boldsymbol{\varepsilon}$, stored or transmitted in the n-dimensional form, and then recovered using some appropriate algorithm. In linear regression, a popular approach for model selection and parameter estimation is to impose a suitable penalty on the empirical loss. In a high-dimensional linear regression, p is often large enough thus it is computationally infeasible to select the model minimizing ℓ_0-penalized empirical loss. Tibshirani [22] proposed the Lasso, an ℓ_1-penalized estimator

$$\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1 \right\}$$

for the regression coefficients. For the simplest orthonormal design $\boldsymbol{X}'\boldsymbol{X}/n = \boldsymbol{I}$, the Lasso estimator is the soft-threshold least squares estimator. In the signal

processing literature, the Lasso is known as basis pursuit [6]. The Lasso has the interpretation as boosting [11, 12] and is computationally feasible for high-dimensional data [18, 19, 9]. Recently Candes and Tao [4] proposed another ℓ_1-penalized method called the Dantzig selector,

$$\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda) = \arg\min_{\boldsymbol{b}} \Big\{ \|\boldsymbol{b}\|_1 : |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n| \le \lambda, \forall j \Big\}.$$

It is a simple convex program that can be recast as a convenient linear program. The Dantzig and Lasso estimators share some theoretical properties from the perspective of variable selection and parameter estimation. We consider variable selection and estimation of a sparse vector of regression coefficients in a linear model simultaneously for the Dantzig and Lasso estimators.

An estimator is variable selection consistent if the set of nonzero estimated coefficients matches that of the "true" nonzero regression coefficients with large probability. One of the important properties of the Lasso is that it can be used for variable selection. Donoho *et al* [8] showed that penalties with a discontinuous derivative at the origin possess the variable selection feature of shrinking some coefficients exactly to zero. Meinshausen and Buhlmann [16], Tropp [23], Zhao and Yu [29] and Wainwright [25] proved that the Lasso is variable selection consistent under a strong irrepresentable condition on the Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$ and some other regularity conditions. Zhang and Huang [27] proved the consistency of the Lasso in the order of the dimension and bias of the selected model under a regularity condition on the eigenvalues of sub-Gram matrices. More recently, Candes and Plan [5] proved the selection consistency of the Lasso under random permutation and sign-change of regression coefficients and a mild condition on the maximum absolute correlation among design vectors. Zhang [28] studied the selection consistency of the Lasso through its ℓ_∞ loss. Although it was pointed out in [10, 2] that the Dantzig and Lasso estimators are quite similar, it is still unclear in the existing literature if the Dantzig selector possesses selection consistency properties parallel to those mentioned above for the Lasso.

Another focus of recent research of the ℓ_1-penalized least squares estimators has been on the estimation loss for the regression coefficients. Candes and Tao [4] derived an elegant probabilistic upper bound of the ℓ_2 loss for the Dantzig selector under a condition on the number of nonzero coefficients and a uniform uncertainty principle (UUP) on the Gram matrix. The similar analysis of upper bounds for the ℓ_q loss $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q$ of the Lasso estimator has been studied by Bunea, Tsybakov and Wegkamp [3] and van de Geer [24] for $q = 1$, Zhang and Huang [27] for $q \in [1, 2]$, Meinshausen and Yu [17] for $q = 2$, Bickel, Ritov and Tsybakov [2] for $q \in [1, 2]$, and Zhang [28] for $q \geq 1$. Under different sets of regularity conditions on the Gram matrix and sparsity of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, these results provide upper bounds of the form $\|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) - \boldsymbol{\beta}\|_q \leq O(k^{1/q}\lambda)$, where k is of the order of an intrinsic dimensionality of the sparse estimation problem. For $N(0, \sigma^2)$ errors and standardized designs with $\|\boldsymbol{x}_j\| = \sqrt{n}$, the required penalty levels λ in these studies on the Lasso are all greater by a constant factor than the universal penalty level $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$ in the inequality of Candes and Tao [4]. Different sets of regularity conditions lead to different forms of constant factors in the upper bounds so that the existing upper bounds are typically not directly comparable mathematically.

For the estimation of a target vector $\boldsymbol{\beta}$, we derive oracle inequalities which bounds the ℓ_q loss of the Dantzig and Lasso estimators in terms of the oracle error bound $\rho_k(\boldsymbol{\beta}) = \sum_{j\notin J_k} |\beta_j|$ with $J_k = \arg\max_{|S|=k} \sum_{j\in S} |\beta_j|$ and error measures on $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. For $k \geq 0$ and all $1 \leq q \leq \infty$, in the event $\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty \leq \lambda$

$$\|\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda) - \boldsymbol{\beta}\|_q \leq (1+\tau^q)^{1/q} \max_{A,\boldsymbol{u}} \min_{\boldsymbol{v}} \max_{B} \frac{2\lambda G_{A,\boldsymbol{u},\boldsymbol{v}} + 2k^{1/q-1}\rho_k(\boldsymbol{\beta})}{(1 - F_{A,B,\boldsymbol{u},\boldsymbol{v}})_+}$$

where $|A| = k + \ell$ with $A \supset J_k$ and $1 \leq \ell \leq p - k$, $\boldsymbol{u} \in \mathbb{R}^A$ with $\|\boldsymbol{u}\|_q = 1$, $0 \neq \boldsymbol{v} \in \mathbb{R}^A$, $|B| = \ell$ with $A \cap B = \emptyset$, $F_{A,B,\boldsymbol{u},\boldsymbol{v}} = \ell^{-1}\|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+$, $G_{A,\boldsymbol{u},\boldsymbol{v}} = \|\boldsymbol{v}\|_1/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+$ and $\tau = (k/\ell)^{1-1/q}$. Moreover, for $0 < \alpha < 1$

$$\|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda/\alpha) - \boldsymbol{\beta}\|_q \leq (1+\xi^q\tau^q)^{1/q} \max_{A,\boldsymbol{u}} \min_{\boldsymbol{v}} \max_{B} \frac{\xi'\lambda G_{A,\boldsymbol{u},\boldsymbol{v}} + 2k^{1/q-1}\rho_k(\boldsymbol{\beta})}{(1 - \xi F_{A,B,\boldsymbol{u},\boldsymbol{v}})_+}$$

with $\xi = (1+\alpha)/(1-\alpha)$ and $\xi' = 1 + 1/\alpha$. For the Lasso with $\boldsymbol{v} = f_s(\boldsymbol{u})$ and $f_s(x) = \mathrm{sgn}(x)|x|^s$, $G_{A,\boldsymbol{u},\boldsymbol{v}} = \frac{\|\boldsymbol{v}_{J_k}\|_1}{(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+} + \left\{\frac{(1+1/\alpha)^{-s}\lambda^{-s}\rho_k(\boldsymbol{\beta})\|\boldsymbol{v}\|_{q/s}}{(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+}\right\}^{1/(s+1)}$ is also allowed. Simple upper bounds are obtained for the ℓ_q loss of the Lasso and Dantzig estimator under an assumption on the ℓ_0 sparsity of a target vector of regression coefficients that $\|\boldsymbol{\beta}\|_0 = k$. By explicitly allowing $q = \infty$, our bound implies the variable selection consistency of threshold Dantzig selectors. Our error bounds sharpen and unify a number of existing approaches and extend the inequality of Candes and Tao [4] from $q = 2$ to $1 \le q \le \infty$ and the Lasso. By taking $\boldsymbol{v} = \boldsymbol{u}$, applying Hölder inequality, and taking the worst scenarios in both the numerator and denominator of the constant factors, our bounds improves upon a result of Bickel, Ritov and Tsybakov [2] and the inequality of Candes and Tao [4]. The choices of $\boldsymbol{v} = f_{q-1}(\boldsymbol{u})$ and $\boldsymbol{v} = \boldsymbol{\Sigma}_A^{-1} f_{q-1}(\boldsymbol{u})$ provides upper bounds of $\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_q$, slightly improving upon a result of Zhang [28]. Although the error bounds for the Dantzig and Lasso estimators are of the same format, the Lasso bounds require a larger penalty level λ/α ($\alpha < 1$) compared with the panelty level λ for the Dantzig bounds.

We also prove that both the Dantzig and Lasso estimators achieve rate minimaxity in the ℓ_q risk $E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q$ and loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q$ for the estimation of regression coefficients in ℓ_r balls, by providing lower bounds for general estimators and matching upper bounds. We extend the results of Donoho and Johnstone [7] from orthonormal design to linear regression design. For $0 < r \le q$, we prove that the minimax ℓ_q risk $\inf_{\boldsymbol{\delta}} \sup_{\|\boldsymbol{\beta}\|_r \le R} E_{\boldsymbol{\beta}}\|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_q^q$ and loss in ℓ_r balls are bounded from below by $R^r \lambda_{mm}^{q-r}$, where $\lambda_{mm} = \sigma\left\{\frac{2}{n}\log\left(\frac{\sigma^r p}{n^{r/2}R^r}\right)\right\}^{1/2}$ is a certain minimax penalty level and R is the radius of the ℓ_r ball. When λ_{mm} is of the same order as $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$, we prove that the Dantzig and Lasso estimators both attain the rate of the minimax risk and loss for $0 < r \le 1 \le q$. In simulation studies and applications, a penalty level $\lambda < \lambda_{univ}$ is often empirically the best choice. For $\lambda_{mm}/\lambda_{univ} = o(1)$, performance bounds requiring penalty

levels $\lambda \geq \lambda_{univ}$ do not match the lower bounds of the minimax rate $R^r \lambda_{mm}^{q-r}$. We close this gap by providing a minimax upper bound for the tail quantile of the ℓ_q loss for the Lasso estimator with $\lambda \asymp \lambda_{mm} = o(\lambda_{univ})$ for $0 < r \leq 1 \leq q \leq 2$.

Chapter 2

Imputation of Blood Glucose Levels and HbA1c's for Automatic Control Algorithms

2.1 Introduction

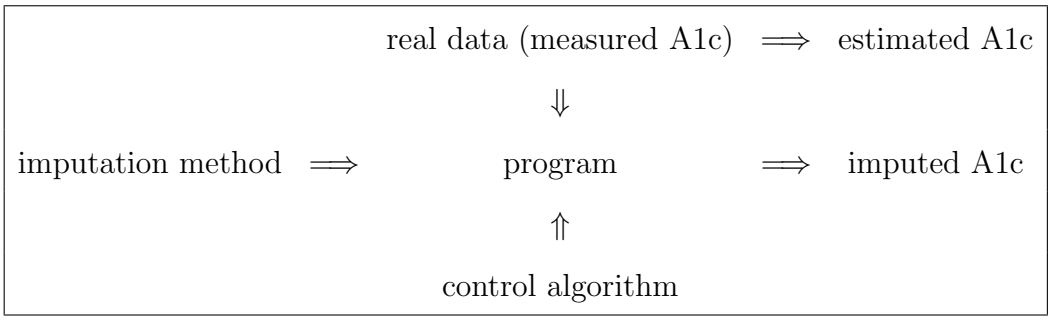
Diabetes is a disorder of metabolism preventing the human body from using digested food for energy and growth. Type 1 diabetes is a form of diabetes mellitus, an autoimmune disease that results in the permanent destruction of insulin-producing beta cells of the pancreas. The lack of insulin holds glucose in the blood, which would have been transfered into cells if enough insulin works in an appropriate way. When the blood glucose level rises above 180 mg/dL, it starts to appear in the urine making it sweet "mellitus". Type 1 diabetes is lethal unless treatment of insulin injection or a functional replacement of pancreatic insulin-producing beta cells is provided. On the other hand, overdose of insulin injection may cause ketoacidotic coma, an extremely dangerous situation. There is currently no preventive action that can be taken against type 1 diabetes. Although the cause of type 1 diabetes is still not fully understood, wise diet and appropriate exercise may help insulin to act more effectively.

Hemoglobin (Hb) is an oxygen-transporting compound in red blood cells. In the normal 60-120 day life span of a red blood cell, glucose molecules join hemoglobin, forming glycated hemoglobin. Once a hemoglobin molecule is glycated, it remains that way. HbA1c, or A1c, is the percentage of glycated hemoglobin molecules in a red blood cell. The higher blood glucose level, the more glucose molecules would join hemoglobin, resulting in a higher A1c level. Increases in

the quantities of glycated hemoglobins are observed for individuals with poorly controlled diabetes. Accumulation of glycated hemoglobins within a red blood cell reflects the average blood glucose level to which the cell has been exposed during its life cycle. The A1c level of a normal person is about 5 and an out-of-control diabetic patient has an A1c level of 8 or more. While hemoglobin A1c test is now considered the best and standard measure of diabetic hyperglycemia, effective management of type 1 diabetes usually indicates a lower A1c level while controlling hypoglycemia risk.

In the past few years many researchers have been working on the development of a feasible automated closed-loop insulin delivery system to replace the open-loop control based on prespecified basal rate and self-adjusted bolus rate. In self-adjusted open-loop control, continuous glucose monitor provides the possibility to alert patients to the presence of high and low blood glucose levels. From time to time additional bolus injection is performed to avoid hyperglycemia events in the presence of scheduled meals and extra food or snacks are eaten to avoid hypoglycemia events. However, many other factors other than carbohydrate intake affect glucose levels, including hormones, exercise, stress, illness, etc. Moreover, it is almost impossible for a patient to read his glucose levels "continuously" and take corresponding actions of "eat" and/or "inject". Patients are often distracted, misguided by what he thinks he should do, and overreact. For a "careless" diabetic patient, it could be extremely helpful to use a closed-loop system that automatically determines and delivers the necessary amount of insulin to control hyperglycemia events. To this end, it is necessary to have three major design elements for an automated closed-loop insulin delivery system: an insulin pump to deliver precise amount of insulin on time, a real-time glucose monitor to accurately measure blood glucose levels, and an effective algorithm with a few prespecified parameters to calculate the amount of insulin delivery from continuous glucose measurements.

Inspired by the novel nonparametric statistical imputation method of Mastrototaro *et al* [15], we study a semiparametric imputation method to simulate a time series of blood glucose level had certain closed-loop algorithm been in control of the diabetic patients, from the "frozen" measurements under self-adjusted open-loop control. The semiparametric imputation method utilizes an insulin absorption model and an insulin action model as parametric components and a nonparametric life-style component. Palerm, Shepp, Cabrera and Zhang [20] proposed a mathematical formula to estimate A1c based on continuous blood glucose levels. The A1c's under certain automatic control algorithm are imputed by applying the mathematical formula for A1c to the imputed time series of blood glucose level. The mathematical formula for A1c makes it possible to continuously evaluate the performance of a closed-loop control algorithm and the open-loop control. As an example in this chapter we impute the blood glucose levels under a simple proportional-integral-derivative (PID) algorithm. To this end, we will have three A1c's to compare for each subject. Measured A1c is the laboratory measured value at the end of the period. Estimated A1c is calculated from the "frozen" blood glucose measurements. PID-imputed A1c is calculated from the imputed blood glucose levels under the simple PID algorithm. We define a number of hypoglycemia measures and compare each of them between the PID algorithm and the open-loop control. The process can be summarized as in the following diagram:



The Star One data set of 137 diabetic patients is provided by Medtronic Inc.

Each of the 137 patients is equipped with an insulin pump and a continous glucose monitor. The monitor reads the blood glucose level $G(t)$ every 5 minutes. The pump delivers the insulin according to a prespecified basal rate and self-controlled bolus rate for meals. The patients receive no other source of insulin except through their insulin pumps. The times and dosages of the insulin deliveries are determined only by each patient. Although the Star One data set contains some additional information for each patient, for example, patient-estimated carbohydrate intake, we use three most relevant variables, the continuous blood glucose measurements (CGM) $G(t)$, the subcutaneous insulin deliveries $u(t)$, and the laboratory measured A1c's.

The rest of the chapter is organized as follows. In Section 2.2, we explain the research design and methods, including a semiparametric imputation method, a simple PID algorithm, and the mathematical formula for A1c. In Section 2.3, we screen the data set to obtain the subjects with the best quality of data and interpolate the missing values of $G(t)$ within each subject. In Section 2.4, numerical experiments are carried out to impute blood glucose levels and calculate A1c values of the corresponding subjects. In Section 2.5, we define hypoglycemia measures and evaluate hypoglycemia risk under the simple PID algorithm and the open-loop control. In Section 2.6, we make a few remarks.

2.2 Research Design and Methods

In this section, we propose a semiparametric imputation method under a simple PID algorithm and describe the mathematical formula for A1c.

2.2.1 Semiparametric Imputation Method

The semiparametric imputation method utilizes an insulin absorption model and an insulin action model as parametric components and a nonparametric life-style

component. For definiteness, we describe here the statistical methods based on a two-compartment plasma insulin concentration model of Hovorka *et al* [14] and a revised minimal insulin action model of Steil *et al* [21].

In Hovorka *et al* [14], the two-compartment modeling of insulin absorption subsystem is proposed to calculate the insulin concentration $I(t)$ from subcutaneous insulin delivery $u(t)$ (basal and bolus infusion). The pump may cause some delay because it is placed in subcutaneous fat rather than in bloodstream to avoid infection. For now we will use the two-compartment modeling of insulin subsystem and keep the values of model parameters. Insulin absorption and plasma insulin concentration are modeled as

$$\begin{aligned} &\frac{dS_1(t)}{dt} = u(t) - \frac{S_1(t)}{t_{max,I}}, \quad \frac{dS_2(t)}{dt} = \frac{S_1(t)}{t_{max,I}} - \frac{S_2(t)}{t_{max,I}}, \\ &U_I(t) = \frac{S_2(t)}{t_{max,I}}, \quad \frac{dI(t)}{dt} = \frac{U_I(t)}{V_I} - k_e I(t), \end{aligned} \tag{2.1}$$

where $S_1(t)$ and $S_2(t)$ are a two-compartment chain representing absorption of subcutaneous delivery of insulin and $U_I(t)$ is the insulin absorption rate (appearance of insulin in plasma). $t_{max,I} = 55$ min is the time-to-maximum insulin absorption, $k_e = 0.138$ min^{-1} is the fractional elimination rate and $V_I = 0.12$ L/kg is the distribution volume. Figure 2.1 illustrates the effect of 1 unit subcutaneous delivery $u(0)$ of insulin on the plasma insulin concentration $I(t)$ over the next 24 hours. Note that there is a gap of about an hour between the subcutaenous delivery and the peak of the plasma insulin concentration, and the coverage of subcutaneous delivery of insulin lasts about 5 hours. For simplicity we denote the insulin absorption system (2.1) as $H(\cdot)$ such that

$$I(t) = H(u(s), s \leq t)$$

Once the plasma insulin concentration $I(t)$ is calculated from the subcutaneous insulin delivery $u(t)$, a natural next-step is to model the action of insulin on the blood glucose level, along with the meal absorption. How insulin regulates glucose levels has not been fully understood and usually it is modeled as a

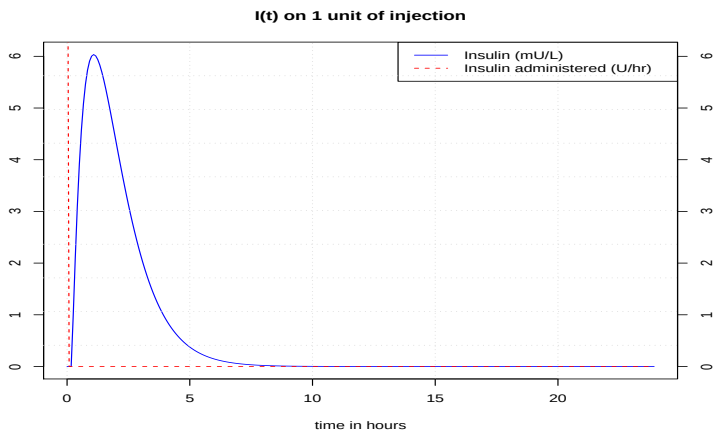


Figure 2.1: The insulin concentration after 1 unit subcutaneous delivery of insulin.

dynamic system. Blood glucose levels are quite volatile, even for healthy people without diabetic diseases. The ups and downs of glucose levels are similar to stock prices. Carbohydrate intake drives up the glucose level just as bid order pushes up the stock price, while insulin pulls down the glucose level as ask order does. And blood glucose levels are cycled daily, rise and fall serveral times due to scheduled meals. One example is the insulin action subsystem and the glucose subsystem of Hovorka *et al* [14], where the insulin action subsystem outputs the (remote) effects of insulin on glucose distribution/transport, glucose disposal and endogenous glucose production based on the input of plasma insulin concentration $I(t)$. And the glucose subsystem ouputs the masses of glucose in the accessible and non-accessible compartments based on the inputs from the insulin action subsystem. It is plausible but has many model constants to measure and many model parameters to calibrate. To keep model simple and make model parameters identifiable, we model the meal absorption and insulin-meal-glucose kinetics together as a semiparametric model. The nonparametric part is called the life-style component, inspired by the revised minimal insulin action model in Steil *et al* [21]. To account for conditions other than IVGTT such as meals and snacks, we could assume an arbitrary rate of glucose appearance $R_a(t)$ that is the

sum of endogenous and exogenous glucose sources. The revised minimal insulin action model could be written as

$$\frac{dG(t)}{dt} = -[p_1 + x(t)]G(t) + \frac{R_a(t)}{V_T}, \quad G(0) = G_0$$
$$\frac{dx(t)}{dt} = -p_2 x(t) + p_3 I(t), \quad x(0) = 0,$$

where p_1 is the "glucose effectiveness at zero insulin" in [21], or equivalently,

$$\frac{dG(t)}{dt} = \frac{R_a(t)}{V_T} - p_1 G(t) - p_3 G(t) \int_0^t e^{p_2(s-t)} I(s) ds.$$

The discretization of the above integro-differential equation could be written as

$$\begin{aligned} \frac{G(t+\Delta_t) - G(t)}{\Delta_t} &= \frac{R_a(t)}{V_T} - p_1 G(t) - p_3 \Delta_t G(t) \sum_{j=1}^{t/\Delta_t} e^{p_2(j\Delta_t - t)} I(j\Delta_t) \\ &= \frac{R_a(t)}{V_T} - p_1 G(t) - p_3 \Delta_t G(t) \sum_{s \le t} I(s) e^{-p_2(t-s)}. \end{aligned} \tag{2.2}$$

Under a regression setup, (2.2) suggests a simple semiparametric model

$$G(t+\Delta_t) - G(t) = \beta_0(t) + \beta_1 G(t) + \beta_2 G(t) \sum_{s \le t} I(s) e^{-w(t-s)}, \tag{2.3}$$

where $\beta_0(t)$ is the nonparametric life-style component of the patient, β_1, β_2 are unknown parameters and w is an unknown weight of the exponential moving average of $\{I(s), s \le t\}$. Before we calibrate the parameters $\{\beta_1, \beta_2, w\}$ and estimate the nonparametric $\beta_0(\cdot)$, it is necessary to remove the abnormal days that do not reflect the particular lifecycle of the patient during the period under investigation, based on the continuous blood glucose measurements. The data during these abnormal days are not included in estimating the life-style component or used in the imputation. K-means clustering on CGM's is applied to seperate the normal days of the subject.

Consider an individual patient in the Star One data set. Suppose measurements are taken in days $d = 1, \ldots, m$ at time points t_k, $k = 1, \ldots, n$ with fixed $\Delta_t = t_{k+1} - t_k$ (5 minutes). Let $u_d(t_k)$ be the actual amounts of subcutaneous insulin delivery and $G_d(t_k)$ be the actual blood glucose measurements. From $u_d(t_k)$,

the plasma insulin concentration $I_d(t_k)$ are computed using the insulin absorption model (2.1). We describe the semiparametric imputation method as follows:

Step 1: Estimate β_1, β_2, w and the life-style component $\beta_0(\cdot)$ using (2.3) and all time/day points. In this step, $\beta_0(\cdot)$ is approximated by a cubic spline function $\widehat{\beta}_0(\cdot)$ with appropriate number of knots to represent the carbohydrate intake and exercise as well as to achieve smoothness.

Step 2: Compute the adjustment $\epsilon_d(t_k)$ of life-style component for each day using the estimated parameters $\{\widehat{\beta}_1, \widehat{\beta}_2, \widehat{w}\}$, $\widehat{\beta}_0(\cdot)$ from Step 1 and the actual data for that day,

$$\epsilon_d(t_k) = G_d(t_{k+1}) - G_d(t_k) - \widehat{\beta}_0(t_k) - \widehat{\beta}_1 G_d(t_k) - \widehat{\beta}_2 G_d(t_k) \sum_{j \le k} I_d(t_j) e^{-\widehat{w}(t_k - t_j)}.$$

Step 3: Dynamically update $\widetilde{u}_d(t_k)$, $\widetilde{I}_d(t_k)$ and impute the blood glucose level $\widetilde{G}_d(t_{k+1})$ for a given closed-loop control algorithm $A(\cdot)$,

$$\begin{aligned}
&\widetilde{u}_d(t_k) = A(\widetilde{G}_d(t_j), j \le k) \\
&\widetilde{I}_d(t_k) = H(\widetilde{u}_d(t_j), j \le k) \\
&\widetilde{G}_d(t_{k+1}) = \widetilde{G}_d(t_k) + \widehat{\beta}_0(t_k) + \widehat{\beta}_1 \widetilde{G}_d(t_k) + \widehat{\beta}_2 \widetilde{G}_d(t_k) \sum_{j \le k} \widetilde{I}_d(t_j) e^{-\widehat{w}(t_k - t_j)} + \epsilon_d(t_k).
\end{aligned}$$

2.2.2 PID Algorithm

As an example we impute the glucose levels under a simple proportional-integral-derivative (PID) algorithm. PID algorithm is a generic closed-loop control algorithm widely used in engineering applications. PID algorithm could be used to control the blood glucose level toward a desired target by providing suitable dose of insulin delivery. In PID, the insulin dose is calculated using continuous glucose measurements $G(t)$ as a linear combination of three components: proportional (P), integral (I) and derivative (D). The proportional component delivers insulin to bring down the glucose level to the glucose target G_{target}. The integral and derivative components deal with the slow rise and rapid rise of glucose levels. PID

algorithm can be written as $PID(t) = Pcontrol(t) + Icontrol(t) + Dcontrol(t)$ where $Pcontrol(t) = \kappa_p(G(t) - G_{target})$, $Icontrol(t) = Icontrol(t-1) + \kappa_i(G(t) - G_{target})$ and $Dcontrol(t) = \kappa_d G'(t)$ with $G'(t)$ being the derivative of $G(t)$.

In our specific application, the target G_{target} is fixed at 100 mg/dL, $\kappa_p = 4.44 \times 10^{-4} TDD$, $\kappa_i = \kappa_p/150$ and $\kappa_d = 60\kappa_p$, where TDD denotes the estimated total daily dose and is approximately 0.6 of the body weight in kilograms, the time interval between two measurements is 1 minute, values of which are suggested by experts of Medtronic Inc.

The three components of PID algorithm can be understood from an intuitive perspective. The proportional component increases insulin delivery when the current glucose level is above the glucose target and reduces insulin delivery when glucose is below target. When the glucose level is at the target level, the integral component provides insulin for fasting glucose, and $Icontrol(t)$ adjusts upward when glucose level is above target, downward when glucose level is below target, helping to stabilize the system. The derivative component increases insulin delivery when the glucose level is rising and reduces insulin delivery when the glucose level is falling, using momentum indicator to stabilize the system.

2.2.3 Mathematical Formula for A1c

Parlerm, Shepp, Cabrera and Zhang [20] derived a simple but accurate theoretical formula to estimate A1c based on continuous blood glucose measurements. Once the relationship between A1c and blood glucose measurements is established, some questions can be answered, for instance, which one is A1c more related to, short-term or long-term glucose levels. More importantly, the semiparametric imputation method allows us to evaluate and compare the performance of different control algorithms, open-loop or closed-loop, in controlling hyperglycemia risk from the perspective of A1c value. For completeness, we briefly reproduce the mathematical A1c formula as follows.

Consider a hemoglobin molecule in a red cell. Suppose the survival model for a red blood cell is an exponential distribution with mean $\frac{1}{\gamma}$ days.

$$A1c/100 = P\left(glycated\right) = \int_0^\infty P\left(glycated|age = t\right)\gamma e^{-\gamma t}dt. \tag{2.4}$$

Another assumption is that glycation satisfies an inhomogeneous Poisson process with the hazard rate proportional to glucose level $G(t)$, i.e.,

$$P\left(glycated|age = t\right) = 1 - e^{-\int_0^t G(x)\alpha dx}. \tag{2.5}$$

Combine (2.4) and (2.5),

$$\begin{aligned} P\left(glycated\right) &= \int_0^\infty \left(1 - e^{-\int_0^t G(x)\alpha dx}\right)\gamma e^{-\gamma t}dt \\ &= 1 - \gamma\int_0^\infty e^{-\gamma t - \int_0^t G(x)\alpha dx}dt \end{aligned} \tag{2.6}$$

To estimate the parameters γ and α of (2.6), it is necessary to make some assumptions. Assume that the average lifetime of red blood cells is 120 days, i.e., $\gamma = 1/120$. Also assume that the glycation rate α is a constant for everyone, so that the data of a normal person can be used to estimate α. It is known that normals have $A1c = 5$ and assume $G(t) \equiv 100$ mg/dL. Newton's method is applied to calculate the reasonable α for a given subject in practice.

2.3 Data Screening and Interpolation of Missing Values

Due to unknown causes, there are quite a few missing blood glucose measurements $G(t)$ listed as "NA" in the Star One data set. This section explains how we screen the data and select 21 subjects with the best quality of data and interpolate the missing values of $G(t)$ within each subject. Since the goal is to compare PID-imputed A1c's with estimated and laboratory measured A1c's, it is natural to partition the data set so that each 3-month data $G(t)$, $u(t)$ of a patient is one subject under investigation with the corresponding laboratory measured A1c at the end of the period.

The mathematical formula for A1c in [20] implies that the weights of recent (late) glucose levels in determining A1c's are exponentially larger than the distant (early) ones. If there are a lot of missing CGM values near the end of the 3-month period, it would bring larger bias for estimation of the "true" A1c's. Therefore we use the following criterion to assign a score of data quality to each subject. If the percentage of the total missing glucose levels for a day is less than 25%, we tag that day with "not missing". The score of data quality for a subject is defined as $\sum_{j=0}^{J-1} I\{\text{day}_{J-j} \text{ not missing}\}R^j$, where R satisfies $\sum_{j=0}^{J-1} R^j = 20$ and J is the length of the period. We use 18 out of 20 (90%) as a cut-off threshold for good data quality. Under this criterion, we have a total of 21 subjects out of 137 patients. The "Score" column of Table 2.1 summarizes the scores of these 21 subjects with good quality of data.

For each subject, it is possible that CGM's of certain day tagged with "not missing" still have some missing values, but the missing percentage is less than 25%. Then we use cubic splines to interpolate these missing values. Figure 2.2 is a typical example of the cubic spline interpolation. Note the difference had linear interpolation been used instead of cubic spline interpolation. For the missing values of a day tagged with "missing", we keep them as is and include them in the A1c calculation but not the imputation.

2.4 Numerical Experiments

After the data is well screened and part of missing CGM values are interpolated, we impute the blood glucose levels under the simple PID algorithm and calculate A1c's of the imputed glucose levels as well as the "frozen" ones.

Figure 2.3 illustrates the life-style component $\widehat{\beta}_0(\cdot)$ of patient 14 approximated by a cubic spline with 23 internal knots, and the average blood glucose levels between 8/8/2005 and 11/3/2005. The life-style component has 3 peaks, indicating

Patient	Period	Score	PID-imputed A1c	Estimated A1c	Measured A1c
14	1	19.244	6.501161	6.926936	7.0
14	2	19.065	6.790962	7.365180	7.0
17	1	19.710	6.780308	7.613608	7.7
17	2	18.662	6.619161	7.212557	7.3
25	3	19.516	10.08607	8.114699	7.4
25	4	18.907	7.623908	8.311716	7.5
32	1	19.514	6.611458	7.439253	6.6
32	2	18.163	6.646665	7.253497	6.5
37	1	18.717	6.531904	6.994606	7.2
37	2	19.441	6.818421	7.403225	7.4
55	1	19.518	6.549876	7.492541	6.7
55	2	19.855	6.656363	7.675616	6.6
55	3	19.731	6.637304	7.841469	7.0
79	1	19.466	7.053136	8.457546	7.9
79	2	18.167	6.606708	7.649595	7.7
114	1	18.284	10.84451	11.100841	9.1
122	1	18.277	6.642880	7.671015	7.7
122	2	19.823	6.666476	7.541406	7.8
125	1	19.338	6.470646	7.009722	6.2
125	2	18.659	6.516138	6.989586	6.5
134	1	18.739	8.803752	11.738476	9.7

Table 2.1: Scores of data quality, 3 types of A1c's of 21 subjects: PID-imputed A1c's, estimated A1c's based on "frozen" CGM's and laboratory measured A1c's.

patient 14 normally had 3 meals per day during that period that are responsible for the corresponding climb-ups of the average glucose level. Note that $\widehat{\beta}_0(\cdot)$ is not always positive. Carbohydate intake would push up the curve and exercise would have the opposite effect. Regular snacks should be taken by type 1 diabetes patients to avoid low blood glucose levels due to the continuous basal infusion of insulin. There is some delay between carbohydrate intake and glucose level climb-up. $\widehat{\beta}_2 = -0.00026$ is expected to be negative, indicating that insulin does help bring down glucose level.

Table 2.1 compares the A1c's under the PID and the open-loop control. PID-imputed A1c's are estimated A1c values based on PID-imputed blood glucose levels and (2.6). Estimated A1c's aer estimated A1c values based on the "frozen"

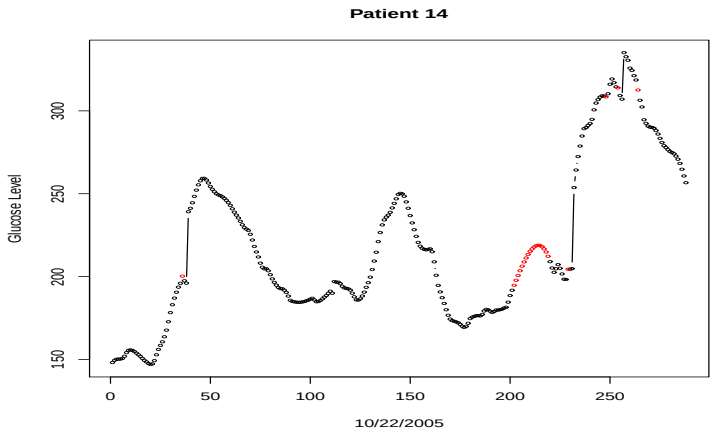


Figure 2.2: The illustrative example of cubic spline interpolation of missing values of glucose measurements of patient 14 on 10/22/2005.

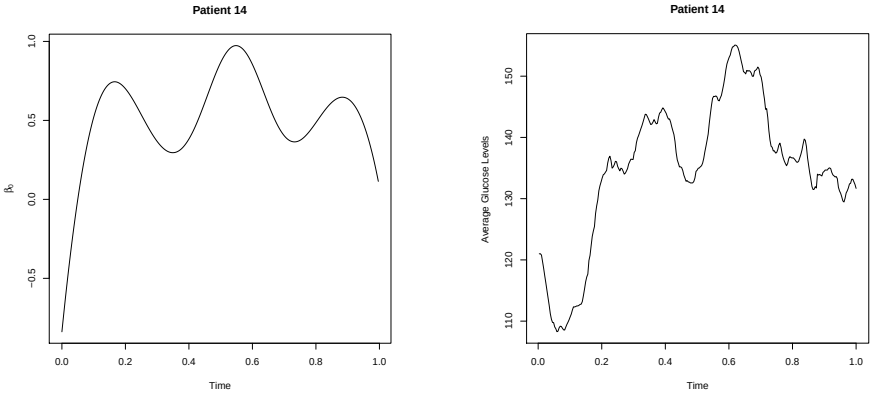


Figure 2.3: Left: the life-style component of patient 14; right: average blood glucose levels of patient 14 from 8/8/2005 to 11/3/2005.

blood glucose levels and (2.6). The laboratory measured A1c's are the true A1c's at the end of a 3-month period. Figure 2.4 visualizes the comparison between PID-imputed and measured A1c's. The blue straight line is a diagonal line of equal PID-imputed and measured A1c's. Out of the 21 subjects with good quality of data, there are 8 subjects when the open-loop control performs better than the simple PID algorithm, and for the rest 13 subjects the PID decreases A1c's significantly. A careful look into these patients where the PID loses tells us that the estimated A1c's differ from the laboratory measured A1c's significantly. This phenomenon occurs in different A1c periods of the same few patients. Palerm,

Shepp, Cabrera and Zhang [20] claims that the A1c value treats some individuals differently than it treats others. Among the most significant determinants of A1c is the biologic difference in the rate of glycation. In our cases, patient 114 and 134 might be "overwarned" by the mathematically estimated A1c's in that the mathematically estimated A1c's are at least 2 points higher than the laboratory measured ones.

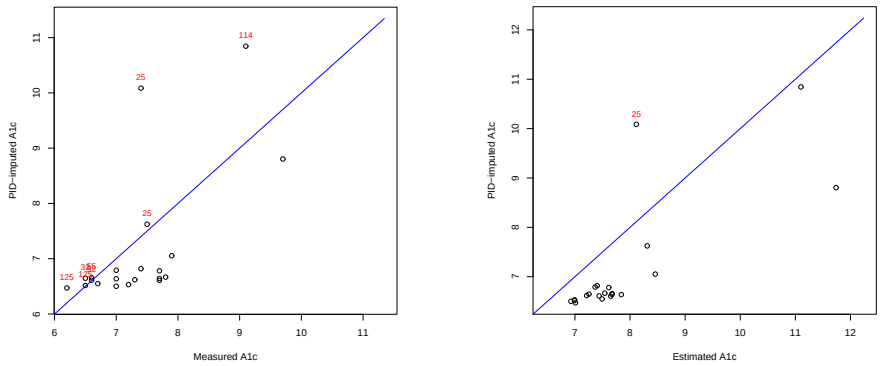


Figure 2.4: Left: PID-imputed A1c's and measured A1c's; right: PID-imputed A1c's and estimated A1c's.

It is not fair to compare the PID-imputed and measured A1c's, because the former are estimated using the mathematical formula for A1c (2.6) and the latter are true values without estimation error. Thus the comparison should be made between PID-imputed and estimated A1c's. Figure 2.4 provides the comparison between PID-imputed A1c's and mathematically estimated ones. It is found that the simple PID algorithm lowers A1c's compared with the open-loop control, expect during period 3 of patient 25. This finding agrees with the results of Mastrototaro *et al* [15]. A simple t-test showes the difference between PID-imputed A1c's and estimated A1c's is significant. Table 2.2 shows that period 3 of patient 25 has the smallest sample standard deviation 44.3 among all 21 subjects. It is possible that patient 25 was a careful diabetic patient during period 3 who knew what had been eaten, what would be eaten, what exercise had been and would

be performed, and what the stress and illness level was, while the PID had no information other than the continuous glucose measurements. Therefore the simple PID algorithm without tuning the parameters cannot guarantee a better solution for this subject.

Table 2.2 shows that the PID reduces the average blood glucose levels with only a relatively small contribution to the variability of glucose levels within the 21 subjects under investigation.

Patient	Period	CGM	CGM (PID)
14	1	138.96 (48.3)	127.54 (49.7)
14	2	147.93 (57.0)	135.28 (57.9)
17	1	153.92 (58.6)	135.24 (64.3)
17	2	147.15 (52.6)	131.77 (57.8)
25	3	162.22 (44.3)	200.71 (63.4)
25	4	168.50 (51.4)	154.34 (52.7)
32	1	151.15 (56.3)	133.75 (58.9)
32	2	145.35 (55.4)	132.35 (60.1)
37	1	145.39 (53.4)	133.07 (54.9)
37	2	150.84 (52.8)	137.45 (54.4)
55	1	151.41 (44.1)	131.25 (49.3)
55	2	158.25 (48.4)	135.00 (54.8)
55	3	159.37 (48.0)	134.98 (54.9)
79	1	173.76 (58.3)	142.36 (63.4)
79	2	158.00 (54.7)	138.06 (57.9)
114	1	227.62 (48.4)	221.68 (48.3)
122	1	152.76 (57.5)	132.75 (62.0)
122	2	154.35 (58.0)	135.36 (63.7)
125	1	148.65 (50.9)	133.83 (52.9)
125	2	140.69 (49.9)	130.98 (54.0)
134	1	244.04 (52.4)	182.49 (71.9)

Table 2.2: Some descriptive statistics of "frozen" CGM's and PID-imputed CGM's for 21 subjects: sample mean and sample standard deviation.

Let us take a close look at how differently the simple PID algorithm would work from the open-loop control. Figure 2.5 illustrates the glucose levels, subcutaneous insulin deliveries and plasma insulin concentration for patient 14 on 8/9/2005. As we can see, continuous insulin deliveries calculated from our semi-parametric imputation model under the PID have rapid effect on reducing the

blood glucose levels.

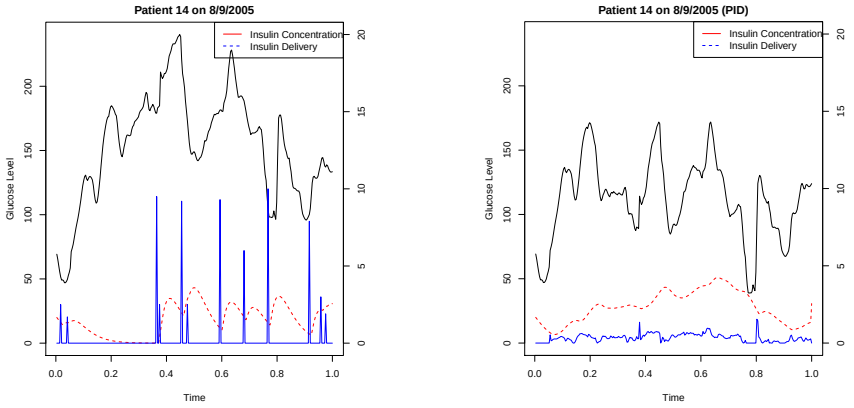


Figure 2.5: Left: glucose levels, insulin deliveries and insulin concentration of patient 14 on 8/9/2005; right: imputed glucose levels, calculated insulin deliveries and insulin concentration of patient 14 on 8/9/2005.

Before the feasibility of an automated system of insulin delivery has been fully validated, it is difficult as well as dangerous to test the performance of an automatic control algorithm on diabetic patients for a long period in a clinical trial. The semiparametric imputation method makes it possible to evaluate the performance of different control algorithms using a "frozen" data set. The Star One data set provides evidence that the simple PID algorithm has good performance in controlling hyperglycemia. However, we have yet tested any other closed-loop control algorithm thus there is no evidence that PID is the best. And the question whether closed-loop or open-loop control is better for a careful diabetic patient is not fully answered.

2.5 Hypoglycemia Measures

Subsection 2.2.3 quantifies hyperglycemia risk as A1c and Section 2.4 evaluates the numerical performance of PID algorithm in controlling A1c from above. Besides hyperglycemia, the other risk of diabetic disorder is hypoglycemia, which could lead to ketoacidotic coma. Hypoglycemia happens when a diabetic patient

injects too much insulin and has no matching carbohydrate intake to spend the insulin. A diabetic patient would feel uncomfortable or hungry if the period when glucose level dives below 60 mg/dL lasts for a long time and happens frequently. A reasonable closed-loop control algorithm should keep low both risks of hyperglycemia and hypoglycemia. To this end, we define a number of hypoglycemia measures: the number of hypoglycemia episodes of a subject, the duration of each hypoglycemia episode, and the average area underneath the threshold of 60 mg/dL, $AUT = \frac{\sum\{60-G(t)\}_+}{\sum I_{G(t)<60}}$, whereas the latter two measures are more important than the first one.

Table 2.3 summarizes three hypoglycemia measures for the PID and open-loop control. For almost all subjects, medians of PID-imputed AUT's are smaller than 10 and medians of PID-imputed durations underneath the threshold of 60 mg/dL are shorter than 1 hours. Incidence of hypoglycemia is similar for the PID and open-loop control. There are no episodes of severe hypoglycemia under the PID. Although the simple PID algorithm has good performance in controlling hyperglycemia, we find it difficult for the PID to further lower hypoglycemia risk, compared with the open-loop control. The reason is that the Star One data set has the survival bias. When diabetic patients feel uncomfortable or hungry, they would simply eat some additional food to keep up their blood glucose levels. It is challenging for closed-loop control algorithms to control hypoglycemia risk because algorithms do not deliver food to spend the insulin. This component is not imputed. Figure 2.6 is an example of patient 79 on 1/4/2006 illustrating when and how hypoglycemia events happen under the PID. From 5:00 to 8:30, the PID worked well to reduce the glucose level towards the glucose target 100 mg/dL slowly. When the patient ate the breakfast during 8:30 and 9:00, the glucose level rose rapidly and the PID algorithm started to alert the insulin pump to deliver more insulin. While there was some delay before insulin action on the glucose

level, the pump kept delivering insulin so that the insulin concentration accumulated and kept reducing the glucose level below 60 mg/dL until the algorithm found out the hypoglycemia event or the patient had lunch to keep up the glucose level.

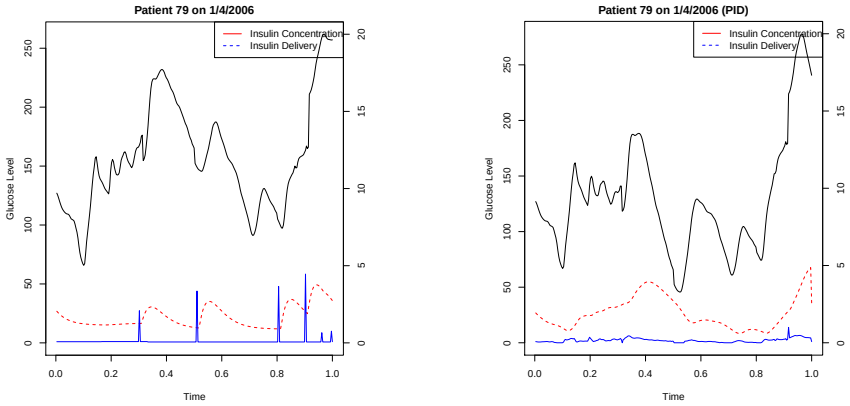


Figure 2.6: Left: glucose levels, insulin deliveries and insulin concentration of patient 79 on 1/4/2006; right: imputed glucose levels, calculated insulin deliveries and insulin concentration of patient 79 on 1/4/2006.

Patient	Period	# Episodes (PID)	Duration (PID)	AUT (PID)
14	1	34 (85)	0.38 (0.50)	2.12 (3.71)
14	2	46 (101)	0.42 (0.58)	4.19 (4.94)
17	1	53 (165)	0.50 (0.75)	4.39 (8.46)
17	2	46 (127)	0.29 (0.92)	2.56 (9.24)
25	3	4 (3)	0.21 (0.17)	3.56 (1.84)
25	4	0 (33)	0 (0.33)	0 (2.85)
32	1	54 (142)	0.38 (0.54)	3.65 (7.26)
32	2	51 (141)	0.46 (0.67)	4.12 (6.60)
37	1	34 (96)	0.46 (0.75)	3.80 (7.00)
37	2	20 (98)	0.33 (0.58)	5.14 (6.15)
55	1	9 (114)	0.08 (0.71)	2.18 (7.65)
55	2	5 (126)	0.17 (1.00)	2.35 (10.42)
55	3	7 (103)	0.42 (1.08)	5.14 (9.35)
79	1	1 (107)	0.75 (1.17)	2.31 (12.70)
79	2	40 (126)	0.17 (0.58)	3.99 (7.17)
114	1	0 (0)	0 (0)	0 (0)
122	1	43 (131)	0.25 (0.83)	4.27 (9.11)
122	2	41 (143)	0.42 (0.92)	4.10 (7.43)
125	1	18 (88)	0.50 (0.58)	3.70 (6.26)
125	2	84 (196)	0.25 (0.42)	3.26 (5.75)
134	1	0 (42)	0 (0.50)	0 (10.00)

Table 2.3: Some hypoglycemia measures of "frozen" CGM's and PID-imputed CGM's for 21 subjects: number of hypoglycemia eposides, median of durations (in hours), median of average areas underneath the threshold of glucose level 60 mg/dL.

2.6 Discussion

This chapter proposes an imputation method to simulate a time series of blood glucose level under certain closed-loop control algorithm, from the "frozen" measurements. Use the mathematical formula for A1c to quantify hyperglycemia risk and define a few hypoglycemia measures, it is possible to evaulate and compare the performance of different closed-loop control algorithms and self-adjusted open-loop control. We apply a simple PID algorithm as a closed-loop control algorithm while the glucose target G_{target} and 3 parameters κ_p, κ_i, κ_d of the algorithm are fixed arbitrarily in the imputation of glucose levels. These parameters could be tuned for individual diabetic patient to minimize a combined risk of

hyperglycemia and hypoglycemia, for example, duration-penalized and/or AUT-penalized A1c value. However, we do not include the optimization of the simple PID algorithm because the main message we want to deliver in this chapter is the idea of imputation from "frozen" measurements.

In our numerical experiment, we have seen that for a careful type 1 diebetic patient 25, it is difficult for the PID to gain any edge from the perspective of lowering the blood glucose level. Can any closed-loop control algorithm beat self-adjusted open-loop control? This kind of question is hard to answer. One of the reasons is that there is quite some delay between the subcutaneous insulin delivery and the insulin action on glucose levels. Therefore a closed-loop control algorithm based only on continuous glucose measurements would have difficulty competing with the action a careful diabetic patient would take if the patient has some experience dealing with type 1 diabetes and keeps a regular and healthy life style.

Chapter 3
Selection and Estimation with the ℓ_1 Penalty

3.1 Introduction

As modern information technologies relentlessly generate increasingly voluminous and complex data, penalized high-dimensional regression methods have been the focus of intense research activities in machine learning and statistics in the past few years.

In many statistical and engineering applications, the number p of design variables (features, covariates) can be larger or even of larger order than the sample size n, but the number of important variables is still smaller than the sample size. In the microarray technology, the expression levels of thousands of genes are collected simultaneously from a relatively small number of samples. In signal processing (sparse recovery, compressed sensing), a p-dimensional signal is encoded through a linear transformation to an n-dimensional vector, stored or transmitted in the n-dimensional form, and then recovered. In such applications, one seeks a parsimonious model that fits the data well. Many applications also require an easy interpretation of the selected model. In linear regression, a popular approach for model selection is to impose a suitable penalty on the empirical loss.

This chapter considers variable selection and estimation of a sparse vector of regression coefficients in a linear model. Specifically, we are interested in the variable selection consistency of threshold Dantzig selectors and the rate minimaxity of the Dantzig and Lasso estimators under the ℓ_q loss for regression coefficients in ℓ_r balls. The first goal requires an upper bound for the ℓ_∞ loss of the Dantzig

selector. The second goal requires lower bounds of the minimax ℓ_q risk and minimax (tail quantiles of the) ℓ_q loss over all estimators as well as matching upper bounds for the Dantzig and Lasso estimators.

Let $\boldsymbol{y} \in \mathbb{R}^n$ be a response vector and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ be a design matrix. The Lasso [22] is an ℓ_1-penalized estimator

$$\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda \|\boldsymbol{b}\|_1 \right\} \tag{3.1}$$

for the regression coefficients. In the signal processing literature, the Lasso is known as basis pursuit [6]. The Lasso has the interpretation as boosting [11, 12] and is computationally feasible for high-dimensional data [18, 19, 9]. Recently Candes and Tao [4] proposed another ℓ_1-penalized method called the Dantzig selector,

$$\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda) = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{b}\|_1 : |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n| \le \lambda, \forall j \right\}. \tag{3.2}$$

It is a simple convex program that can be recast as a convenient linear program.

Variable selection is fundamental for the interpretation of models with hign-dimensional data in statistical, engineering and social science applications. An estimator is variable selection consistent if the set of nonzero estimated coefficients matches that of the "true" nonzero regression coefficients with large probability. In general, consistent variable selection implies near optimal parameter estimation and prediction. One of the important features of the Lasso is that it can be used for variable selection. Meinshausen and Buhlmann [16], Tropp [23], Zhao and Yu [29] and Wainwright [25] proved that the Lasso is variable selection consistent under a strong irrepresentable condition on the Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$ and some other regularity conditions. Zhang and Huang [27] proved the consistency of the Lasso in the order of the dimension and bias of the selected model under a regularity condition on the eigenvalues of sub-Gram matrices. More recently, Candes and Plan [5] proved the selection consistency of the Lasso under random permutation and sign-change of regression coefficients and a mild condition on

the maximum absolute correlation among design vectors. Zhang [28] studied the selection consistency of the Lasso through its ℓ_∞ loss. Although the Dantzig and Lasso estimators are quite similar, it is still unclear in the existing literature if the Dantzig selector possesses selection consistency properties parallel to those mentioned above for the Lasso.

Another focus of recent studies of the ℓ_1-penalized least squares estimators has been on the estimation loss for the regression coefficients. Candes and Tao [4] derived an elegant probabilistic upper bound of the ℓ_2 loss for the Dantzig selector under a condition on the number of nonzero coefficients and a uniform uncertainty principle (UUP) on the Gram matrix. Efron, Hastie and Tibshirani [10] questioned whether a similar performance bound holds for the Lasso estimator as well. Upper bounds for the ℓ_q loss of the Lasso estimator has being studied by Bunea, Tsybakov and Wegkamp [3] and van de Geer [24] for $q = 1$, Zhang and Huang [27] for $q \in [1, 2]$, Meinshausen and Yu [17] for $q = 2$, Bickel, Ritov and Tsybakov [2] for $q \in [1, 2]$ with a parallel analysis of the Dantzig selector, and Zhang [28] for $q \geq 1$. Under different sets of regularity conditions on the Gram matrix and sparsity of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, these results provide upper bounds of the form $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \leq O(k^{1/q}\lambda)$, where k is of the order of an intrinsic dimensionality of the sparse estimation problem. For $N(0, \sigma^2)$ errors and standardized designs with $\|\boldsymbol{x}_j\| = \sqrt{n}$, the required penalty levels λ in these studies on the Lasso are all greater by a constant factor than the universal penalty level $\sigma\sqrt{(2/n)\log p}$ in the inequality of Candes and Tao [4]. Different sets of regularity conditions lead to different forms of constant factors in the upper bounds so that the existing upper bounds are typically not directly comparable mathematically. Technical discussion on the constant factors and conditions on the Gram matrix will be presented after we introduce the necessary terminologies and specific results.

Although this chapter focuses on the selection of variables and estimation

of regression coefficients, we would like to mention that the prediction of future responses is another important question in high-dimensional data. In a vague sense, the estimation of regression coefficients is related to the prediction or the estimation of the mean response $E\boldsymbol{y}$ as it could be viewed as a careful application of a suitable partial inversion of the design matrix $\boldsymbol{X}$ to a good predictor, in view of the persistency of the Lasso [13] and the convergence rate $k^{1/q}\lambda$ discussed in the previous paragraph.

The main results of this chapter contribute to two specific problems in high-dimensional regression. Firstly, under an assumption on the ℓ_0 sparsity of a target vector of regression coefficients, we obtain simple upper bounds for the ℓ_q loss of the Dantzig and Lasso estimators. Our upper bounds sharpen and unify a number of existing approaches and extend the inequality of Candes and Tao [4] from $q = 2$ to $1 \le q \le \infty$ and the Lasso. By explicitly allowing $q = \infty$ and thresholding the Dantzig selector, the bounds imply variable selection consistency. Secondly, we prove that both the Dantzig and Lasso estimators are rate minimax in the ℓ_q risk and loss for the estimation of regression coefficients in ℓ_r balls. This requires lower bounds for general estimators and matching upper bounds for the Dantzig and Lasso estimators. For $0 < r \le q$, we prove that the minimax ℓ_q risk and loss in ℓ_r balls are bounded from below by $R^r\lambda_{mm}^{q-r}$, where λ_{mm} is a certain minimax penalty level and R is the radius of the ℓ_r ball. These lower bounds extend the results of Donoho and Johnstone [7] from orthonormal designs. When λ_{mm} is of the same order as $\sigma\sqrt{(2/n)\log p}$, we prove that the Dantzig and Lasso estimators attain the rate of the minimax risk and loss for $0 < r \le 1 \le q$. We also prove that the Lasso attains the minimax rate for the ℓ_q loss in ℓ_r balls for $0 < r \le 1 \le q \le 2$ in the difficult case of $\lambda \asymp \lambda_{mm} = o(\sigma\sqrt{(2/n)\log p})$.

The rest of the chapter is organized as follows. In Section 3.2, we study error bounds and variable selection under the ℓ_0 sparsity of regression coefficients. In Section 3.3, we study the estimation of regression coefficients under the ℓ_q loss in

ℓ_r balls and provide non-probabilistic oracle inequalities for approximation of a given target vector $\boldsymbol{\beta}^*$. In Section 3.4, we provide all proofs. In Section 3.5, we make a few final remarks.

We use the following notation throughout the sequel. For vectors $\boldsymbol{v} = (v_1, \ldots, v_p)'$, $\|\boldsymbol{v}\|_0 = \#\{j : v_j \neq 0\}$ and $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2$ and the usual extension to $q = \infty$. Functions are applied to vectors in individual components, $f(\boldsymbol{v}) = (f(v_1), \ldots, f(v_p))'$. For matrices $\boldsymbol{M}$ and $0 \leq a, b \leq \infty$, $\|\boldsymbol{M}\|_{a,b} = \max\{\|\boldsymbol{M}\boldsymbol{v}\|_b : \|\boldsymbol{v}\|_a = 1\}$ is the operator norm from ℓ_a to ℓ_b. For subsets A and B of $\{1, \ldots, p\}$, $\boldsymbol{X}_A = (\boldsymbol{x}_j, j \in A)$, $\boldsymbol{\Sigma}_{A,B} = \boldsymbol{X}'_A\boldsymbol{X}_B/n$, $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{A,A}$, and $\boldsymbol{P}_A$ is the projection from $\mathbb{R}^n$ to the linear span of $\{\boldsymbol{x}_j : j \in A\}$. For real x, $x_+ = \max(x, 0)$ and $1/x_+ = \infty$ for $x \leq 0$. For simplicity, the dependence of the Dantzig and Lasso estimators on the penalty level λ is suppressed unless otherwise stated.

3.2 Error bounds and variable selection under ℓ_0 sparsity

Two types of error bounds will be considered in this chapter. The first type specifies sparse vectors $\boldsymbol{\beta}$ of regression coefficients (or targets $\boldsymbol{\beta}^*$) and conditions on the data $(\boldsymbol{X}, \boldsymbol{y})$ for upper bounds of the ℓ_q loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q$. The second type provides upper bounds for the ℓ_q risk $E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q$ and quantiles of the ℓ_q loss under certain probability measures $P_{\boldsymbol{\beta}}$. For simplicity, we assume

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j + \boldsymbol{\varepsilon} \tag{3.3}$$

with $\|\boldsymbol{x}_j\|^2 = n$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ under $P_{\boldsymbol{\beta}}$, although the normality and unbiasedness assumptions can be weaken as discussed in Section 3.5. Theorem 1 below presents both types of error bounds respectively in two parts under the ℓ_0 sparsity of $\boldsymbol{\beta}$. Define

$$\lambda_{mm} = \sigma\Big\{\frac{2}{n}\log\Big(\frac{\sigma^r p}{n^{r/2}R^r}\Big)\Big\}^{1/2}, \quad \lambda_{univ} = \sigma\sqrt{(2/n)\log p}, \tag{3.4}$$

as penalty levels associated with the variance σ^2 under $P_{\boldsymbol{\beta}}$ and the radius R of ℓ_r balls.

Theorem 1. *(i) Let $q \in [1,\infty]$, $\boldsymbol{\beta} \in \mathbb{R}^p$ with $J = \{j : \beta_j \neq 0\}$ and $\|\boldsymbol{\beta}\|_0 = |J| = k \geq 0$, $1 \leq \ell \leq p-k$ and $z_\infty^* = \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n\|_\infty$. Let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda)$ with $\{\widetilde{\alpha}, \xi, \widetilde{C}\} = \{0, 1, 2\}$ or $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}(\lambda/\alpha)$ with $\alpha \in (0,1)$ and $\{\widetilde{\alpha}, \xi, \widetilde{C}\} = \{\alpha, (1+\alpha)/(1-\alpha), 1 + 1/\alpha\}$. For $A \subset \{1, \ldots, p\}$, let $T_A(\boldsymbol{u})$ be mappings from $\mathbb{R}^A$ to $\mathbb{R}^A$ and define $\boldsymbol{w} = T_A(\boldsymbol{u})/(\boldsymbol{u}'\boldsymbol{\Sigma}_A T_A(\boldsymbol{u}))_+$ as a function of $\{A, \boldsymbol{u}\}$. Then, in the event $z_\infty^* \leq \lambda$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_s \leq \max_{A,B,\boldsymbol{u}} \frac{\{(1+\xi)\|\boldsymbol{u}_J\|_1\}^{(q/s-1)/(q-1)} G_{A,\boldsymbol{u}}}{(1 + \xi^q (k/\ell)^{q-1})^{(1/s-1)/(q-1)} (1 - \xi F_{A,B,\boldsymbol{u}})_+}, \ 1 \leq s \leq q, \quad (3.5)$$

where $F_{A,B,\boldsymbol{u}} = \|\boldsymbol{u}_J\|_1 \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell$, $G_{A,\boldsymbol{u}} = \widetilde{C}\lambda\|\boldsymbol{w}\|_1 \min\{1, (1+\xi)\|\boldsymbol{u}_J\|_1/\|\boldsymbol{u}\|_1\}$ and the maximum is taken over $A \supset J$ with $|A| = k+\ell$, $B \cap A = \emptyset$ with $|B| \leq \ell$ and $\boldsymbol{u} \in \mathbb{R}^A$ with $\|\boldsymbol{u}\|_q = 1$. Moreover, if $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}(\lambda/\alpha)$ and $\mathrm{sgn}(T_A(\boldsymbol{u})) = \mathrm{sgn}(\boldsymbol{u})$, then (3.5) also holds with $G_{A,\boldsymbol{u}} = \widetilde{C}\lambda\|\boldsymbol{w}_J\|_1$.
(ii) Let $P_{\boldsymbol{\beta}}$ be probabilities giving the linear model (3.3). For $\lambda = \sigma\sqrt{(2/n)\log(p/\epsilon)}$, (3.5) holds with at least probability $P_{\boldsymbol{\beta}}\{z_\infty^* \leq \lambda\} \geq 1 - \epsilon/\sqrt{\pi \log(p/\epsilon)}$.

In the rest of the section, we discuss an implication of Theorem 1 on variable selection and its connections to the results of Candes and Tao [4], Bickel, Ritov and Tsybakov [2] and Zhang [28] via different choices of the mappings $T_A(\boldsymbol{u})$. We shall focus on the Dantzig selector and omit parallel statements for the Lasso with different $(\widetilde{\alpha}, \xi, \widetilde{C})$.

For $(|A|, |B|, \|\boldsymbol{u}\|, \|\boldsymbol{v}\|) = (d, \ell, 1, 1)$ with $A \cap B = \emptyset$, define

$$\delta_d^{\pm} = \max_{A,\boldsymbol{u}} \Big\{ \pm \Big(\|\boldsymbol{\Sigma}_A \boldsymbol{u}\| - 1 \Big) \Big\}, \ \delta_d = \delta^+ \vee \delta^-, \ \theta_{d,\ell} = \max_{A,B,\boldsymbol{u},\boldsymbol{v}} \boldsymbol{v}'\boldsymbol{\Sigma}_{B,A}\boldsymbol{u}. \quad (3.6)$$

For $q = 2$, (3.5) with the option $T_A(\boldsymbol{u}) = \boldsymbol{u}$ gives

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}}_{Dantzig} - \boldsymbol{\beta}\|_s &\leq \max_{A,B,\boldsymbol{u}} \frac{2^{2/s}\{\|\boldsymbol{u}\|_1 \wedge (2\|\boldsymbol{u}_J\|_1)\}(1 + k/\ell)^{1-1/s}\|\boldsymbol{u}_J\|_1^{2/s-1}\lambda}{(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u} - \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{u}\|_1\|\boldsymbol{u}_J\|_1/\ell)_+} \\ &\leq \frac{2^{2/s}\{(1+\ell/k)^{1/2} \wedge 2\}(1+k/\ell)^{1-1/s} k^{1/s}\lambda}{(1 - \delta_{k+\ell}^-)_+\{1 - (k^{1/2}/\ell)\max_{A,B,\boldsymbol{u}} \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{u}\|_1/\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u}\}_+}, \end{aligned}$$

where the worst $\{A, \boldsymbol{u}\}$ are taken separately with $\|\boldsymbol{u}\|_1 \le \sqrt{|A|} = \sqrt{k+\ell}$ and $\|\boldsymbol{u}_J\|_1 \le \sqrt{k}$ by Cauchy-Schwarz, and the lower bound $\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u} \ge 1 - \delta^-_{k+\ell}$ is used.

Corollary 1. *Suppose* $\|\boldsymbol{\beta}\|_0 = k \ge 0$. *Then, for* $\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty \le \lambda$ *and* $1 \le q \le 2$,

$$\|\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda) - \boldsymbol{\beta}\|_q \le \frac{2^{2/q}(1 + k/\ell)^{1-1/q}\{(1+\ell/k)^{1/2} \wedge 2\}k^{1/q}\lambda}{(1 - \delta^-_{k+\ell})_+\{1 - \widetilde{F}\}_+}, \tag{3.7}$$

where $1 \le \ell \le p - k$ *and for* $|A| = k + \ell$, $|B| = \ell$ *and* $A \cap B = \emptyset$,

$$\widetilde{F} = \min\Big\{\frac{k^{1/2}\max_{A,B}\|\boldsymbol{\Sigma}_{B,A}\|_{2,1}}{\ell(1 - \delta^-_{k+\ell})_+}, \frac{\sqrt{k/\ell}\,\theta_{\ell,k+\ell}}{(1 - \delta^-_{k+\ell})_+}, \sqrt{\frac{k(1+\delta^+_\ell)}{\ell(1 - \delta^-_{k+\ell})_+}}\Big\}. \tag{3.8}$$

The first upper bound in (3.8) is of the sharpest form among the three due to the factor ℓ in the denominator. For $(\ell, q) = (k, 2)$, inserting the second upper bound in (3.8) and the inequality $\delta^-_{2k} \le \delta_{2k}$ to (3.7) yields the upper bound $4\sqrt{k}/(1 - \delta_{2k} - \theta_{k,2k})_+$ of Candes and Tao [4]. The upper bound (7.6) of Bickel, Ritov and Tsybakov [2] is of a different form. However, inserting the third upper bound in (3.8) to (3.7) improves upon inserting Lemma 4.1 (ii) into (7.6) in their paper by a factor greater than $1/\{1 - \widetilde{F}\}_+$.

Another option in Theorem 1 is $T_A(\boldsymbol{u}) = \boldsymbol{\Sigma}_A^{-1} f_{q-1}(\boldsymbol{u})$, where $f_s(x) = \mathrm{sgn}(x)|x|^s$. For this option, $\boldsymbol{u}'\boldsymbol{\Sigma}_A T_A(\boldsymbol{u}) = \|\boldsymbol{u}\|_q^q = 1$ and $\boldsymbol{w} = T_A(\boldsymbol{u})$. Since $G_{A,\boldsymbol{u}} \le 2\lambda\|\boldsymbol{w}\|_1$ and $\|\boldsymbol{w}\|_1 \le \|\boldsymbol{\Sigma}_A^{-1}\|_{q/(q-1),1} = \|\boldsymbol{\Sigma}_A^{-1}\|_{\infty,q}$, separate maximization over $\boldsymbol{u}$ in (3.5) yields

$$\|\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda) - \boldsymbol{\beta}\|_q \le \max_{A,B}\frac{2\lambda\|\boldsymbol{\Sigma}_A^{-1}\|_{\infty,q}(1 + (k/\ell)^{q-1})^{1/q}}{\{1 - (k^{1-1/q}/\ell)\|\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\Sigma}_{A,B}\|_{\infty,q}\}_+}, \ 1 \le q \le \infty, \tag{3.9}$$

in the event $\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty \le \lambda$, where A and B are as in (3.5).

For $q = \infty$, this option provides the selection consistency of threshold Dantzig selectors and an oracle property of the Gauss-Dantzig selector of Candes and Tao [4],

$$\widehat{\boldsymbol{\beta}}_{GD} = \arg\min_{\boldsymbol{b}}\Big\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\| : |\widehat{\beta}_j| \le \lambda' \Rightarrow b_j = 0, \forall j, \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda)\Big\}. \tag{3.10}$$

For any threshold function $t(x;\lambda)$ satisfying $\{x : t(x;\lambda) = 0\} = \{x : |x| \le \lambda\}$ and $xt(x;\lambda) \ge 0$, define the threshold Dantzig selector as

$$\widehat{\boldsymbol{\beta}}_{TD} = t\big(\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda); \lambda'\big). \tag{3.11}$$

Examples include the hard threshold function $t(x;\lambda) = xI\{|x| > \lambda\}$ and the soft threshold function $t(x;\lambda) = \mathrm{sgn}(x)(|x| - \lambda)_+$. Define the oracle estimator

$$\widehat{\boldsymbol{\beta}}_{oracle} = \arg\min_{\boldsymbol{b}} \Big\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\| : \beta_j = 0 \Rightarrow b_j = 0, \forall j \Big\}. \tag{3.12}$$

Theorem 2. *Suppose (3.3) holds with $\|\boldsymbol{\beta}\|_0 = k$. Let $\widehat{\boldsymbol{\beta}}_{GD}$ and $\widehat{\boldsymbol{\beta}}_{TD}$ be as in (3.10) and (3.11) respectively with the penalty level $\lambda = \lambda_{univ}$ and a threshold level λ' satisfying*

$$\max_{A,B,\boldsymbol{v}} \frac{2\|\boldsymbol{\Sigma}_A^{-1}\boldsymbol{v}\|_1\{1 \vee (k/\ell)\}\lambda_{univ}}{(1 - (k/\ell)\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{\Sigma}_A^{-1}\boldsymbol{v}\|_1)_+} \le \lambda' < \min_{\beta_j \ne 0} |\beta_j|/2,$$

where λ_{univ} is as in (3.4) and the maximum is taken over $A \supset \{j : \beta_j \ne 0\}$ with $|A| = k + \ell$, $B \cap A = \emptyset$ with $|B| = \ell$, and $\boldsymbol{v} \in \mathbb{R}^A$ satisfying $\|\boldsymbol{v}\|_1 = 1$. Then,

$$P\Big\{ sgn(\widehat{\boldsymbol{\beta}}_{TD}) \ne sgn(\boldsymbol{\beta}) \text{ or } \widehat{\boldsymbol{\beta}}_{GD} \ne \widehat{\boldsymbol{\beta}}_{oracle} \Big\} \le 1/\sqrt{\pi \log(p)} \to 0.$$

Remark 1. *The basic requirement on $\boldsymbol{X}$ in Theorem 2 is $(k/\ell)\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{\Sigma}_A^{-1}\|_{1,1} < 1$ uniformly. Meanwhile, the strong irrepresentable condition for the selection consistency of the Lasso without post-thresholding is $\|\boldsymbol{\Sigma}_{J^c,J}\boldsymbol{\Sigma}_J^{-1}\|_{\infty,\infty} < 1$ uniformly.*

Inequality (3.9) and another version of (3.5) with the option $T_A(\boldsymbol{u}) = f_{q-1}(\boldsymbol{u})$ and $G_{A,\boldsymbol{u}} = \widetilde{C}\lambda\|\boldsymbol{w}_J\|_1 \le \widetilde{C}\lambda k^{1/q}\|\boldsymbol{w}\|_{q/(q-1)}$ are related to the error bounds in Zhang [28] for the Lasso, where the connection between ℓ_∞ bounds and variable selection has been explored. For $\|\boldsymbol{\beta}\|_0 = k$, error bounds of [28] can be written as

$$\big\|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda/t^*) - \boldsymbol{\beta}\big\|_q \le \frac{32(1 + F^*)G^*}{\widetilde{C}(1 - F^*)_+^2} \quad \text{for } z_\infty^* \le \lambda,\ 1 \le q \le \infty,\ \ell \ge k, \tag{3.13}$$

with $F^* = \max_{A,B,\boldsymbol{u}}(k^{1-1/q}/\|\boldsymbol{u}_J\|_1)F_{A,B,\boldsymbol{u}}$, $G^* = \widetilde{C}\lambda(k+\ell)^{1/q}\max_{A,\boldsymbol{u}}\|\boldsymbol{w}\|_{q/(q-1)}$ for $T_A(\boldsymbol{u}) = \boldsymbol{\Sigma}_A^{-1}f_{q-1}(\boldsymbol{u})$, $G^* = \widetilde{C}\lambda k^{1/q}\max_{A,\boldsymbol{u}}\|\boldsymbol{w}\|_{q/(q-1)}$ for $T_A(\boldsymbol{u}) = f_{q-1}(\boldsymbol{u})$,

$\widetilde{C} = (1 + 1/t^*)$ and $t^* = (1 - F^*)/\{4(1 + F^*)\}$, where $\boldsymbol{w} = T_A(\boldsymbol{u})/(\boldsymbol{u}'\boldsymbol{\Sigma}_A T_A(\boldsymbol{u}))_+$ is as in (3.5). In either cases, $G^* \geq \max_{A,\boldsymbol{u}} G_{A,\boldsymbol{u}}$ and $F^* \geq \max_{A,B,\boldsymbol{u}} F_{A,B,\boldsymbol{u}}$ due to applications of the Hölder inequality. It turns out that for the Lasso with $\alpha = t^*$, the right-hand side of (3.5) is smaller than 5/12 of the right-hand side of (3.13). For small k/ℓ, Zhang [28] pointed out the smaller order $k^{1/q}\lambda$ of G^* for $T_A(\boldsymbol{u}) = f_{q-1}(\boldsymbol{u})$ as an advantage for the Lasso, compared with the order $(k + \ell)^{1/q}\lambda$. The cost of this advantage is the square of $(1 - F^*)_+$ in the denominator of (3.13), compared with (3.7) and (3.9) for the Dantzig selector. Moreover, the error bound in (3.7) for the Dantzig selector is also of the order $k^{1/q}\lambda$ for $q \leq 2$ with much smaller constants, and the difference between $k^{1/q}$ and $(k + \ell)^{1/q}$ diminishes for large q as in Theorem 2. Thus, the advantage of the Lasso in this aspect has some limitations.

We have observed that Theorem 1 sharpens and unifies a number of existing error bounds for the Dantzig and Lasso estimators after applying the Hölder inequality and taking the worst scenarios in both the numerator and denominator in (3.5). We would like to mention that without additional applications of the Hölder inequality, (3.5) typically gives error bounds of a sharper form such as (3.9) involving the dimension-normalized $\|\cdot\|_{\infty,q}$ norm for matrices instead of the $\|\cdot\|_{q,q}$ norm. In addition, since the mappings $\boldsymbol{v} = T_A(\boldsymbol{u})$ are allowed to depend on $\{A, \boldsymbol{u}\}$, our approach actually allows to replace "$\max_{A,B,\boldsymbol{u}}$" in (3.5) with potentially much smaller "$\max_{A,\boldsymbol{u}} \inf_{\boldsymbol{v}} \max_B$". More general error bounds for $\|\boldsymbol{\beta}\|_0 > k$ are given in Subsection 3.3.4 as oracle inequalities. Although the error bounds in Theorem 1 for the Dantzig and Lasso estimators are of the same format, the Lasso bounds require a larger penalty level λ/α and larger $\{\widetilde{\alpha}, \xi, \widetilde{C}\}$. This theoretical advantage of the Dantzig selector reverses when $\|\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda)\|_1 \leq \|\boldsymbol{\beta}\|_1$ for $z^*_\infty \leq \lambda$ is replaced by $\|\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda)\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda)\|_1$ for $z^*_\infty > \lambda$ in our proofs. See Section 3.5.

3.3 Estimation with ℓ^q loss in ℓ^r balls

We state in four subsections our results on lower bounds for the minimax risk and loss in ℓ_r balls, upper bounds for the maxima of the estimation risk in ℓ_r balls for the Dantzig and Lasso estimators, upper bounds for the Lasso estimation loss, and oracle inequalities used to derive the upper bounds.

3.3.1 Lower bounds for the estimation risk and loss

Donoho and Johnstone [7] proved that for $0 < r < q$ and based on a p-vector $\widetilde{\boldsymbol{y}} \sim N(\boldsymbol{\beta}, \sigma_n^2 \boldsymbol{I}_p)$, the minimax ℓ_q risk in the ℓ_r ball $\Theta_{r,R} = \{\boldsymbol{v} : \|\boldsymbol{v}\|_r \le R\}$ is approximately

$$\inf_{\boldsymbol{\delta}} \sup_{\boldsymbol{\beta} \in \Theta_{r,R}} E_{\boldsymbol{\beta}} \big\| \boldsymbol{\delta}(\widetilde{\boldsymbol{y}}) - \boldsymbol{\beta} \big\|_q^q = (1 + o(1)) R^r \lambda_{mm}^{q-r}$$

and achieved within an infinitesimal fraction by threshold estimators at the threshold level λ_{mm}, provided that $\lambda_{mm}/\sigma_n \to \infty$ and $R^r/\lambda_{mm}^r \to \infty$. Here λ_{mm} is as in (3.4) with $\sigma_n = \sigma/\sqrt{n}$ and the infimum is taken over all Borel mappings $\boldsymbol{\delta}$ of proper dimensions. The following theorem extends their result to the estimation of regression coefficients in the linear model (3.3).

Theorem 3. *Let $P_{\boldsymbol{\beta}}$ be as in (3.3), $R > 0$, $q \ge r > 0$, $\sigma_n = \sigma/\sqrt{n}$ and λ_{mm} be as in (3.4). Suppose $R^r/\lambda_{mm}^r \to \infty$ and $\lambda_{mm}/\sigma_n \to \infty$. Then,*

$$\mathscr{R}(\Theta_{r,R}; \boldsymbol{X}) = \inf_{\boldsymbol{\delta}} \sup_{\|\boldsymbol{\beta}\|_r \le R} E_{\boldsymbol{\beta}} \|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_q^q \ge (1 + o(1)) R^r \lambda_{mm}^{q-r}, \quad (3.14)$$

and for all $0 \le \epsilon \le 1$,

$$\inf_{\boldsymbol{X}} \inf_{\boldsymbol{\delta}} \sup_{\|\boldsymbol{\beta}\|_r \le R} P_{\boldsymbol{\beta}} \big\{ \|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_q^q \ge (1 - \epsilon) R^r \lambda_{mm}^{q-r} \big\} \ge \frac{\epsilon + o(1)}{3^q}. \quad (3.15)$$

Remark 2. *By (3.4), $\lambda_{mm} \le \lambda_{univ}$ iff $R \ge \sigma_n$. For $\lambda_{univ} \asymp \lambda_{mm}$, theorems in Subsection 3.3.2 provide conditions under which both the Dantzig and Lasso estimators attain the minimax rate in ℓ_r balls. For smaller λ_{mm}, the rate minimaxity of the Lasso estimator is provided in Subsection 3.3.3.*

3.3.2 Upper bounds for the Dantzig and Lasso estimation risk

Here we present upper bounds for the minimax ℓ_q risk $E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q^q$ for the Dantzig and Lasso estimators. Our upper bounds match the lower bound (3.14) up to constant factors of the form

$$M_{q,\ell,*}(C_1,C_2)=\max_{A,B,\boldsymbol{v}}\left\{C_1\|\boldsymbol{v}\|_1/\ell^{1/q}+C_2(1+\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1^q/\ell)^{1/q}\right\} \tag{3.16}$$

where the maximum is taken over $|A|=|B|=\ell$, $A\cap B=\emptyset$ and $\|\boldsymbol{\Sigma}_A\boldsymbol{v}\|_{q/(q-1)}=1$, and

$$M_{q,d,\ell,\tau}(C_1,C_2)=\max_{A,B,\boldsymbol{v}}\frac{(1+\tau^q)^{1/q}(C_1\|\boldsymbol{v}\|_1/\ell^{1/q}+C_2)}{(1-\tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1/\ell^{1/q})_+}. \tag{3.17}$$

where the maximum is taken over $|A|=d$, $|B|=\ell$, $A\cap B=\emptyset$ and $\|\boldsymbol{\Sigma}_A\boldsymbol{v}\|_{q/(q-1)}=1$.

Theorem 4. *Let λ_{mm} and λ_{univ} be as in (3.4) and $q\ge 1$. Suppose $(\log p)/n=O(1)$ and $R^r/\lambda_{mm}^r\asymp d\le n\wedge p$ for some integer $d\to\infty$ satisfying $(\log d)/\log p\le c_0<1$. Let $0<\alpha_0<1$ and $\widehat{\boldsymbol{\beta}}$ be either the Dantzig or the Lasso estimator with $\lambda=\lambda_{univ}/\alpha_0$. Suppose $p^{1-(\alpha_1/\alpha_0)^2}(n^q+d^{q/r})/d\to 0$ for a certain $\alpha_1\in(\alpha_0,1)$. For given $\{k,\ell\}$, define*

$$C_1=(1+1/\alpha_0)(\ell\lambda_{mm}^r/R^r)^{1/q}/\sqrt{1-c_0},\quad C_2=2\{R^r/(\lambda_{mm}^r k)\}^{1/r-1/q}.$$

(i) Set $\ell=\lceil d/(1-\alpha_0)^{q/(q-1)}\rceil$ for the Lasso estimator and $\ell=d$ for the Dantzig selector. Set $k=d$ in the definition of C_2. Then,

$$\sup_{\|\boldsymbol{\beta}\|_r\le R}E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q^q\le(1+o(1))M_{q,\ell,*}^q\big(C_1,C_2\big)R^r\lambda_{mm}^{q-r},\quad r=1. \tag{3.18}$$

(ii) Let $\tau>0$ and $\{k_\alpha,\ell_\alpha\}$ be integers satisfying $0<k_\alpha^{1-1/q}(1+\alpha)\le\tau\ell_\alpha^{1-1/q}(1-\alpha)$ and $k_\alpha+\ell_\alpha=d$. For the Lasso estimator, let $(k,\ell)=(k_{\alpha_0},\ell_{\alpha_0})$ and assume $M_{q,d,\ell_\alpha,\tau}(C_1,C_2)=O(1)M_{q,d,\ell,\tau}(C_1,C_2)$, $\ell_\alpha=O(\ell)$ and $1/k_\alpha=O(1/k)$ for a

certain $\alpha \in (\alpha_1, 1)$. *For the Dantzig selector, let* $\{k, \ell\}$ *be integers satisfying* $k^{1-1/q} \le \tau \ell^{1-1/q}$ *and* $d = k + \ell$. *Then,*

$$\sup_{\|\boldsymbol{\beta}\|_r \le R} E_{\boldsymbol{\beta}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \le (1 + o(1)) M_{q,d,\ell,\tau}^q \big(C_1, C_2\big) R^r \lambda_{mm}^{q-r}, \quad 0 < r \le 1. \tag{3.19}$$

Remark 3. *The risk bounds in Theorem 4 match the minimax risk in (3.14) up to constant factors. The constant factors are particularly simple when* $R^r/\lambda_{mm}^r = k$, *where* $C_1 = (1 + 1/\alpha_0)(\ell/k)^{1/q}/\sqrt{1 - c_0}$ *and* $C_2 = 2$.

Remark 4. *Since* $\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1$ *is increasing in* $|B|$ *and* $\ell \le d$, *it follows from (3.17) that*

$$M_{q,d,\ell,\tau}(C_1, C_2) \le (d/\ell)^{1/q} M_{q,d,d,\tau(d/\ell)^{1/q}}(C_1, C_2),$$

so that we may pick a suitable τ *in Theorem 4 (ii) to control the denominator of (3.17).*

Remark 5. *If* $(n + d^{1/r})/p = O(1)$, *then* $p^{1-(\alpha_1/\alpha_0)^2}(n^q + d^{q/r})/d \to 0$ *for a certain fixed* $\alpha_1 \in (\alpha_0, 1)$ *when* $\alpha_0 \in (0, 1/\sqrt{q+1})$ *is fixed. The proof of Theorem 4 (ii) actually provides slightly stronger results where the set* A *in (3.17) is restricted to contain the indices of the* $k = d - \ell$ *largest* $|\beta_j|$ *for given* $\boldsymbol{\beta}$. *The condition* $(\log d)/\log p \le c_0 < 1$ *fails and* $\lambda_{mm}/\lambda_{univ} = o(1)$ *when* $\log p = (1 + o(1)) \log d$. *This difficult case will be considered in the next subsection.*

Theorem 4 differs from existing results by directly comparing the ℓ_q risk of estimators with the minimax risk, instead of finding upper bounds for the ℓ_q loss. The quantities (3.16) and (3.17) are best understood by comparisons with functions of (3.6) and

$$\eta_{q,d} = \max_A \|\boldsymbol{\Sigma}_A^{-1}\|_{\infty,q}/d^{1/q}, \ \kappa_{q,d,\ell} = \max_{A,B} \|\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\Sigma}_{A,B}\|_{\infty,q}/\ell^{1/q}, \tag{3.20}$$

$$\eta_{q,d}^* = \max_A \|\boldsymbol{\Sigma}_A^{-1}\|_{q,q}, \ \kappa_{q,d,\ell}^* = \max_{A,B} \|\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\Sigma}_{A,B}\|_{q,q}, \ \gamma_{q,d} = \max_A \|\boldsymbol{\Sigma}_A\|_{q,q}, \tag{3.21}$$

where the maxima are taken over $|A| = d$, $|B| = \ell$ and $A \cap B = \emptyset$. These quantities also facilitate comparisons between our and existing upper bounds on

the loss as in the derivation of Corollary 1. In such comparisons, the Hölder inequality and (3.6) give

$$\eta_{q,d} \le \eta^*_{q,d}, \quad \kappa_{q,d,\ell} \le \kappa^*_{q,d,\ell}, \quad \eta^*_{2,d} \le 1/(1-\delta^-_d), \quad \kappa^*_{2,d,\ell} \le \theta_{d,\ell}\eta^*_{2,d}. \tag{3.22}$$

The constant factors (3.16) and (3.17) are bounded from the above by functions of quantities in (3.20), (3.21) and (3.6) if we take the maxima with individual norms before arithmetic operations and apply the Hölder inequality as in (3.22). If maxima are taken over $\|\boldsymbol{\Sigma}_A \boldsymbol{v}\|_{q/(q-1)} = 1$ as in (3.16), $\max_{\boldsymbol{v}} \|\boldsymbol{M}\boldsymbol{v}\|_1 = \max_{\|\boldsymbol{u}\|_{q/(q-1)}=1} \|\boldsymbol{M}\boldsymbol{\Sigma}_A^{-1}\boldsymbol{u}\|_1 = \|\boldsymbol{\Sigma}_A^{-1}\boldsymbol{M}'\|_{\infty,q}$ for all matrices $\boldsymbol{M}$. Taking $\boldsymbol{M} = \boldsymbol{I}_\ell$ or $\boldsymbol{M} = \boldsymbol{\Sigma}_{B,A}$, we find

$$M_{q,\ell,*}(C_1, C_2) \le C_1\eta_{q,\ell} + C_2\big(1 + \kappa^q_{q,\ell,\ell}\big)^{1/q} \tag{3.23}$$

with $M_{2,\ell,*}(C_1, C_2) \le \{C_1 + C_2((1-\delta^-_\ell)^2 + \theta^2_{\ell,\ell})^{1/2}\}/(1-\delta^-_\ell)_+$, and

$$M_{q,d,\ell,\tau}(C_1, C_2) \le \frac{C_1\eta_{q,d}(d/\ell)^{1/q} + C_2}{(1+\tau^q)^{-1/q}(1-\tau\kappa_{q,d,\ell})_+} \tag{3.24}$$

with $M_{2,d,\ell,\tau}(C_1, C_2) \le \sqrt{1+\tau^2}\{C_1\sqrt{d/\ell} + C_2(1-\delta^-_d)\}/(1-\delta^-_d - \tau\theta_{d,\ell})_+$. The upper bounds for $q = 2$ follows from $\kappa^*_{2,d,\ell} \le \theta_{d,\ell}\eta^*_{2,d}$ and $\eta^*_{2,d} \le 1/(1-\delta^-_d)$. In view of (3.24) and Remark 4, Theorem 4 immediately implies the following theorem on rate minimaxity.

Theorem 5. *Suppose the conditions of Theorem 4 with fixed c_0 and α_0 in $(0,1)$. Suppose $\eta_{q,d} + \kappa_{q,d,d} = O(1)$. Then, for both the Dantzig and Lasso estimators at $\lambda = \lambda_{univ}/\alpha_0$,*

$$\sup_{\|\boldsymbol{\beta}\|_r \le R} E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^q_q = O(1)\inf_{\boldsymbol{\delta}} \sup_{\|\boldsymbol{\beta}\|_r \le R} E_{\boldsymbol{\beta}}\|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|^q_q, \ 0 < r \le 1 \le q.$$

Remark 6. *For $q = 2$, the conditions on (3.20) for the rate minimaxity in Theorem 5 hold when $\delta^-_d \le \delta^* < 1$ for a fixed δ^*. For $p \gg n$, random matrix theory can be applied to validate conditions on (3.16), (3.17), (3.6), (3.20) and (3.21) up to $k \asymp \ell \asymp d \asymp n/\log(p/n)$.*

3.3.3 Upper bounds for the Lasso estimation loss

The upper bounds for the estimation risk in Subsection 3.3.2 are obtained from an oracle inequality which also provide upper bounds for the tail probability of the estimation loss at penalty level $\lambda = \lambda_{univ}/\alpha_0$, with $0 < \alpha_0 < 1$ for the Lasso and $0 < \alpha_0 \le 1$ for the Dantzig selector. However, in applications and simulation studies, a penalty level $\lambda < \lambda_{univ}$ is often empirically the best choice. As we mentioned in Remark 2, $\lambda_{mm} < \lambda_{univ}$ iff $R > \sigma/\sqrt{n}$. For $\lambda_{mm}/\lambda_{univ} = o(1)$, performance bounds requiring penalty levels $\lambda \ge \lambda_{univ}$ do not match the lower bounds for the minimax rates in Theorem 3. For example, when $p = n \log n$, the order of $d \asymp R^r/\lambda_{mm}^r$ could be as large as $n/\log\log n$ for regularity conditions on $\boldsymbol{X}$ to hold, so that $\lambda_{mm}/\lambda_{univ} \to 0$ as $n \to \infty$. Theorem 6 below closes this gap by providing a minimax upper bound for the tail quantile of the ℓ_q loss for the Lasso estimator with $\lambda \asymp \lambda_{mm} = o(\lambda_{univ})$.

For $1 \le \ell \le d \le p$ and $\tau > 0$ define

$$N_{d,\ell,\tau}(C_1, C_2) = \max_{A,B,\boldsymbol{u}} \frac{C_1 + C_2 \boldsymbol{u}' \boldsymbol{\Sigma}_A \boldsymbol{u}}{(\boldsymbol{u}' \boldsymbol{\Sigma}_A \boldsymbol{u} - \tau \|\boldsymbol{\Sigma}_{B,A} \boldsymbol{u}\|_1/\sqrt{\ell})_+}, \tag{3.25}$$

where the maximum is taken over $|A| = d$, $|B| = \ell$, $A \cap B = \emptyset$ and $\|\boldsymbol{u}\| = 1$.

Theorem 6. *Let λ_{mm} and λ_{univ} be as in (3.4) and $\gamma_{2,\ell}$ be as in (3.21). Let $\{d, k, \ell\}$ be positive integers with $k + \ell = d$ and $k \asymp \ell \asymp d$, $\lambda = \min\big(\lambda_{univ}, (1 + \epsilon_0)\gamma_{2,\ell}^{1/2}\lambda_{mm}\big)/\alpha$ with $0 < \epsilon_0 \le \alpha < 1$. Let $0 < r \le 1 \le q \le 2$, $\tau = \sqrt{k/\ell}(1 + \epsilon_0)/(1 - \alpha)$ and $\widetilde{G} = (1 + \epsilon_0) + (1 + \alpha)/2$ for $\lambda = \lambda_{univ}/\alpha$, $\tau = \{(1 + \epsilon_0)^2 k/\ell + \alpha^2\}^{1/2}/(1-\alpha)$ and $\widetilde{G} = \tau(1-\alpha)\sqrt{\ell/k}+1/2$ for $\lambda < \lambda_{univ}/\alpha$, $C_* = (1+\tau^2)^{1-1/q}(1+ \tau\sqrt{\ell/k})^{2/q-1}$ and*

$$C_1 = C_* \widetilde{G}(\lambda/\lambda_{mm})(k\lambda_{mm}^r/R^r)^{1/q},\ C_2 = (5/2)C_*(k\lambda_{mm}^r/R^r)^{1/q-1/r}. \tag{3.26}$$

Suppose $n \wedge p \ge d \asymp R^r/\lambda_{mm}^r \to \infty$ and $\lambda_{mm} n^{1/2}/\sigma \to \infty$. Then,

$$\sup_{\|\boldsymbol{\beta}\|_r \le R} P_{\boldsymbol{\beta}}\Big\{ \|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_q^q \ge N_{d,\ell,\tau}^q\big(C_1, C_2\big) R^r \lambda_{mm}^{q-r} \Big\} \to 0.$$

Remark 7. *The upper bound in Theorem 6 matches the minimax lower bound in (3.15) up to a constant factor. The constant factor $N^q_{q,d,\ell,\tau}(C_1, C_2)$ is particularly simple when $R^r/\lambda^r_{mm} = k$, where $C_1 = C_*\widetilde{G}\lambda/\lambda_{mm}$ and $C_2 = (5/2)C_*$.*

Similar to (3.23) and (3.24), upper bounds for (3.25) can be obtained by bounding the ℓ_1 norm with the ℓ_2 norm and taking maxima before arithmetic operations. In fact,

$$N_{d,\ell,\tau}(C_1, C_2) \le \frac{C_1 + C_2(1-\delta_d^-)}{(1-\delta_d^- - \tau\theta_{d,\ell})_+}. \tag{3.27}$$

Theorem 7. *Let $1 \le q \le 2$. Suppose $n \wedge p \ge d \asymp R^r/\lambda^r_{mm} \to \infty$, $\lambda_{mm}n^{1/2}/\sigma \to \infty$, $\delta_d^- < 1$ and $\theta_{d,\ell} = O(1)$. Then, with the λ in Theorem 6, the Lasso estimator is rate minimax in the following sense:*

$$\begin{aligned} & \inf\Big[t : \sup_{\|\boldsymbol{\beta}\|_r \le R} P_{\boldsymbol{\beta}}\Big\{\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_q^q \ge t^q R^r \lambda_{mm}^{q-r}\Big\} \le \epsilon\Big] \\ = \;& O(1)\inf\Big[t : \inf_{\boldsymbol{\delta}} \sup_{\|\boldsymbol{\beta}\|_r \le R} P_{\boldsymbol{\beta}}\Big\{\|\boldsymbol{\delta}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}\|_q^q \ge t^q R^r \lambda_{mm}^{q-r}\Big\} \le \epsilon\Big], \quad \forall\ \epsilon > 0. \end{aligned}$$

3.3.4 Oracle inequalities

If one is allowed to approximate a target $\boldsymbol{\beta}^*$ with a vector with at most k nonzero entries, the best can be done under the ℓ_q losses is to pick the k largest elements of $\boldsymbol{\beta}^*$ in absolute value. The ℓ_q loss of this oracle approximation is

$$\rho_{q,k}(\boldsymbol{\beta}^*) = \sum_{j \notin J_k} |\beta_j^*|^q,\ \rho_k(\boldsymbol{\beta}^*) = \rho_{1,k}(\boldsymbol{\beta}^*),\ \text{where}\ \ J_k = \arg\max_{|S|=k} \sum_{j \in S} |\beta_j^*|. \tag{3.28}$$

Here we provide oracle inequalities which bounds the ℓ_q loss of the Dantzig and Lasso estimators in terms of $\rho_{s,k}(\boldsymbol{\beta}^*)$ and error measures on $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*$. The oracle inequalities make assertions about $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q$ in certain domain of $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}^*\}$ and thus do not require distributional assumptions about the error $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*$.

For $q > 0$ and $\ell \ge 1$, define

$$z^*_{q,\ell} = \max_{|A|=\ell} z^*_{q,A},\ z^*_{q,A} = \|\boldsymbol{X}'_A(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n\|_q/|A|^{1/q}, \quad z^*_\infty = z^*_{\infty,1}. \tag{3.29}$$

Since $z^*_{q,\ell}$ is the length normalized ℓ_q norm of the ℓ largest elements of $\{|\boldsymbol{x}'_j(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n|, j \le p\}$, $z^*_{q,\ell}$ is increasing in q and decreasing in ℓ, and $z^*_{q,\ell} \le z^*_\infty$ for all $\ell \ge 1$. We first deal with the case where the penalty level λ is no smaller than z^*_∞.

Theorem 8. *Let $q \in [1,\infty]$, $k \ge 0$, $1 \le \ell \le p-k$, $0 \le \alpha_* \le \alpha$, and $\{J_k, \rho_k(\boldsymbol{\beta}^*), z^*_{\infty,J_k}, z^*_\infty\}$ be as in (3.28) and (3.29). Let $\{\widehat{\boldsymbol{\beta}}, \widetilde{\alpha}, \xi\} = \{\widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda), 0, 1\}$ with $\alpha \le 1$ or $\{\widehat{\boldsymbol{\beta}}, \widetilde{\alpha}, \xi\} = \{\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda), \alpha, (1+\alpha_*)/(1-\alpha)\}$ with $\alpha < 1$. For $A \subset \{1,\ldots,p\}$, let $T_A(\boldsymbol{u})$ be mappings from $\mathbb{R}^A$ to $\mathbb{R}^A$ and define $\boldsymbol{w} = T_A(\boldsymbol{u})/(\boldsymbol{u}'\boldsymbol{\Sigma}_A T_A(\boldsymbol{u}))_+$ as a function of $\{A, \boldsymbol{u}\}$. Let $s > 0$ and $f_s(x) = \mathrm{sgn}(x)|x|^s$. Then, in the event $\{z^*_\infty \le \alpha\lambda, z^*_{\infty,J_k} \le \alpha_*\lambda\}$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \le \frac{G_{A,\boldsymbol{u}}\{1 + (\xi\|\boldsymbol{u}_{J_k}\|_1/\ell^{1-1/q})^q\}^{1/q}}{(1 - \xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)\big(1 + \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1^q/\ell\big)^{1/q}}{\ell^{1-1/q}(1-\widetilde{\alpha})(1 - \xi F_{A,B,\boldsymbol{u}})_+} \quad (3.30)$$

with $F_{A,B,\boldsymbol{u}} = \|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell$ and $G_{A,\boldsymbol{u}} = (1+\alpha)\lambda\|\boldsymbol{w}\|_1$, and

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{G_{A,\boldsymbol{u}}\|\boldsymbol{u}_{J_k}\|_1(1+\xi)}{(1 - \xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)(1 + F_{A,B,\boldsymbol{u}})}{(1-\widetilde{\alpha})(1 - \xi F_{A,B,\boldsymbol{u}})_+}, \quad (3.31)$$

for certain $A \supset J_k$ with $|A| = k + \ell$, $B \cap A = \emptyset$ with $|B| \le \ell$ and $\boldsymbol{u} \in \mathbb{R}^A$ with $\|\boldsymbol{u}\|_q = 1$. Moreover, (3.30) and (3.31) also hold with

$$G_{A,\boldsymbol{u}} = (1+\alpha)\lambda\|\boldsymbol{w}\|_1(1+\xi)\|\boldsymbol{u}_{J_k}\|_1/\|\boldsymbol{u}\|_1 + \Big\{\frac{2\rho_k(\boldsymbol{\beta}^*)(1+\alpha)\lambda\|\boldsymbol{w}\|_1}{(1-\widetilde{\alpha})\|\boldsymbol{u}\|_1}\Big\}^{1/2} \quad (3.32)$$

and for $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}(\lambda)$ with $T_A(\boldsymbol{u}) = f_s(\boldsymbol{u})$

$$G_{A,\boldsymbol{u}} = (1+\alpha_*)\lambda\|\boldsymbol{w}_{J_k}\|_1 + \big\{(1+\alpha)\lambda\rho_{s,k}(\boldsymbol{\beta}^*)\|\boldsymbol{w}\|_{q/s}\big\}^{1/(s+1)}. \quad (3.33)$$

Remark 8. *Let $\tau = \xi(k/\ell)^{1-1/q}$. Since $\|\boldsymbol{u}_{J_k}\|_1 \le k^{1-1/q}$, $\xi F_{A,B,\boldsymbol{u}} \le \tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell^{1/q}$. For $\tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell^{1/q} \le 1$, $\xi\|\boldsymbol{u}_{J_k}\|_1/\ell^{1-1/q} \le \tau$, $\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1^q/\ell \le 1/\tau^q$ and $\tau\ell^{1-1/q}(1-\widetilde{\alpha}) = k^{1-1/q}\xi(1-\widetilde{\alpha}) \ge k^{1-1/q}$. Thus, (3.30) and (3.31) implies that for all $1 \le s \le q$*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_s \le \frac{(1+\tau^q)^{(1-1/s)/(q-1)}k^{1/s-1/q}}{(1+\xi)^{(1/q-1/s)/(1-1/q)}} \max_{A,B,\boldsymbol{u}} \frac{\{G_{A,\boldsymbol{u}} + 2\rho_k(\boldsymbol{\beta}^*)/k^{1-1/q}\}}{(1 - \tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell^{1/q})_+}, \quad (3.34)$$

with an application of the Hölder inequality $\|\boldsymbol{h}\|_s \le \|\boldsymbol{h}\|_1^{(1/s-1/q)/(1-1/q)} \|\boldsymbol{h}\|_q^{(1-1/s)/(1-1/q)}$. *This is especially simple for* $\boldsymbol{v} = T_A(\boldsymbol{u}) = \boldsymbol{\Sigma}_A^{-1} f_{q-1}(\boldsymbol{u})$, *with* $\boldsymbol{w} = \boldsymbol{v}$ *as in (3.9) and the equivalence of the maximizations over* $\|\boldsymbol{u}\|_q = 1$ *and* $\|\boldsymbol{\Sigma}_A \boldsymbol{v}\|_{q/(q-1)} = 1$.

Our next theorem provides error bounds for the Lasso estimator under the weaker condition $z^*_{1,\ell} < \lambda$ instead of $z^*_\infty < \lambda$ in Theorem 8.

Theorem 9. *Let* $\{q, k, \ell, J_k, \rho_k(\boldsymbol{\beta}^*), T_A(\boldsymbol{u}), \boldsymbol{w}, f_s\}$ *be as in Theorem 8 and* $\{z^*_{s,\ell}, z^*_{s,J_k}\}$ *as in (3.29). Let* $q' = q/(q-1)$, $\tau = \xi(k/\ell)^{1/q'}$ *and* $\xi = \{(\alpha_*+1)^{q'} + \alpha_{q'}^{q'} \ell/k\}^{1/q'}/(1-\alpha)$ *with* $\alpha = \alpha_1 \in (0,1)$ *and positive* $\{\alpha_*, \alpha_{q'}\}$. *Then, in the event* $\{z^*_{q'\vee q, J_k} \le \alpha_* \lambda, z^*_{s,\ell} \le \alpha_s \lambda, s = 1, q, q'\}$,

$$\|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) - \boldsymbol{\beta}\|_q \le \frac{G_{A,\boldsymbol{u}}(1+\tau^q)^{1/q}}{(1-\xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)((\xi F_{A,B,\boldsymbol{u}})^q + \tau^q)^{1/q}}{\tau \ell^{1/q'}(1-\alpha)(1-\xi F_{A,B,\boldsymbol{u}})_+} \tag{3.35}$$

with $F_{A,B,\boldsymbol{u}} = k^{1/q'} \|\boldsymbol{\Sigma}_{B,A} \boldsymbol{w}\|_1/\ell$ *and* $G_{A,\boldsymbol{u}} = \|\boldsymbol{w}\|_{q'} \{(\alpha_*+1)^q k + (\alpha_q+1)^q \ell\}^{1/q} \lambda$, *and*

$$\|\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda) - \boldsymbol{\beta}\|_1 \le \frac{G_{A,\boldsymbol{u}}(\|\boldsymbol{u}_{J_k}\|_1 + \xi k^{1/q'})}{(1-\xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)\{k^{1/q'} + \|\boldsymbol{u}_{J_k}\|_1 F_{A,B,\boldsymbol{u}}\}}{k^{1/q'}(1-\alpha)(1-\xi F_{A,B,\boldsymbol{u}})_+} \tag{3.36}$$

for certain $A \supset J_k$ *with* $|A| = k + \ell$, $B \cap A = \emptyset$ *with* $|B| = \ell$, *and* $\boldsymbol{u} \in \mathbb{R}^A$ *with* $\|\boldsymbol{u}\|_q = 1$. *If* $T_A(\boldsymbol{u}) = f_s(\boldsymbol{u})$ *with* $s > 0$, *then (3.35) and (3.36) hold with*

$$G_{A,\boldsymbol{u}} = \|\boldsymbol{w}\|_{q'} \big\{(\alpha_*+1)^q k + \alpha_q^q \ell\big\}^{1/q} \lambda + \big\{\lambda \rho_{s,k}(\boldsymbol{\beta}^*) \|\boldsymbol{w}\|_{q/s}\big\}^{1/(s+1)}. \tag{3.37}$$

Remark 9. *Similar to Remark 8, inserting* $\|\boldsymbol{u}_{J_k}\|_1 \le k^{1/q'}$ *and* $\xi F_{A,B,\boldsymbol{u}} \le 1$ *in the numerators of (3.35) and (3.36) yields (3.34) as a simple version of Theorem 9 with the respective* ξ. *Consider the special case where* $\boldsymbol{\beta}^* = \boldsymbol{\beta}$ *and* $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. *Since* J_k *is fixed given* $\boldsymbol{\beta}$, *we typically have* $\alpha_* = o(\alpha_{q'})$, *which implies* $\tau \approx (k/\ell + \alpha_{q'}^{q'})^{1/q'}/(1-\alpha)$ *for* $k/\ell = O(1)$.

3.4 Proofs

We prove the lower bounds, the oracle inequalities, and then the upper bounds. Lemmas are stated and proved as needed.

3.4.1 Proofs of lower bounds

Let $P_{\mu,w}$ be a (prior) probability distribution under which (z_j, β_j) are iid vectors with

$$z_j|\beta_j \sim N(\beta_j, \sigma_n^2),\ P_{\mu,w}\{\beta_j = \mu\} = w = 1 - P_{\mu,w}\{\beta_j = 0\},$$

where $\mu = \lambda_{mm}(1-\epsilon)$ and $w = (1-\epsilon)(R/\lambda_{mm})^r/p$. Since $\widetilde{z}_j = \boldsymbol{x}_j'(\boldsymbol{y} - \sum_{k\neq j}\beta_k\boldsymbol{x}_k)/n$ is sufficient for β_j given $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}_k, k \neq j)$ and $\widetilde{z}_j$ and z_j are iid given $\boldsymbol{\beta}$. The minimum Bayes risk is bounded from below by

$$\begin{aligned}
& E_{\mu,w}\sum_{j=1}^{p}\min_t E_{\boldsymbol{\beta}}\Big[|t-\beta_j|^q\Big|\boldsymbol{X},\boldsymbol{y}\Big] \\
\geq\ & E_{\mu,w}\sum_{j=1}^{p}\min_t E_{\boldsymbol{\beta}}\Big[|t-\beta_j|^q\Big|\boldsymbol{X},\boldsymbol{y},\beta_k,k\neq j\Big] \\
=\ & E_{\mu,w}\sum_{j=1}^{p}\min_t E_{\boldsymbol{\beta}}\Big[|t-\beta_j|^q\Big|z_j\Big] \\
=\ & (1+o(1))R^r\lambda_{mm}^{q-r}
\end{aligned}$$

as $(R^r/\lambda_{mm}^r, \lambda_{mm}/\sigma_n) \to (\infty,\infty)$ and then $\epsilon \to 0$. The approximation in the last step above is given in Donoho and Johnstone [7]. Let

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} E_{\mu,w}E_{\boldsymbol{\beta}}\big[\|\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{y}) - \boldsymbol{\beta}\|_q^q\big|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\beta}\in\Theta_{r,R}\big].$$

Since the conditional Bayes risk of $\boldsymbol{\delta}^*$ is no greater than the minimax risk in $\Theta_{r,R}$,

$$\begin{aligned}
& (1+o(1))R^r\lambda_{mm}^{q-r} \\
\leq\ & E_{\mu,w}E_{\boldsymbol{\beta}}\big[\|\boldsymbol{\delta}^*-\boldsymbol{\beta}\|_q^q\big|\boldsymbol{\beta}\in\Theta_{r,R}\big] + E_{\mu,w}E_{\boldsymbol{\beta}}\|\boldsymbol{\delta}^*-\boldsymbol{\beta}\|_q^q I\{\boldsymbol{\beta}\notin\Theta_{r,R}\} \\
\leq\ & \mathscr{R}(\Theta_{r,R};\boldsymbol{X}) + 2^{(q-1)_+}E_{\mu,w}E_{\boldsymbol{\beta}}\big(\|\boldsymbol{\delta}^*\|_q^q + \|\boldsymbol{\beta}\|_q^q\big)I\{\boldsymbol{\beta}\notin\Theta_{r,R}\}. \qquad (3.38)
\end{aligned}$$

Since $E_{\mu,w}E_{\boldsymbol{\beta}}[\|\boldsymbol{\delta}^*-\boldsymbol{\beta}\|_q^q\big|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\beta}\in\Theta_{r,R}\big] \leq E_{\mu,w}[\|\boldsymbol{\beta}\|_q^q\big|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\beta}\in\Theta_{r,R}\big] \leq R^r\mu^{q-r}$, $\|\boldsymbol{\delta}^*\|_q^q \leq 2^{(q-1)_+}R^r\mu^{q-r}$ almost surely. Let $N = \#\{j:\beta_j=\mu\}$. We have $\|\boldsymbol{\beta}\|_q^q = \mu^q N$ and $\boldsymbol{\beta}\notin\Theta_{r,R}$ if and only if $N > R^r/\mu^r = wp/(1-\epsilon)^{1+r}$ under $P_{\mu,w}$. Thus, since $N \sim \text{Binomial}(p,w)$ with $pw\to\infty$,

$$E_{\mu,w}E_{\boldsymbol{\beta}}\big(\|\boldsymbol{\delta}^*\|_q^q + \|\boldsymbol{\beta}\|_q^q\big)I\{\boldsymbol{\beta}\notin\Theta_{r,R}\}$$

$$\leq \quad 2^{(q-1)_+}R^r\mu^{q-r}P\{N > wp/(1-\epsilon)^{1+r}\} + \mu^q E_{\mu,w}NI\{N > wp/(1-\epsilon)^{1+r}\}$$
$$= \quad o(1)R^r\lambda_{mm}^{q-r}. \tag{3.39}$$

The combination of (3.38) and (3.39) gives (3.14).

Now consider the loss $L(\boldsymbol{\delta},\boldsymbol{\beta}) = I\{\|\boldsymbol{\delta}-\boldsymbol{\beta}\|_q > c(R/\lambda_{mm})^{r/q}\lambda_{mm}\}$ in (3.15). Define

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{y})I\Big\{\|\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{y})\|_q \leq (1+c)(R/\lambda_{mm})^{r/q}\lambda_{mm}\Big\}.$$

Since $\|\boldsymbol{\beta}\|_q^q \sim N\mu^q$ and $\|\widehat{\boldsymbol{\beta}}\|_\infty \leq \mu$ under $P_{\mu,w}$, in the event $\|\boldsymbol{\beta}\|_r \leq R$,

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q^q \quad &\leq \quad c^qR^r\lambda_{mm}^{q-r}I\Big\{\|\boldsymbol{\delta}-\boldsymbol{\beta}\|_q \leq c(R/\lambda_{mm})^{r/q}\lambda_{mm}\Big\}\\ &\quad +\Big(\|\boldsymbol{\beta}\|_q + (1+c)(R/\lambda_{mm})^{r/q}\lambda_{mm}\Big)^q I\Big\{\|\boldsymbol{\delta}-\boldsymbol{\beta}\|_q > c(R/\lambda_{mm})^{r/q}\lambda_{mm}\Big\}\end{aligned}$$

and $\|\boldsymbol{\beta}\|_q^q \leq R^r\mu^{q-r} \leq R^r\lambda_{mm}^{q-r}$. It follows that

$$\begin{aligned}E_{\mu,w}E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q^q \quad &\leq \quad c^qR^r\lambda_{mm}^{q-r} + (2+c)^qR^r\lambda_{mm}^{q-r}\max_{\|\boldsymbol{\beta}\|_r\leq R}E_{\boldsymbol{\beta}}L(\delta(\boldsymbol{X},\boldsymbol{y}),\boldsymbol{\beta})\\ &\quad +2^{q-1}E_{\mu,w}\Big(\mu^qN + (1+c)^qR^r\lambda_{mm}^{q-r}\Big)I\{\|\boldsymbol{\beta}\|_r > R\}.\end{aligned}$$

Since $E_{\mu,w}\Big(\mu^qN + (1+c)^qR^r\lambda_{mm}^{q-r}\Big)I\{\|\boldsymbol{\beta}\|_r > R\} = o(1)R^r\lambda_{mm}^{q-r}$,

$$\sup_{\|\boldsymbol{\beta}\|_r\leq R}E_{\boldsymbol{\beta}}L(\delta(\boldsymbol{X},\boldsymbol{y}),\boldsymbol{\beta}) \geq \frac{1-c^q+o(1)}{(2+c)^q}$$

Since the $o(1)$ is uniform in the choice of $\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{y})$, we find

$$\inf_{\boldsymbol{\delta}}\sup_{\|\boldsymbol{\beta}\|_r\leq R}P_{\boldsymbol{\beta}}\Big\{\|\boldsymbol{\delta}(\boldsymbol{X},\boldsymbol{y})-\boldsymbol{\beta}\|_q^q > (1-\epsilon)R^r\lambda_{mm}^{q-r}\Big\} \geq \frac{\epsilon+o(1)}{3^q},\ \forall 0<\epsilon<1.$$

This gives (3.15) and completes the proof of Theorem 3. □

3.4.2 Proofs of oracle inequalities

As mentioned at the beginning of Subsection 3.3.4, we consider the estimation of an arbitrary target $\boldsymbol{\beta}^*$ from data points $(\boldsymbol{X},\boldsymbol{y})$ without any distributional assumption on the error $\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*$. The following lemma, which provides upper bounds in our proofs for tails of various inner products, can be viewed as a variation of Lemma 3.1 of Candes and Tao [4].

Lemma 1. *Let $\boldsymbol{h} \in \mathbb{R}^p$, $J_k \subset \{1, \ldots, p\}$ with $|J_k| = k$, and A be the union of J_k and the indices of the ℓ largest $|h_j|$ with $j \notin J_k$, $1 \le \ell \le p - k$. Then, for any vector $\boldsymbol{w} \in \mathbb{R}^p$,*

$$\sum_{j \notin A} w_j h_j \le \|\boldsymbol{h}_{J_k^c}\|_1 \max\Big\{\|\boldsymbol{w}_B\|_1/\ell : B \cap A = \emptyset, |B| \le \ell\Big\}.$$

Proof. Let $B_1, \ldots, B_m$ form a partition of J_k^c with decreasing values of $|h_j|$ such that $B_1 = A \setminus J_k$, $|B_j| = \ell$ for $j < m$ and $|B_m| \le \ell$. Since $\|\boldsymbol{h}_{B_j}\|_\infty \le \|\boldsymbol{h}_{B_{j-1}}\|_1/\ell$,

$$\sum_{j \notin A} w_j h_j = \sum_{j=2}^{m} \boldsymbol{w}'_{B_j} \boldsymbol{h}_{B_j} \le \sum_{j=2}^{m} \|\boldsymbol{w}_{B_j}\|_1 \|\boldsymbol{h}_{B_j}\|_\infty \le \max_B \|\boldsymbol{w}_B\|_1 \sum_{j=2}^{k} \|\boldsymbol{h}_{B_{j-1}}\|_1/\ell.$$

The proof is complete, since $\sum_{j=2}^{k} \|\boldsymbol{h}_{B_{j-1}}\|_1 \le \|\boldsymbol{h}_{J_k^c}\|_1$. □

Proof of Theorem 8. Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ for both estimators. The negative gradient is

$$\boldsymbol{g} = \boldsymbol{X}'\big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big)/n = \boldsymbol{X}'\big(\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{X}\boldsymbol{h}\big)/n, \quad \text{with } \|\boldsymbol{g}\|_\infty \le \lambda,$$

where $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*$. Define $\zeta_S(\boldsymbol{a}) = \{\boldsymbol{a}'(\boldsymbol{X}'_S\widetilde{\boldsymbol{\varepsilon}}/n - \boldsymbol{g}_S)\}_+$ for $S \subset \{1, \ldots, p\}$, $\boldsymbol{v} = T_A(\boldsymbol{u})$,

$$A = \operatorname*{arg\,max}_{A: A \supset J_k, |A| = k+\ell} \sum_{j \in A \setminus J_k} |h_j|, \ \boldsymbol{u} = \frac{\boldsymbol{h}_A}{\|\boldsymbol{h}_A\|_q}, \ B = \operatorname*{arg\,max}_{B: B \cap A = \emptyset, |B| \le \ell} \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1. \quad (3.40)$$

Since $\boldsymbol{X}'_A\boldsymbol{X}\boldsymbol{h}/n = \boldsymbol{X}'_A\widetilde{\boldsymbol{\varepsilon}}/n - \boldsymbol{g}_A$, (3.40) and Lemma 1 give

$$\boldsymbol{v}'\boldsymbol{\Sigma}_A\boldsymbol{h}_A = \boldsymbol{v}'\boldsymbol{X}'_A(\boldsymbol{X}\boldsymbol{h} - \boldsymbol{X}_{A^c}\boldsymbol{h}_{A^c})/n \le \zeta_A(\boldsymbol{v}) + (\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1/\ell)\big\|\boldsymbol{h}_{J_k^c}\big\|_1. \quad (3.41)$$

Since $\boldsymbol{v}'\boldsymbol{\Sigma}_A\boldsymbol{h}_A = \boldsymbol{v}'\boldsymbol{\Sigma}_A\boldsymbol{u}\|\boldsymbol{h}_A\|_q$ and $\boldsymbol{w} = \boldsymbol{v}/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+$,

$$\|\boldsymbol{h}_A\|_q \le \zeta_A(\boldsymbol{w}) + (\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\big\|\boldsymbol{h}_{J_k^c}\big\|_1. \quad (3.42)$$

Let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}$. Since $\operatorname{sgn}(\widehat{\beta}_j)g_j = \lambda$ for $\widehat{\beta}_j \ne 0$, for $f_s(x) = \operatorname{sgn}(x)|x|^s$

$$f_s(h_j)(z_j - g_j) \le \{(|\beta_j^*|^s \wedge |h_j|^s)(|z_j| + \lambda)\} \wedge \{2|\beta_j^*|^s\lambda + |h_j|^s(|z_j| - \lambda)\} \quad (3.43)$$

for all $1 \le j \le p$, $|z_j| \le \lambda$ and $s > 0$. Since $z^*_\infty \le \alpha\lambda$ and $z^*_{\infty,J_k} \le \alpha_*\lambda$, this gives

$$\|\boldsymbol{X}\boldsymbol{h}\|^2/n = \boldsymbol{h}'(\boldsymbol{X}'\widetilde{\boldsymbol{\varepsilon}}/n - \boldsymbol{g}) \le \|\boldsymbol{h}_{J_k}\|_1(\alpha_*\lambda + \lambda) + \|\boldsymbol{h}_{J_k^c}\|_1(\alpha\lambda - \lambda) + 2\lambda\rho_k(\boldsymbol{\beta}^*).$$

Since $\|\boldsymbol{h}_{J_k}\|_1 = \|\boldsymbol{h}_A\|_q\|\boldsymbol{u}_{J_k}\|_1$ and $z^*_\infty \le \alpha\lambda < \lambda$, the above inequality implies

$$\begin{aligned} \|\boldsymbol{h}_{J_k^c}\|_1 &\le 2\rho_k(\boldsymbol{\beta}^*)/(1-\alpha) + \|\boldsymbol{h}_{J_k}\|_1(1+\alpha_*)/(1-\alpha) \\ &= 2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha}) + \xi\|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{h}_A\|_q. \end{aligned} \tag{3.44}$$

The combination of (3.42) and (3.44) yields

$$(1 - \xi F_{A,B,\boldsymbol{u}})\|\boldsymbol{h}_A\|_q \le \zeta_A(\boldsymbol{w}) + (\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha})$$

with the factor $\xi F_{A,B,\boldsymbol{u}} = \xi\|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell$, and

$$(1 - \xi F_{A,B,\boldsymbol{u}})\|\boldsymbol{h}_{J_k^c}\|_1 \le 2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha}) + \xi\|\boldsymbol{u}_{J_k}\|_1\zeta_A(\boldsymbol{w}).$$

Since $\|\boldsymbol{h}\|_q^q \le \|\boldsymbol{h}_A\|_q^q + \big(\|\boldsymbol{h}_{J_k^c}\|_1/\ell^{1-1/q}\big)^q$ by Lemma 1, these inequalities imply

$$\|\boldsymbol{h}\|_q \le \frac{\zeta_A(\boldsymbol{w})(1 + (\xi\|\boldsymbol{u}_{J_k}\|_1/\ell^{1-1/q})^q)^{1/q}}{(1 - \xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)\big(1 + \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1^q/\ell\big)^{1/q}}{\ell^{1-1/q}(1-\widetilde{\alpha})(1 - \xi F_{A,B,\boldsymbol{u}})_+}. \tag{3.45}$$

Moreover, since $\|\boldsymbol{h}\|_1 = \|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{h}_A\|_q + \|\boldsymbol{h}_{J_k^c}\|_1$, they also imply

$$\begin{aligned} \|\boldsymbol{h}\|_1 &\le \frac{\zeta_A(\boldsymbol{w})\|\boldsymbol{u}_{J_k}\|_1(1+\xi)}{(1 - \xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)\{1 + \|\boldsymbol{u}_{J_k}\|_1(\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\}}{(1-\widetilde{\alpha})(1 - \xi F_{A,B,\boldsymbol{u}})_+}. \\ &= \frac{\zeta_A(\boldsymbol{w})\|\boldsymbol{u}_{J_k}\|_1(1+\xi)}{(1 - \xi F_{A,B,\boldsymbol{u}})_+} + \frac{2\rho_k(\boldsymbol{\beta}^*)(\xi F_{A,B,\boldsymbol{u}} + \xi)}{\xi(1-\widetilde{\alpha})(1 - \xi F_{A,B,\boldsymbol{u}})_+}. \end{aligned} \tag{3.46}$$

Thus, (3.30) and (3.31) hold if $\zeta_A(\boldsymbol{w})$ can be replaced by $G_{A,\boldsymbol{u}}$ in (3.42).

For $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Dantzig}$, $z^*_\infty \le \lambda$ implies $\|\widehat{\boldsymbol{\beta}}\|_1 \le \|\boldsymbol{\beta}^*\|_1$, so that $\|\boldsymbol{h}_{J_k^c}\|_1 \le \|\boldsymbol{\beta}^*_{J_k^c}\|_1 + \|\widehat{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}_{J_k}\|_1 \le \|\boldsymbol{\beta}^*_{J_k^c}\|_1 + \|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}_{J_k}\|_1 \le 2\|\boldsymbol{\beta}^*_{J_k^c}\|_1 + \|\boldsymbol{h}_{J_k}\|_1$. Thus,

$$\|\boldsymbol{h}_{J_k^c}\|_1 \le 2\rho_k(\boldsymbol{\beta}^*) + \|\boldsymbol{h}_{J_k}\|_1 = 2\rho_k(\boldsymbol{\beta}^*) + \|\boldsymbol{u}_{J_k}\|_1\|\boldsymbol{h}_A\|_q. \tag{3.47}$$

This effectively drops $\{\alpha, \alpha_*\}$ from (3.44), or replaces $\{\widetilde{\alpha}, \xi\} = \{\alpha, (1+\alpha_*)/(1-\alpha)\}$ with $\{\widetilde{\alpha}, \xi\} = \{0, 1\}$. Thus, (3.45) and (3.46) hold with $\{\widetilde{\alpha}, \xi\} = \{0, 1\}$. Again, (3.30) and (3.31) hold if $\zeta_A(\boldsymbol{w})$ can be replaced by $G_{A,\boldsymbol{u}}$ in (3.42).

It remains to prove that $\zeta_A(\boldsymbol{w})$ can be replaced by $G_{A,\boldsymbol{u}}$ in (3.42) under respective conditions. Since $\zeta_A(\boldsymbol{w}) = \{\boldsymbol{w}'(\boldsymbol{X}'_A\widetilde{\boldsymbol{\varepsilon}}/n - \boldsymbol{g}_A)\}_+ \le \|\boldsymbol{w}\|_1(z^*_\infty + \lambda)$, $G_{A,\boldsymbol{u}} = \|\boldsymbol{w}\|_1(1+\alpha)\lambda$ is always allowed.

For the Lasso estimator with $\boldsymbol{v} = T_A(\boldsymbol{u}) = f_s(\boldsymbol{u})$, (3.43) and the condition $z^*_{\infty,J_k} \le \alpha_*\lambda$ yield $\zeta_A(f_s(\boldsymbol{h}_A)) \le \lambda\{(1+\alpha_*)\|f_s(\boldsymbol{h}_{J_k})\|_1 + (1+\alpha)\rho_{s,k}(\boldsymbol{\beta}^*)\}$. Thus, as in (3.41),

$$\begin{aligned}
\|\boldsymbol{h}_A\|_q^{s+1}\boldsymbol{v}'\boldsymbol{\Sigma}_A\boldsymbol{u} &= f_s(\boldsymbol{h}_A)'\boldsymbol{\Sigma}_A\boldsymbol{h}_A \\
&\le \zeta_A(f_s(\boldsymbol{h}_A)) + (\|\boldsymbol{\Sigma}_{B,A}f_s(\boldsymbol{h}_A)\|_1/\ell)\|\boldsymbol{h}_{J_k^c}\|_1 \\
&\le \lambda\{(1+\alpha_*)\|\boldsymbol{h}_A\|_q^s\|\boldsymbol{v}_{J_k}\|_1 + (1+\alpha)\rho_{s,k}(\boldsymbol{\beta}^*)\} + \|\boldsymbol{h}_A\|_q^s(\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1/\ell)\|\boldsymbol{h}_{J_k^c}\|_1.
\end{aligned}$$

Since $ax^{s+1} \le bx^s + c$ implies $x \le b/a + (c/a)^{1/(s+1)}$ for positive $\{a,b,c\}$,

$$\|\boldsymbol{h}_A\|_q \le \frac{(1+\alpha_*)\lambda\|\boldsymbol{v}_{J_k}\|_1}{(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+} + \Big(\frac{(1+\alpha)\lambda\rho_{s,k}(\boldsymbol{\beta}^*)}{(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{v})_+}\Big)^{1/(s+1)} + (\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\|\boldsymbol{h}_{J_k^c}\|_1.$$

Thus, since $\|\boldsymbol{w}\|_{q/s} = 1/(\boldsymbol{u}\boldsymbol{\Sigma}_A\boldsymbol{v})_+$, $\zeta_A(\boldsymbol{w})$ can be replaced by (3.33) in (3.42).

In general, the combination of (3.44) and (3.47) imply

$$\|\boldsymbol{h}_A\|_1 \le (1+\xi)\|\boldsymbol{h}_{J_k}\|_1 + 2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha}) = (1+\xi)\|\boldsymbol{h}_A\|_q\|\boldsymbol{u}_{J_k}\|_1 + 2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha}).$$

Since $\|\boldsymbol{u}\|_1\|\boldsymbol{h}_A\|_q = \|\boldsymbol{h}_A\|_1$ and (3.42) holds with $\zeta_A(\boldsymbol{w}) \le \|\boldsymbol{w}\|_1(1+\alpha)\lambda$,

$$\begin{aligned}
\|\boldsymbol{u}\|_1\|\boldsymbol{h}_A\|_q^2 &\le \|\boldsymbol{h}_A\|_1\{\|\boldsymbol{w}\|_1(1+\alpha)\lambda + (\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\|\boldsymbol{h}_{J_k^c}\|_1\} \\
&\le \{(1+\xi)\|\boldsymbol{h}_A\|_q\|\boldsymbol{u}_{J_k}\|_1 + 2\rho_k(\boldsymbol{\beta}^*)/(1-\widetilde{\alpha})\}\|\boldsymbol{w}\|_1(1+\alpha)\lambda \\
&\quad + \|\boldsymbol{h}_A\|_q\|\boldsymbol{u}\|_1(\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\|\boldsymbol{h}_{J_k^c}\|_1.
\end{aligned}$$

Thus, it follows from the argument for the Lasso with $T_A(\boldsymbol{u}) = f_s(\boldsymbol{u})$ for $s = 1$ that $\zeta_A(\boldsymbol{w})$ can be replaced by (3.32) in (3.42). This completes the proof. □

Proof of Theorem 9. We use the notation of the proof of Theorem 8. Since (3.42) still holds, we need a version of (3.44) and an upper bound for $\zeta_A(\boldsymbol{w})$.

Let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}(\lambda)$. Set $\zeta_{A,J_k}(\boldsymbol{a}) = \boldsymbol{a}'_A\boldsymbol{X}'_A\widetilde{\boldsymbol{\varepsilon}}/n - \boldsymbol{a}'_{J_k}\boldsymbol{g}_{J_k}$. Since $\|\boldsymbol{u}\|_q = 1$,

$$\zeta_{A,J_k}(\boldsymbol{u}) \le \|\boldsymbol{u}_{J_k}\|_q(z^*_{q',J_k} + \lambda)k^{1/q'} + \|\boldsymbol{u}_{A\setminus J_k}\|_q z^*_{q',\ell}\ell^{1/q'}$$

$$\leq \{(z^*_{q',J_k}+\lambda)^{q'}(k/\ell)+(z^*_{q',\ell})^{q'}\}^{1/q'}\ell^{1/q'} \leq \xi k^{1/q'}(1-\alpha)\lambda, \quad (3.48)$$

due to $\xi = \{(\alpha_*+1)^{q'}+\alpha_{q'}^{q'}\ell/k\}^{1/q'}/(1-\alpha)$. It follows from Lemma 1 and (3.29) that $(\boldsymbol{X}'_{A^c}\widetilde{\boldsymbol{\varepsilon}}/n)'\boldsymbol{h}_{A^c} \leq z^*_{1,\ell}\|\boldsymbol{h}_{J_k^c}\|_1 \leq \alpha\lambda\|\boldsymbol{h}_{J_k^c}\|_1$. Thus, by (3.43)

$$\begin{aligned}
\|\boldsymbol{X}\boldsymbol{h}\|^2/n &= \boldsymbol{h}'(\boldsymbol{X}'\widetilde{\boldsymbol{\varepsilon}}/n-\boldsymbol{g}) \\
&= \|\boldsymbol{h}_A\|_q\zeta_{A,J_k}(\boldsymbol{u})+(\boldsymbol{X}'_{A^c}\widetilde{\boldsymbol{\varepsilon}}/n)'\boldsymbol{h}_{A^c}-\boldsymbol{h}'_{J_k^c}\boldsymbol{g}_{J_k^c} \\
&\leq \|\boldsymbol{h}_A\|_q\xi k^{1/q'}\lambda(1-\alpha)+\alpha\|\boldsymbol{h}_{J_k^c}\|_1\lambda-\|\boldsymbol{h}_{J_k^c}\|_1\lambda+2\rho_k(\boldsymbol{\beta}^*)\lambda.
\end{aligned}$$

This gives as a version of (3.44) in the form

$$\|\boldsymbol{h}_{J_k^c}\|_1 \leq 2\rho_k(\boldsymbol{\beta}^*)/(1-\alpha)+\xi k^{1/q'}\|\boldsymbol{h}_A\|_q.$$

Replacing $\xi\|\boldsymbol{u}_{J_k}\|_1$ by $\xi k^{1/q'}$ in the derivation of (3.45) and (3.46), we find that

$$\|\boldsymbol{h}\|_q \leq \frac{\zeta_A(\boldsymbol{w})(1+\tau^q)^{1/q}}{(1-\xi F_{A,B,\boldsymbol{u}})_+}+\frac{2\rho_k(\boldsymbol{\beta}^*)\big(1+\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1^q/\ell\big)^{1/q}}{\ell^{1/q'}(1-\alpha)(1-\xi F_{A,B,\boldsymbol{u}})_+}.$$

with $F_{A,B,\boldsymbol{u}} = k^{1/q'}\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell$ and $\tau = \xi(k/\ell)^{1/q'}$, and

$$\|\boldsymbol{h}\|_1 \leq \frac{\zeta_A(\boldsymbol{w})(\|\boldsymbol{u}_{J_k}\|_1+\xi k^{1/q'})}{(1-\xi F_{A,B,\boldsymbol{u}})_+}+\frac{2\rho_k(\boldsymbol{\beta}^*)\{1+\|\boldsymbol{u}_{J_k}\|_1(\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{w}\|_1/\ell)\}}{(1-\alpha)(1-\xi F_{A,B,\boldsymbol{u}})_+}.$$

It remains to prove that $\zeta_A(\boldsymbol{w})$ can be replaced by $G_{A,\boldsymbol{u}}$. Similar to (3.48),

$$\zeta_A(\boldsymbol{w}) \leq \|\boldsymbol{w}\|_{q'}\{(z^*_{q,J_k}+\lambda)^q k+(z^*_{q,\ell}+\lambda)^q\ell\}^{1/q},$$

so that $G_{A,\boldsymbol{u}} = \|\boldsymbol{w}\|_{q'}\{(\alpha_*+1)^q k+(\alpha_q+1)^q\ell\}^{1/q}\lambda$ is always valid. For $T_A(\boldsymbol{u}) = f_s(\boldsymbol{u})$,

$$\zeta_A(f_s(\boldsymbol{h}_A)) \leq \|\boldsymbol{h}_A\|_q^s\|\boldsymbol{v}\|_{q'}\big\{(\alpha_*+1)^q k+\alpha_q^q\ell\big\}^{1/q}\lambda+\lambda\rho_{s,k}(\boldsymbol{\beta}^*),$$

so that (3.37) is allowed as in the proof of the validity of (3.33) in Theorem 8. □

3.4.3 Proofs of upper bounds

The oracle inequalities are used to prove upper bounds on $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_q$. Since the purpose here is to estimate $\boldsymbol{\beta}$, we set the target $\boldsymbol{\beta}^* = \boldsymbol{\beta}$.

Proof of Theorem 1. Since $\|\boldsymbol{\beta}\|_0 = k$, we are allowed to apply the oracle inequalities in Theorem 8 with $\{\boldsymbol{\beta}^*, \rho_k(\boldsymbol{\beta}^*), J_k\} = \{\boldsymbol{\beta}, 0, J\}$. Since $\|\boldsymbol{u}_J\|_1 \le k^{1-1/q}$, this gives

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \le \max_{A,B,\boldsymbol{u}} \frac{G_{A,\boldsymbol{u}}(1 + \xi^q (k/\ell)^{q-1})^{1/q}}{(1 - \xi F_{A,B,\boldsymbol{u}})_+}, \ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \le \max_{A,B,\boldsymbol{u}} \frac{G_{A,\boldsymbol{u}} \|\boldsymbol{u}_J\|_1 (1+\xi)}{(1 - \xi F_{A,B,\boldsymbol{u}})_+},$$

with $\widetilde{C} = (1 + 1/\alpha)$ in $G_{A,\boldsymbol{u}}$ for the penalty level λ/α for the Lasso. This implies (3.5) by the Hólder inequality as in Remark 8. Finally, consider the Lasso in the case where $\boldsymbol{v} = T_A(\boldsymbol{u})$ agrees in sign with $\boldsymbol{u}$. Since $\{w_j, v_j, u_j, h_j, \widehat{\beta}_j, g_j\}$ have the same sign for $j \notin J$, $w_j(\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n - g_j) \le |w_j|(z_\infty^* - \lambda/\alpha) < 0$. Thus, $\zeta_A(\boldsymbol{w}) \le \sum_{j \in J} w_j(\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n - g_j) \le \|\boldsymbol{w}_J\|_1(1 + 1/\alpha)\lambda$ for $z_\infty^* \le \lambda$ as in the proof of Theorem 8. This allows $G_{A,\boldsymbol{u}} = \widetilde{C}\lambda\|\boldsymbol{w}_J\|_1$. Part (ii) follows directly from $\boldsymbol{x}_j'\boldsymbol{\varepsilon}/n \sim N(0, \sigma^2/n)$ under $P_{\boldsymbol{\beta}}$. □

Proof of Theorem 2. This theorem is a direct consequence of Theorem 1 with $q = \infty$, since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty \le \lambda' < \min_{\beta_j \neq 0} |\beta_j|/2$ guarantees $\{j : |\widehat{\beta}_j| > \lambda'\} = \{j : \beta_j \neq 0\}$. □

Our proofs of risk bounds require the following lemma.

Lemma 2. *Let $\widehat{\boldsymbol{\beta}}$ be either the Dantzig or the Lasso estimator at penalty level λ. Suppose $\|\boldsymbol{\beta}\|_r \le R$ with $0 < r \vee 1 \le q$. For any event Ω_0 with $t_* = \sqrt{2\log(1/P_{\boldsymbol{\beta}}(\Omega_0))} \ge 1$,*

$$E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I_{\Omega_0} \le 2^{q-1} P_{\boldsymbol{\beta}}(\Omega_0)\Big\{\frac{\Gamma(2q+1)}{(t_*^2 n\lambda/\sigma^2)^q} + \Big(\frac{(t_* + \sqrt{n})^2}{n\lambda/\sigma^2} + 2p^{(1-1/r)_+}R\Big)^q\Big\} \quad (3.49)$$

In particular, if $(\log p)/n + \sigma^2/(n\lambda^2) + R^r/(n\lambda^r) + \lambda^r/R^r = O(1)$, then

$$E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I_{\Omega_0} = o(1) R^r \lambda^{q-r}, \quad (3.50)$$

provided that $P_{\boldsymbol{\beta}}(\Omega_0)(\lambda^r/R^r)\{(\sigma/\lambda)^{2q} + p^{q(1-1/r)_+}(R/\lambda)^q\} = o(1)$.

Remark 10. *Since the unit sphere $S^{n-1} \subset \mathbb{R}^n$ is covered by $(2/\epsilon + 1)^n$ ϵ-balls for all $\epsilon > 0$, a certain ϵ ball contains at least m unit vectors $\boldsymbol{x}_j/\|\boldsymbol{x}_j\|$ for $\epsilon =$*

$(\log(p/m))/(2n)$. *It follows that the set of design vectors* $\boldsymbol{x}_j$ *contains some highly correlated clusters when* $(\log p)/n \geq 2$. *Thus, the condition* $(\log p)/n = O(1)$ *is natural for the estimation of* $\boldsymbol{\beta}$.

Proof of Lemma 2. Let $\widehat{\boldsymbol{\beta}}$ be the Lasso estimator. Since $\widehat{\boldsymbol{\beta}}$ minimizes the penalized loss, $\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\varepsilon}\|^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1$, so that

$$\|\widehat{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\beta}\|_1 \leq \frac{\|\boldsymbol{\varepsilon}\|^2}{2n\lambda} + 2\|\boldsymbol{\beta}\|_1 \leq \frac{(\|\boldsymbol{\varepsilon}\|/\sigma - t_* - \sqrt{n})_+^2}{n\lambda/\sigma^2} + \frac{(t_* + \sqrt{n})^2}{n\lambda/\sigma^2} + 2p^{(1-1/r)_+}R.$$

Since $\|\boldsymbol{\varepsilon}/\sigma\|$ is a Lip(1) function of $\boldsymbol{\varepsilon}/\sigma \sim N(0, \boldsymbol{I}_n)$ and $E_{\boldsymbol{\beta}}\|\boldsymbol{\varepsilon}\|/\sigma \leq \sqrt{n}$, the Gaussian isoperimetric theorem gives $P_{\boldsymbol{\beta}}\{\|\boldsymbol{\varepsilon}\|/\sigma - \sqrt{n} > t\} \leq e^{-t^2/2}$, so that

$$\begin{aligned} E_{\boldsymbol{\beta}}(\|\boldsymbol{\varepsilon}\|/\sigma - t_* - \sqrt{n})_+^{2q} &\leq \int_0^\infty P_{\boldsymbol{\beta}}\{\|\boldsymbol{\varepsilon}\|/\sigma - t_* - \sqrt{n} > t\} dt^{2q} \\ &\leq \int_0^\infty e^{-t_*^2/2 - t_* t} dt^{2q} = P_{\boldsymbol{\beta}}(\Omega_0)\Gamma(2q+1)/t_*^{2q}. \end{aligned}$$

The above inequalities yield (3.49) due to $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \leq (\|\widehat{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\beta}\|_1)^q$ for $q \geq 1$.

It follows from (3.49) that

$$\frac{E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I_{\Omega_0}}{R^r \lambda^{q-r}} = O(\lambda^r/R^r) P_{\boldsymbol{\beta}}(\Omega_0)\Big\{O(1) + (t_*^2/n + 1)^q(\sigma/\lambda)^{2q} + p^{q(1-1/r)_+}R^q/\lambda^q\Big\}.$$

Since the right-hand side is of no greater order than $P_{\boldsymbol{\beta}}(\Omega_0)\{(t_*^2 + n)^q + p^q n^{q/r}\} = o(1)$ for $t_*^2/(n \vee \log p) \to \infty$, it suffices to consider the case $t_*^2/n = O(1)$. Hence, (3.50) holds under the specified conditions. The same conclusions hold for the Dantzig selector, since $\|\widehat{\boldsymbol{\beta}}_{Dantzig}\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{Lasso}\|_1$. □

Proof of Theorem 4. We first bound $\lambda_{univ}/\lambda_{mm}$ and the expected loss for large $z_\infty^* = \|\boldsymbol{X}'\boldsymbol{\varepsilon}/n\|_\infty$. Let $\sigma_n = \sigma/\sqrt{n}$. Since $R^r/\lambda_{mm}^r \asymp d$ and $\lambda_{mm}^2 = 2\sigma_n^2 \log(p\sigma_n^r/R^r)$,

$$\begin{aligned} 2\sigma_n^2 \log(p/d) &= 2\sigma_n^2\Big\{\log(p\lambda_{mm}^r/R^r) + O(1)\Big\} \qquad (3.51) \\ &= \lambda_{mm}^2\Big[1 + 2(\sigma_n/\lambda_{mm})^2\Big\{\log(\lambda_{mm}^r/\sigma_n^r) + O(1)\Big\}\Big] \approx \lambda_{mm}^2. \end{aligned}$$

Thus, since $(\log d)/\log p \leq c_0 < 1$, $(1-c_0)\lambda_{univ}^2 = (1-c_0)2\sigma_n^2 \log p \leq 2\sigma_n^2 \log(p/d) \approx \lambda_{mm}^2$.

Let $\Omega_0 = \{z^*_\infty/\lambda > \alpha\}$ with any $\alpha \in (\alpha_1, 1)$. Since z^*_∞ is the maximum of p variables from $N(0, \sigma_n^2)$, $P_{\boldsymbol{\beta}}\{\Omega_0\} \le p\exp(-n(\alpha\lambda)^2/(2\sigma^2)) \le p^{1-(\alpha_1/\alpha_0)^2}$ for large n. Thus, due to $\lambda^2/\sigma_n^2 \asymp \log p$ and $n \ge d \asymp R^r/\lambda_{mm}^r \asymp R^r/\lambda^r \to \infty$, we have

$$P_{\boldsymbol{\beta}}(\Omega_0)(\lambda^r/R^r)\{(\sigma/\lambda)^{2q} + (R/\lambda)^q\} = O(1)p^{1-(\alpha_1/\alpha_0)^2}(n^q/d + d^{q/r-1}) = o(1).$$

Since $0 < r \le 1$, Lemma 2 gives $E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I\{z^*_\infty/\lambda > \alpha\} = o(R^r\lambda_{mm}^{q-r})$.

(i) Let $k = 0$, $\ell = \lceil d/(1-\alpha_0)^{q/(q-1)}\rceil$ and $\boldsymbol{v} = T_A(\boldsymbol{u}) = \boldsymbol{\Sigma}_A^{-1} f_{q-1}(\boldsymbol{u})$ in Theorem 8, so that $\|\boldsymbol{u}_{J_k}\|_1 = F_{A,B,\boldsymbol{u}} = 0$ and $\boldsymbol{w} = \boldsymbol{v}$ as in (3.9). For $z^*_\infty \le \alpha\lambda < \lambda$, (3.30) implies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \le \max_{A,B,\boldsymbol{v}} \Big\{(1+\alpha)\lambda\|\boldsymbol{v}\|_1 + 2Rd^{1/q-1}(1 + \|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1^q/\ell)^{1/q}\Big\} \tag{3.52}$$

for both the Dantzig and Lasso estimators. Thus, since $P_{\boldsymbol{\beta}}\{z^*_\infty > \lambda_{univ} = \alpha_0\lambda\} \to 0$,

$$\begin{aligned} & E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I\{z^*_\infty/\lambda \le \alpha\} \\ \le\ & (1+o(1))M^q_{q,\ell,*}\big((1+\alpha_0)(\ell\lambda_{mm}/R)^{1/q}\lambda/\lambda_{mm}, 2\{R/(d\lambda_{mm})\}^{1-1/q}\big)R\lambda_{mm}^{q-1} \end{aligned}$$

for both the estimators with the $M_{q,\ell,*}(x,y)$ in (3.16). This proves (3.18), due to the proven $\sqrt{1-c_0}\lambda_{univ} \le (1+o(1))\lambda_{mm}$ and $E_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q I\{z^*_\infty/\lambda > \alpha\} = o(R^r\lambda_{mm}^{q-r})$.

(ii) Since $k_\alpha^{1-1/q}(1+\alpha) \le \tau\ell_\alpha^{1-1/q}(1-\alpha)$, (3.34) with $s = q$ yields

$$\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_q^q \le (1+\tau^q)\max_{A,B,\boldsymbol{v}}\Big[\frac{(1+\alpha)\lambda\|\boldsymbol{v}\|_1 + 2\rho_{k_\alpha}(\boldsymbol{\beta}){k_\alpha}^{1/q-1}}{(1-\tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{v}\|_1/\ell_\alpha^{1/q})_+}\Big]^q$$

in the event $z^*_\infty \le \alpha\lambda$. For $0 < r < 1$, the $(k_\alpha+1)$-th largest $|\beta_j|$ is no greater than $R/(k_\alpha+1)^{1/r}$, so that $\rho_{k_\alpha}(\boldsymbol{\beta}) \le R^r\{R/(k_\alpha+1)^{1/r}\}^{1-r} \le R{k_\alpha}^{1-1/r}$. It follows that

$$\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}\|_q^q \le M^q_{q,d,\ell_\alpha,\tau}(C_{1,\alpha}, C_{2,\alpha})R^r\lambda_{mm}^{q-r} \tag{3.53}$$

with $C_{1,\alpha} = (1+\alpha)\lambda\ell_\alpha^{1/q}\lambda_{mm}^{-1}(\lambda_{mm}^r/R^r)^{1/q}$, $C_{2,\alpha} = 2R{k_\alpha}^{1/q-1/r}\lambda_{mm}^{-1}(\lambda_{mm}^r/R^r)^{1/q}$ and the $M_{q,k,\ell,\tau}(x,y)$ in (3.17). Moreover, since $M_{q,d,\ell_\alpha,\tau}(C_1,C_2)/M_{q,d,\ell,\tau}(C_1,C_2)$

and $C_{j,\alpha}/C_j$ are all bounded and we have already dealt with the case $z^*_\infty/\lambda > \alpha$, (3.53) implies (3.19).

Similarly, for the Dantzig selector with $k+\ell = d$ and $k^{1-1/q} \le \tau \ell^{1-1/q}$, (3.34) yields $M^q_{q,d,\ell,\tau}(C_1,C_2)R^r\lambda^{q-r}_{mm}$ as an upper bound for $\|\widehat{\boldsymbol{\beta}}_{Dantzig} - \boldsymbol{\beta}\|^q_q$. □

The proof of Theorem 6 requires the following lemma.

Lemma 3. *Let $\widetilde{p}_\ell$ be the positive number satisfying $2\log\widetilde{p}_\ell - 1 - \log(2\log\widetilde{p}_\ell) = (2/\ell)\log\binom{p}{\ell}$. Suppose $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ under probability P. Then,*

$$P\Big\{\max_{|A|=\ell}\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\| \ge \sigma\sqrt{2\ell\log\widetilde{p}_\ell}\Big\} \le \frac{1}{2\sqrt{\log\widetilde{p}_\ell}} \le \frac{1}{\sqrt{2}},$$

where $\boldsymbol{P}_A = \boldsymbol{X}_A(\boldsymbol{X}'_A\boldsymbol{X}_A)^{-1}\boldsymbol{X}'_A$ is the projection to the linear span of $\{\boldsymbol{x}_j, j\in A\}$.

Proof. Since $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\sigma^2\boldsymbol{I}_n)$, $\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2/\sigma^2 \sim \chi^2_\ell$ variables. Let $x>0$. Since $\chi^2_\ell/(1+x)$ has a gamma distribution, change of variable $t \to t^2$ and some algebra give

$$P\Big\{\frac{\chi^2_\ell}{1+x} \ge \ell\Big\} = \frac{e^{-\ell(1+x)/2}(1+x)^{\ell/2}}{\Gamma(\ell/2)2^{\ell/2}}\int_{\sqrt{\ell}}^\infty 2\exp\Big\{-(1+x)(t^2-\ell)/2 + (\ell-1)\log t\Big\}dt$$

Since the derivatives of $f(t) = (1+x)(t^2-\ell)/2 - (\ell-1)\log t$ satisfy $\{(\partial f)/(\partial t)\}(\sqrt{\ell}) \ge 0$ and $\{(\partial^2 f)/(\partial t)^2\}(t) \ge (1+x)$, the Stirling formula gives

$$\begin{aligned} P\Big\{\frac{\chi^2_\ell}{1+x} \ge \ell\Big\} &\le \frac{e^{-\ell(1+x)/2}(1+x)^{\ell/2}}{\Gamma(\ell/2)2^{\ell/2}}e^{-f(\sqrt{\ell})}\int_{\sqrt{\ell}}^\infty 2\exp\Big\{-(1+x)(t-\sqrt{\ell})^2/2\Big\}dt \\ &\le \frac{e^{-\ell(1+x)/2}(1+x)^{\ell/2}}{(\ell/2)^{\ell/2-1/2}e^{-\ell/2}\sqrt{2\pi}2^{\ell/2}}\ell^{(\ell-1)/2}\Big(\frac{2\pi}{1+x}\Big)^{1/2} \\ &= 2^{-1/2}e^{-x\ell/2}(1+x)^{(\ell-1)/2}. \end{aligned}$$

Setting $x = 2\log\widetilde{p}_\ell - 1$, we have $(x-\log(1+x))(\ell/2) = \log\binom{p}{\ell}$, so that

$$P\Big\{\max_{|A|=\ell}\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\| \ge \sigma\sqrt{2\ell\log\widetilde{p}_\ell}\Big\} \le \binom{p}{\ell}\frac{e^{-(x-\log(1+x))\ell/2}}{2^{1/2}(1+x)^{1/2}} = \frac{1}{2\sqrt{\log\widetilde{p}_\ell}}.$$

The conclusion follows from $2\log\widetilde{p}_\ell \ge 1$. □

Proof of Theorem 6. There are two cases. We first apply Theorem 9 to the case $\lambda < \lambda_{univ}/\alpha$, i.e. $(1+\epsilon_0)\gamma^{1/2}_{2,\ell}\lambda_{mm} < \lambda_{univ}$. Since $\|\boldsymbol{X}'_A\boldsymbol{\varepsilon}/n\| \le \gamma^{1/2}_{2,\ell}\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|/\sqrt{n}$,

$$P_{\boldsymbol{\beta}}\Big\{z^*_{1,\ell} \le z^*_{2,\ell} \le \gamma^{1/2}_{2,\ell}\sigma\sqrt{(2/n)\log\widetilde{p}_\ell}\Big\} \ge 1 - 1/(2\sqrt{\log\widetilde{p}_\ell}) \to 1.$$

Since $R^r/\lambda_{mm}^r \asymp d \asymp \ell$ and $\lambda_{mm} n^{1/2}/\sigma \to \infty$, $\lambda_{mm} = (1+o(1))\sigma\sqrt{(2/n)\log(p/\ell)}$ by (3.51). By Stirling, $\log\binom{p}{\ell} = (1+o(1))\ell\log(p/\ell)$ for $p/\ell \to \infty$. It follows that $\lambda_{mm} = (1+o(1))\sigma\sqrt{(2/n)\log\widetilde{p}_\ell}$. Thus, $z_{1,\ell}^* \le z_{2,\ell}^* \le \gamma_{2,\ell}^{1/2}\sigma\sqrt{(2/n)\log\widetilde{p}_\ell} \le \alpha\lambda$ with large probability. Moreover, since $E_{\boldsymbol{\beta}}(z_{2,J_k}^*)^2 = \mathrm{trace}(\sigma^2\boldsymbol{\Sigma}_{J_k}/(nk)) = (\sigma^2/n) = o(\lambda^2)$, $z_{2,J_k}^* \le \alpha_*\lambda$ with large probability for certain $\alpha_* = o(1)$. It follows that (3.34) is valid with large probability for the $\tau = \{(1+\epsilon_0)^2 k/\ell + \alpha^2\}^{1/2}/(1-\alpha)$.

Let $q = 2$ and $T_A(\boldsymbol{u}) = \boldsymbol{u}$ in (3.34) and switch notation $(q,s) \to (2,q)$. By (3.37),

$$G_{A,\boldsymbol{u}} + 2\rho_k(\boldsymbol{\beta})/\sqrt{k} \le \|\boldsymbol{w}\|\{\tau(1-\alpha)\sqrt{\ell} + \sqrt{k}/2\}\lambda + (2+1/2)\rho_k(\boldsymbol{\beta})/\sqrt{k}.$$

Since $\rho_k(\boldsymbol{\beta}) \le R^r(R^r/k)^{(1-r)/r} = Rk^{1-1/r}$ and $\boldsymbol{w} = \boldsymbol{u}/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u})$, we have

$$G_{A,\boldsymbol{u}} + 2\rho_k(\boldsymbol{\beta})/\sqrt{k} \le \widetilde{G}\sqrt{k}\lambda/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u}) + (5/2)Rk^{1/2-1/r} \tag{3.54}$$

for the $\widetilde{G} = \tau(1-\alpha)\sqrt{\ell/k} + 1/2$. Thus, (3.34) gives

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \;\le\; C_* k^{1/q} \max_{A,B,\boldsymbol{u}} \frac{\widetilde{G}\lambda/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u}) + (5/2)R/k^{1/r}}{\{1 - \tau\|\boldsymbol{\Sigma}_{B,A}\boldsymbol{u}\|_1\ell^{-1/2}/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u})\}_+}, \; 1 \le q \le 2,$$

for the $C_* = (1+\tau^2)^{1-1/q}(1+\tau\sqrt{\ell/k})^{2/q-1}$. In view of the definitions in (3.25) and (3.26), this gives $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q \le N_{d,\ell,\tau}(C_1,C_2))R^{r/q}\lambda_{mm}^{1-r/q}$.

We apply Theorem 8 to the second case where $\lambda = \lambda_{univ}/\alpha$. Since $E(z_{2,J_k}^*)^2 = (\sigma^2/n) = o(\lambda^2)$, $P_{\boldsymbol{\beta}}\{z_\infty^* \le \alpha\lambda, z_{\infty,J_k}^* \le \alpha_*\lambda\} \to 1$ with $\alpha_* = \epsilon_0$. Thus, (3.34) is valid with large probability for the $\tau = \xi\sqrt{k/\ell} = \sqrt{k/\ell}(1+\epsilon_0)/(1-\alpha)$ due to $\|\boldsymbol{u}_{J_k}\|_1 \le \sqrt{k}$. By (3.33),

$$G_{A,\boldsymbol{u}} + 2\rho_k(\boldsymbol{\beta})/\sqrt{k} \le (1+\epsilon_0)\lambda\|\boldsymbol{w}_{J_k}\|_1 + (1+\alpha)\lambda\|\boldsymbol{w}\|\sqrt{k}/2 + (2+1/2)\rho_k(\boldsymbol{\beta})/\sqrt{k}.$$

Since $\|\boldsymbol{w}_{J_k}\|_1 \le \sqrt{k}\|\boldsymbol{w}\|$ and $\|\boldsymbol{w}\| = 1/(\boldsymbol{u}'\boldsymbol{\Sigma}_A\boldsymbol{u})$, (3.54) holds for the $\widetilde{G} = (1+\epsilon_0) + (1+\alpha)/2$. The rest of the proof is the same as the first case and omitted. □

3.5 Discussion

Since the oracle inequalities apply directly to data points $(\boldsymbol{X},\boldsymbol{y})$ and target vectors $\boldsymbol{\beta}^*$, the normality assumption on the error in (3.3) is not crucial for our upper

bounds for the estimation risk and loss (not even the condition $E_{\boldsymbol{\beta}}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$). For example, for the estimation of a target $\boldsymbol{\beta}^*$ with $\boldsymbol{X}\boldsymbol{\beta}^* \approx E\boldsymbol{y}$, the upper bounds in Theorem 6 are valid for $\|\widehat{\boldsymbol{\beta}}_{Lasso} - \boldsymbol{\beta}^*\|_q^q$ with large probability under P and $\sigma = \sigma_1 + \sigma_2$, provided that

$$E\exp(\boldsymbol{v}'\boldsymbol{X}'(\boldsymbol{y} - E\boldsymbol{y})) \le \exp(-n\sigma_1^2\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v}/2),\ \max_{|A|=\ell}\|\boldsymbol{P}_A(E\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\| \le \sigma_2\sqrt{2\ell\log(p/\ell)}.$$

For design matrices $\boldsymbol{X}$ with iid sub-Gaussian rows, our results can be extended to $\boldsymbol{\beta}$ in ℓ_r balls with $1 < r \le 2$ due to $\sigma_2 \le O(1)\rho_{2,\ell}(\boldsymbol{\beta})$ when the target is $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{b}}\|\boldsymbol{\beta} - \boldsymbol{b}\|$ subject to $\|\boldsymbol{b}\|_0 = k = \ell$.

The proofs in this chapter do not completely deal with the most difficult case of $q > 2$ and $\lambda_{mm} = o(\lambda_{univ})$. For example, an application of Theorem 9 under $P_{\boldsymbol{\beta}}$ in (3.3) would require an upper bound for the $z^*_{q,\ell}$ with $\boldsymbol{\beta}^* = \boldsymbol{\beta}$ in (3.29).

Let $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Dantzig}(\lambda)$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Lasso}(\lambda)$. Since $\|\widetilde{\boldsymbol{\beta}}\|_1 \le \|\widehat{\boldsymbol{\beta}}\|_1$,

$$\begin{aligned}
\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k^c}\|_1 &\le \|(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})_{J_k^c}\|_1 + \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k^c}\|_1 \\
&\le 2\|\widehat{\boldsymbol{\beta}}_{J_k^c}\|_1 + \|(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})_{J_k}\|_1 + \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k^c}\|_1 \\
&\le \{3\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k^c}\|_1 + 2\rho_k(\boldsymbol{\beta}^*) + \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k}\|_1\} + \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J_k}\|_1.
\end{aligned}$$

This and the results for the Lasso in Theorem 9 would yield slightly worse error bounds of the same type for the Dantzig selector with $\lambda < \lambda_{univ}$. We have decided to omit an explicit statement of this result.

The proofs of Theorem 9 can be modified to extend the oracle inequalities for the Dantzig selector to

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}}\left\{\|\boldsymbol{b}\|_1 : \max_{|A|=d}\|\boldsymbol{X}'_A(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})\| \le \sqrt{d}\lambda\right\} \tag{3.55}$$

in the event $z^*_{2,d} \le \lambda = o(\lambda_{univ})$ in (3.29). This will provide sharper error bounds for the smaller λ and $q \le 2$. We omit this modification since the computational issues with the convex programming for (3.55) and a data-driven choice of d for $d > 1$ are not completely clear to the authors at the time of this writing, although $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{Dantzig}$ for $d = 1$.

References

[1] Bergman, R. N., Finegood D. T. and Ader, M. (1985). Assessment of insulin sensitivity in vivo. *Endocr. Rev.* **6**45-86.

[2] Bickel, P., Ritov Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.

[3] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electron. J. Statist.* **1** 169-194 (electronic).

[4] Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313-2404.

[5] Candes, E. and Plan, Y. (2009) Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145-2177.

[6] Chen, S. and Donoho, D.L. (1994). On basis pursuit. Technical Report, Department of Statistics, Stanford University.

[7] Donoho, D.L. and Johnstone, I. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277-303.

[8] Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992). Maximum entropy and the nearly black object (with discussion). *J. R. Statist. Soc. B* **54** 41-81.

[9] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407-499.

[10] Efron, B., Hastie, T. and Tibshirani, R. (2007). Discussion: The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2358-2364.

[11] Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kauffmann, San Francisco, 148-156.

[12] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* **28** 337-307.

[13] Greenshtein E. and Ritov Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971-988.

[14] Hovorka *et al* (2004). Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiol. Meas.* **25** 905-920.

[15] Mastrototaro J., Palerm C., Shepp L. and Zhang C.-H. (2008). Statistical imputation of the blood glucose levels that would be obtained under a given closed loop algorithm from "frozen data" which was obtained under open loop control. Preprint.

[16] Meinshausen, N. and Buhlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[17] Meinshausen, N. and Yu, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246-270.

[18] Osborne, M., Presnell, B. and Turlach, B. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20** 389-404.

[19] Osborne, M., Presnell, B. and Turlach, B. (2000b). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9** (2) 319-337.

[20] Palerm C., Shepp L., Cabrera J. and Zhang C.-H. (2008). A theoretical formula for HbA1c derived from continuous blood glucose measurements. Technical Report, Department of Statistics and Biostatistics, Rutgers University.

[21] Steil G.M., Clark B., Kanderian S. and Rebrin K. (2005). Modeling insulin action for development of a closed-loop artificial pancreas. *Diabetes Technology & Therapeutics* **7** 1.

[22] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[23] Tropp, J.A. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** 1030-1051.

[24] van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614-645.

[25] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy recovery of sparsity using ℓ_1 constrained quadratic programming. *IEEE Trans. Information Theory*, to appear.

[26] Zhang, C.-H. (2009). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, to appear.

[27] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.

[28] Zhang, T. (2009). Some Sharp Performance Bounds for Least Squares Regression with L1 Regularization. *Ann. Statist.* **37** 2109-2144.

[29] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7** 2541-2567.

Vita

Fei Ye

2005 B.S. in Mathematics, Fudan University, Shanghai, China.

2008 M.S. in Mathematics with an Option in Mathematical Finance, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA.

2010 Ph.D. in Statistics, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA.