Estimation and Testing Methods for Monotone Transformation Models

Junyi Zhang

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Estimation and Testing Methods for Monotone Transformation Models

Junyi Zhang

This thesis deals with a general class of transformation models that contains many important semiparametric regression models as special cases. It develops a self-induced smoothing method for estimating the regression coefficients of these models, resulting in simultaneous point and variance estimations. The self-induced smoothing does not require bandwidth selection, yet provides the right amount of smoothness so that the estimator is asymptotically normal with mean zero (unbiased) and variance-covariance matrix consistently estimated by the usual sandwich-type estimator. An iterative algorithm is given for the variance estimation and shown to numerically converge to a consistent limiting variance estimator. The self-induced smoothing method is also applied to selecting the non-zero regression coefficients for the monotone transformation models. The resulting regularized estimator is shown to be $\sqrt{n}$-consistent and achieve desirable sparsity and asymptotic normality under certain regularity conditions. The smoothing technique is used to estimate the monotone transformation function as well. The smoothed rank-based estimate of the transformation function is uniformly consistent and converges weakly to a Gaussian process which is the same as the limiting process for that without smoothing. An explicit covariance function estimate is obtained by using the smoothing technique, and shown to be consistent. The estimation of the transformation function reduces the multiple hypotheses testing problems for the monotone transformation models to those for linear models. A new hypotheses testing procedure is proposed in this thesis for linear models and shown to be more powerful than some widely-used testing methods when there is a strong collinearity in data. It is proved that the new testing procedure controls the family-wise error rate.

Contents

List of Tables iv

List of Figures v

Acknowledgments vi

Chapter 1 Introduction 1

Chapter 2 Parameter Estimation 5

2.1 Introduction 5

2.2 Main Results 7

2.2.1 Methods 8

2.2.2 Large-sample properties 11

2.2.3 Proofs of the Theorems 14

2.3 Extensions 22

2.4 Numerical Results 25

2.4.1 PBC data 25

2.4.2 Simulation studies 26

2.5 Other Proofs 30

2.5.1 Lemmas and corollaries 30

2.5.2 A sufficient condition for Assumption 2.3 37

2.6 Discussion 38

Chapter 3 Variable Selection 39

3.1 Introduction 39

3.2 Main Results 40

3.2.1 Large-sample properties 43

3.3 Proofs of the Theorems 45

3.4 Numerical Results 48

3.5 Discussion 54

Chapter 4 Estimation of the Monotone Transformation Function 59

4.1 Introduction 59

4.2 Main Results 60

4.2.1 Methods 61

4.2.2 Large-sample properties 62

4.3 Discussion 64

Chapter 5 Multiple Hypotheses Testing 67

5.1 Introduction 67

5.1.1 Background 67

5.1.2 The issue of collinearity 70

5.2 Basics 72

5.2.1 Marginal statistics 72

5.2.2 Chi-squared statistics 75

5.3 Method and Application 76

5.3.1 Generalized marginal statistics 76

5.3.2 MMRD step-down procedure 77

5.3.3 Application to the REE studies 79

5.4 Theory 79

5.4.1 Controlling the error rate 79

5.4.2 Distributional properties of $\left|U_j^B\right|_\infty$ 83

5.5 Other Properties 87

5.5.1 Variance of marginal fit 87

5.5.2 Chi-squared statistics 88

5.5.3 Distributional theories 89

Bibliography **90**

List of Tables

2.1 Regression Analysis of PBC data 25

2.2 The proportional hazard model without censoring 27

2.3 The proportional hazard model with censoring 27

2.4 The linear model with gaussian noise 28

3.1 The simulations for a linear model with gaussian noise 50

3.2 MAD of estimated coefficients for a linear model with gaussian noise . 51

3.3 BIAS of estimated coefficients for a linear model with gaussian noise . 52

3.4 The simulations for a linear model with mixed noise 55

3.5 MAD of estimated coefficients for a linear model with mixed noise . . 56

3.6 BIAS of estimated coefficients for a linear model with mixed noise . . 57

5.1 The testing results of MMRD 80

5.2 The simulated quantiles 86

List of Figures

2.1 The smoothed rank correlation for PBC data. 26

4.1 Estimation of the exponential function 65

4.2 Estimation of the logarithm function 65

4.3 Estimation of the logistic function 66

4.4 Estimation of the mixed function 66

5.1 Nested testing hypotheses 71

5.2 The distribution of U_1 based on the REE data 81

5.3 The impact of correlation on the distribution of U_1 86

Acknowledgments

I would like to express my deeply-felt thanks to my Ph.D advisor, Prof. Zhiliang Ying, for his great encouragement, support and guidance throughout my graduate study and my thesis-writing period. Prof. Ying has been not only a great academic advisor, but a great mentor in my life. He has taught me so much that I can benefit greatly from for my entire life.

I would also like to thank Prof. Daniel Rabinowitz, Prof. Yongzhao Shao, Prof. Zhezhen Jin and Prof. Yang Feng for agreeing to serve on my defense committee and their helpful suggestions. In addition, I would like to thank Prof. Yongzhao Shao, Prof. Zhezhen Jin, Prof. Yang Feng and Dr. Zimian Wang for their great help on my research. It is always an invaluable and enjoyable experience working with them.

I am grateful to all my friends including the faculty, the staff and my classmates in the Department of Statistics at Columbia University for their encouragement and support throughout the past five years. I have had many opportunities to learn from the professors and got inspiration from my Ph.D classmates.

Last but most importantly, I want to express my deepest thanks to my parents, grandparents, and my fiancée, Yun Ling, who is a Ph.D. student in USC Marshall

School of Business, for their tolerance, understanding and love.

To my parents, grandparents and those who educate me

Chapter 1

Introduction

Consider the following class of regression models, known as the monotone transformation models, with response variable denoted by Y and $(d+1)$-dimensional covariate vector by $\mathbf{X}$,

$$Y = H(\mathbf{X}'\boldsymbol{\beta} + \varepsilon), \tag{1.1}$$

where $\boldsymbol{\beta}$ is the unknown regression parameter vector, ε is the unobserved error term that is independent of $\mathbf{X}$ with a completely unspecified distribution, and H is a monotone increasing, but otherwise unspecified function.

It is easily seen that this class of models contains many commonly used regression models as its submodels that are especially important in the econometrics and survival analysis literature. For example, with $H(u) = u$, (1.1) becomes the standard regression model with an unspecified error distribution; with $H(u) = u^{\lambda}$ $(\lambda > 0)$, the Box-Cox transformation model (Box and Cox, 1964); with $H(u) = I[u \geq 0]$, the binary choice model (Maddala, 1983; McFadden, 1984); with $H(u) = uI[u \geq 0]$, a censored regression model (Tobin, 1958; Powell, 1984); with $H(u) = \exp(u)$, the accelerated failure times (AFT) model (Cox and Oakes, 1984; Kalbfleisch and Prentice, 2002); with ε having an extreme value density $f(w) = \exp(w - \exp(w))$, the Cox proportional hazards regression (Cox, 1972); with ε having the standard logistic distribution, the proportional odds regression (Bennett, 1983). In addition to the econometrics, model (1.1) also encompasses the main semiparametric models in survival analysis, where right censoring is a major feature. Under the right censorship, there is a censoring variable C and one observes $\tilde{Y} = Y \wedge C$ and $\Delta_i = I(Y_i \leq C_i)$.

Estimation of the parameter $\boldsymbol{\beta}$ was studied by Han (1987), Sherman (1993), Khan and Tamer (2007). In particular, Han (1987) proposed the maximum rank correlation (MRC) estimator and proved the strong consistency of the MRC estimator; Sherman (1993) showed the $\sqrt{n}$-consistency and the asymptotic normality for Han's MRC estimator and proposed an estimate of the limiting variance-covariance matrix for the MRC estimator by using the finite-difference approximation; Khan and Tamer (2007) proposed the partial rank correlation estimator for $\boldsymbol{\beta}$ when there is censoring in model (1.1). Estimation of the transformation function H was studied by Chen (2002), who constructed a rank-based estimator and established its consistency and asymptotic normality.

This thesis focuses on (1) estimation of the regression coefficient $\boldsymbol{\beta}$, which is of finite dimension, and of the transformation function H, which is of infinite dimension; (2) the variable selection problem for $\boldsymbol{\beta}$, especially when p is larger than n; (3) the multiple hypotheses testing problem related to the linear and monotone transformation models.

In Chapter 2, we develop a self-induced smoothing method for estimating the regression coefficient $\boldsymbol{\beta}$ to address the issue of discreteness in Han's rank correlation objective function. Through the self-induced smoothing method, we bypass the bandwidth selection problem associated with the finite difference approximation. We show that the self-induced smoothing provides the right amount of smoothness so that the estimator is asymptotically normal with mean zero (unbiased) and variance-covariance matrix consistently estimated by the usual sandwich-type estimator. An iterative algorithm is given for the variance estimation and shown to numerically converge to a consistent limiting variance estimator. The approach is applied to a data set involving survival times of primary biliary cirrhosis patients. Simulations results are reported, showing that the new method performs pretty well under a variety of scenarios.

Chapter 3 is concerned with the variable selection problem for the monotone transformation model (1.1). We apply the self-induced smoothing method to Han's rank correlation function and develop a variable selection method which is distribution-free and robust. The new variable selection method consists of the regularized SMRCE and the rank correlation information criteria. For the regularized SMRCE (RSMRCE), we add the SCAD (Fan and Li, 2001) penalty function to the smoothed rank correlation function. The rank correlation information criteria is introduced as a modified rank correlation function, which is adjusted for the dimensional complexity of the predictors been selected. We show that the regularized SMRCE achieves desired sparsity. Moreover, the RSMRCE does not introduce any bias for a proper thresholding level in the sense that the regularized estimator is $\sqrt{n}$-consistent and asymptotically normal. Extensive simulation studies show that the proposed variable selection procedure are more robust than the existing methods such as the LASSO-BIC approach.

Chapter 4 deals with the estimation problem for the monotone transformation function. We apply the self-induced smoothing method to Chen's (2002) rank-based estimator. The smoothed estimate for the monotone function is shown to be uniformly consistent and to converge weakly to a Gaussian process. Through the smoothing technique, we derive a close form covariance formula for the limiting Gaussian process for Chen's rank-based estimate. This covariance estimate is also consistent.

In Chapter 5, we develop a new multiple hypotheses testing procedure, which is called the minimax of marginal regression distances (MMRD) step-down method, for linear models. We prove that the new testing procedure controls the family-wise error rate. The MMRD procedure is shown to be more powerful than Holm's (1979) step-down procedure and Benjamini-Hochberg's (1995) false discovery rate (FDR) controlling procedure when applied to the REE studies.

Chapter 2

Parameter Estimation

2.1 Introduction

A basic estimation method for model (1.1) is the maximum rank correlation (MRC) estimator proposed in the econometrics literature by Han (1987). Because both the transformation function H and the error distribution are unspecified, not all components of $\boldsymbol{\beta}$ are identifiable. Without loss of generality, we shall assume henceforth that the last component, $\beta_{d+1} = 1$. Let $(Y_1, \mathbf{X}_1), ..., (Y_n, \mathbf{X}_n)$ be a random sample from (1.1). Han's MRC estimator, denoted by $\hat{\boldsymbol{\theta}}_n$, is the maximizer of following objective function

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} I[Y_i > Y_j] I[\mathbf{X}_i' \boldsymbol{\beta}(\boldsymbol{\theta}) > \mathbf{X}_j' \boldsymbol{\beta}(\boldsymbol{\theta})], \tag{2.1}$$

where $I[\ \cdot\ ]$ denotes the indicator function, $\mathbf{X}'$ the transpose of $\mathbf{X}$, and $\boldsymbol{\theta}$ the first d components of $\boldsymbol{\beta}$, i.e. $\boldsymbol{\beta}(\boldsymbol{\theta}) = (\theta_1, ..., \theta_d, 1)'$. Han (1987) proved that the MRC estimator $\hat{\boldsymbol{\theta}}_n$ is strongly consistent under certain regularity conditions.

An important subsequent development is due to Sherman (1993), who made use of the empirical process theory and Hoeffding's decomposition to approximate the objective function, viewed as a U-process. He showed that $\hat{\boldsymbol{\theta}}_n$ is, in fact, asymptotically normal under additional regularity conditions.

For the censoring case, Khan and Tamer (2007) constructed the following partial rank correlation function as an extension of the rank correlation objective function (2.1),

$$Q_n^*(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I[\tilde{Y}_i > \tilde{Y}_j] I[\mathbf{X}_i' \boldsymbol{\beta}(\boldsymbol{\theta}) > \mathbf{X}_j' \boldsymbol{\beta}(\boldsymbol{\theta})]. \tag{2.2}$$

They showed that the resulting maximum partial rank correlation estimate (PRCE) $\hat{\boldsymbol{\theta}}_n^*$, as the maximizer of $Q_n^*(\boldsymbol{\theta})$, is consistent and asymptotically normal.

Crucial for the statistical inference of (1.1) based on $\hat{\boldsymbol{\theta}}_n$ is the consistent variance estimation. In standard objective (loss) function derived estimation, the asymptotic variance is usually estimated by a sandwich-type estimator of form $\hat{\mathbf{A}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{A}}^{-1}$ with $\hat{\mathbf{A}}$ being the second derivative of the objective function and $\hat{\mathbf{V}}$ an estimator of the variance of the first derivative (score). The challenge here, however, is that Q_n itself is a (discontinuous) step function that precludes automatic use of differentiation to obtain $\hat{\mathbf{A}}$. Furthermore, $\hat{\mathbf{V}}$ is also difficult to obtain since the score function cannot be derived directly from Q_n via differentiation. Sherman(1993) suggested using numerical derivatives of first and second orders to construct $\hat{\mathbf{A}}$ and $\hat{\mathbf{V}}$. His approach requires bandwidth selection for the derivative functions. It is unclear how stable the resulting variance estimator is. Alternatively, one may resort to bootstrap (Efron, 1979) or other resampling methods (e.g. Jin et al., 2001). These approaches require repeatedly solving the maximization of (2.1), which is discontinuous and often multidimensional when $d > 1$. The computational cost could therefore be prohibitive.

In this chapter, a self-induced smoothing method for rank correlation criterion function (2.1) is developed so that the differentiation can be performed, while bypassing the bandwidth selection. Both point and variance estimators can be obtained simultaneously in a straightforward way that is typically used for smooth objective functions. The new method is motivated by a novel approach proposed in Brown and Wang (2005, 2007), where an elegant self-induced smoothing method was introduced for non-smooth estimating functions. Although our approach bears similarity with that of Brown and Wang (2005), it is far from clear why such self-induced smoothing is suitable for the discrete objective function (rank correlation). In fact, undersmoothing would make the Hessian (second derivative) unstable while oversmoothing would introduce significant bias. Through highly technical and tedious derivations, the author will show that the proposed method does strike a right balance in terms of asymptotic unbiasedness and enough smoothness for differentiation (twice).

The rest of this chapter is organized as follows. In Section 2.2, the new methods are described and related large sample properties are developed. In particular, the construction for simultaneous point and variance estimation is given and it is shown that the resulting point estimator is asymptotically normal and the variance estimator is consistent. In Section 2.3, the approach, along with the algorithm and large sample properties, is extended to handle survival data with right censoring. Simulation results are reported in Section 2.4, where application to a real data set is also given. Section 2.5 contains some concluding remarks.

2.2 Main Results

In this section, a self-induced smoothing method is developed for the rank correlation criterion function defined by (2.1). It is divided into three subsections, with the first introducing the method and the algorithm, the second establishing large sample properties and the third covering proofs.

2.2.1 Methods

Since MRC estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically normal (Sherman, 1993), its difference with the true parameter value, $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$, should approximately be a Gaussian noise $\mathbf{Z}/\sqrt{n}$, where $\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Assume that $\mathbf{Z}$ is independent of data and let $E_{\mathbf{Z}}$ denote the expectation with respect to $\mathbf{Z}$ given data. A self-induced smoothing for Q_n is $\tilde{Q}_n(\boldsymbol{\theta}) = E_{\mathbf{Z}} Q_n(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n})$. The self-induced smoothing using the limiting Gaussian distribution was originally proposed by Brown and Wang (2005) for certain non-smooth estimating functions.

To get an explicit form for $\tilde{Q}_n$, let Φ be the standard normal distribution function, $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j$, $\sigma_{ij} = \sqrt{(\mathbf{X}_{ij}^{(1)})' \boldsymbol{\Sigma} \mathbf{X}_{ij}^{(1)}}$ where $\mathbf{X}_{ij}^{(1)}$ denotes the first d components of $\mathbf{X}_{ij}$. Then, it is easy to see that

$$\tilde{Q}_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} I[Y_i > Y_j] \Phi\left(\sqrt{n} \mathbf{X}_{ij}' \boldsymbol{\beta}(\boldsymbol{\theta}) / \sigma_{ij}\right). \tag{2.4}$$

We shall use $\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\Theta}} \tilde{Q}_n(\boldsymbol{\theta})$ to denote the corresponding estimator, which will be called the smoothed maximum rank correlation estimator (SMRCE). Here and in the sequel, $\boldsymbol{\Theta}$ denotes the parameter space for $\boldsymbol{\theta}$.

Remark 2.1. Smoothing is an appealing way for a simple solution to the inference problem associated with the MRCE. If $\tilde{Q}_n$ were a usual smooth objective function, then its first derivative would become the score function and its second derivative could be used for variance estimation. Specifically, if we use $\mathbf{V}$ to denote the limiting variance of the score scaled by n and $\mathbf{A}$ the limit of the second derivative, then the asymptotic variance of the resulting estimator, scaled by n, should be of form $\mathbf{A}^{-1}\mathbf{V}\mathbf{A}^{-1}$. A consistent estimator could then be obtained by the plug-in method, i.e. replacing unknown parameters by their corresponding empirical estimators.

Remark 2.2. It is unclear, however, whether or not the self-induced smooth will provide a right amount of smoothing, even in view of the results given in Brown and Wang (2005). With over-smoothing, $\tilde{\boldsymbol{\theta}}_n$ may be asymptotically biased, i.e. the bias is not of order $o(n^{-1/2})$; with under-smoothing, the "score" function (first derivative of $\tilde{Q}_n$) may have multiple "spikes" and thus the second derivative matrix (Hessian) of $\tilde{Q}_n$ may not behave properly and certainly cannot be expected to provide a consistent variance estimator.

In Subsection 2.2.2, it is shown that the self-induced smoothing here does result in a right amount of smoothing in the sense that the bias is asymptotically negligible and the Hessian matrix behave properly. Before starting the theoretic developments, the method is described as follows first.

We first differentiate the smoothed objective function $\tilde{Q}_n$ to get score

$$\tilde{\mathbf{S}}_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i<j} H_{ij} \phi\left(\frac{\sqrt{n}\mathbf{X}'_{ij}\boldsymbol{\beta}(\boldsymbol{\theta})}{\sigma_{ij}}\right) \frac{\sqrt{n}\mathbf{X}^{(1)}_{ij}}{\sigma_{ij}},$$

where $H_{ij} = sgn(Y_i - Y_j)$. This is a U-process of order 2 with kernel

$$\mathbf{s}_n(\mathbf{U}_i, \mathbf{U}_j) = \frac{1}{2} H_{ij} \phi\left(\frac{\sqrt{n}\mathbf{X}'_{ij}\boldsymbol{\beta}(\boldsymbol{\theta})}{\sigma_{ij}}\right) \frac{\sqrt{n}\mathbf{X}^{(1)}_{ij}}{\sigma_{ij}},$$

where $\mathbf{U}_i$ denotes the pair $(Y_i, \mathbf{X}_i)$.

By Hoeffding's decomposition, the asymptotic variance of $\sqrt{n}\tilde{\mathbf{S}}_n(\boldsymbol{\theta})$ is approximated by

$$\hat{\mathbf{V}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{n^3} \sum_{i=1}^{n} \left\{ \sum_{j} \left[H_{ij} \times \phi\left(\frac{\sqrt{n}\mathbf{X}'_{ij}\boldsymbol{\beta}}{\sigma_{ij}}\right) \frac{\sqrt{n}\mathbf{X}^{(1)}_{ij}}{\sigma_{ij}} \right] \right\}^{\otimes 2}, \tag{2.5}$$

where, for a vector $\mathbf{v}$, $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$. Thus, $\hat{\mathbf{V}}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\Sigma})$ is used to estimate $\mathbf{V}$, the middle part of the "sandwich" variance formula discussed in Remark 2.1.

As for $\mathbf{A}$, we differentiate $\tilde{\mathbf{S}}_n(\boldsymbol{\theta})$ to get

$$\hat{\mathbf{A}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{2n(n-1)} \sum_{i \neq j} \left\{ H_{ij} \times \dot{\phi}\left(\frac{\sqrt{n}\mathbf{X}'_{ij}\boldsymbol{\beta}}{\sigma_{ij}}\right) \left[\frac{\sqrt{n}\mathbf{X}^{(1)}_{ij}}{\sigma_{ij}}\right]^{\otimes 2} \right\}, \tag{2.6}$$

where $\dot{\phi}(z) = -z\phi(z)$ is the derivative of $\phi(z)$. Although the self-induced smoothing was motivated earlier with $\boldsymbol{\Sigma}$ being the limiting covariance matrix of the estimator, it will be shown later that for any positive definite matrix $\boldsymbol{\Sigma}$, $\hat{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\Sigma})$ converges to $\mathbf{A}$.

Note that the above discussions about $\mathbf{A}$ and $\mathbf{V}$ are not mathematically rigorous. This is because the kernel function for the score process is sample size n-dependent. The usual asymptotic theory for the U-process is not applicable. Indeed, our rigorous derivations, to be given in Subsection 2.2.3, are quite tedious, involving many approximations that are quite delicate.

Let

$$\hat{\mathbf{D}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \hat{\mathbf{A}}_n^{-1}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \times \hat{\mathbf{V}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \times \hat{\mathbf{A}}_n^{-1}(\boldsymbol{\theta}, \boldsymbol{\Sigma}). \tag{2.7}$$

If $\boldsymbol{\theta}$ is the true parameter value, then $\hat{\mathbf{D}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ converges to the limiting covariance matrix, which is the desired choice for $\boldsymbol{\Sigma}$ in the self-induced smoothing. Therefore, (2.7) leads to an iterative algorithm of form $\hat{\boldsymbol{\Sigma}}_n^{(k)} = \hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Sigma}}_n^{(k-1)})$; see also Brown and Wang (2005). Specifically, an iterative algorithm is proposed as follows:

Algorithm 2.1. *(SMRCE)*

1. *Compute the MRC estimator $\hat{\boldsymbol{\theta}}_n$ and set $\hat{\boldsymbol{\Sigma}}^{(0)}$ to be the identity matrix.*

2. *Update variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_n^{(k)} = \hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\Sigma}}_n^{(k-1)})$. Smooth the rank correlation $Q_n(\boldsymbol{\theta})$ using covariance matrix $\hat{\boldsymbol{\Sigma}}_n^{(k)}$. Maximize the resulting smoothed rank correlation to get an estimator $\hat{\boldsymbol{\theta}}_n^{(k)}$.*

3. *Repeat step 2 until $\hat{\boldsymbol{\theta}}_n^{(k)}$ converge.*

2.2.2 Large-sample properties

This subsection is devoted to the large sample theory. The main results are: 1. the smoothed MRC estimator (SMRCE) is asymptotically equivalent to the MRC estimator; 2. the proposed method leads to a consistent variance estimator; and 3. the iterative algorithm for point and variance estimation converges numerically.

First introduce notation as well as assumptions, which are similar to those in Sherman (1993) for the MRC estimator. Let

$$\tau(y, \mathbf{x}, \boldsymbol{\theta}) = E\left[I_{[y>Y]} I_{[(\mathbf{x}-\mathbf{X})'\boldsymbol{\beta}(\boldsymbol{\theta})>0]} + I_{[y<Y]} I_{[(\mathbf{x}-\mathbf{X})'\boldsymbol{\beta}(\boldsymbol{\theta})<0]}\right], \tag{2.8}$$

which is the projection of the kernel of U-process $Q_n(\boldsymbol{\theta})$. The expectation is taken for $(\mathbf{X}, Y)$. Also let

$$|\nabla_m|\tau(y, \mathbf{x}, \boldsymbol{\theta}) = \sum_{i_1,\ldots,i_m} \left| \frac{\partial^m \tau(y, \mathbf{x}, \boldsymbol{\theta})}{\partial\theta_{i_1} \cdots \partial\theta_{i_m}} \right|.$$

The following Assumptions 2.1 and 2.2 are used in Han (1987) (see also Sherman, 1993) to establish consistency for the MRC estimator. For asymptotic normality, we need an additional regularity condition (Assumption 2.3) given in Sherman (1993).

Assumption 2.1. The true parameter value $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$, which is a compact subset of the d-dimensional Euclidean space $\mathbb{R}^d$.

Assumption 2.2. The support of $\mathbf{X}$ is not contained in any linear subspace of $\mathbb{R}^{d+1}$. Conditional on the first d components of $\mathbf{X}$, the last component of $\mathbf{X}$ has a density function with respect to the Lebesgue measure.

Assumption 2.3. There exists a neighborhood, $\mathcal{N}$, of $\boldsymbol{\theta}_0$ such that for each pair $(y, \mathbf{x})$ of possible values of $(Y, \mathbf{X})$,

(i) The second derivatives of $\tau(y, \mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist in $\mathcal{N}$.

(ii) There is an integrable function $M_1(y, \mathbf{x})$ such that for all $\boldsymbol{\theta}$ in $\mathcal{N}$,

$$\|\nabla_2 \tau(y, \mathbf{x}; \boldsymbol{\theta}) - \nabla_2 \tau(y, \mathbf{x}; \boldsymbol{\theta}_0)\|_2 \leq M_1(y, \mathbf{x})|\boldsymbol{\theta} - \boldsymbol{\theta}_0|.$$

(iii) $E(|\nabla_1|\tau(Y, \mathbf{X}; \boldsymbol{\theta}_0))^2 < +\infty$.

(iv) $E|\nabla_2|\tau(Y, \mathbf{X}; \boldsymbol{\theta}_0) < +\infty$.

(v) The matrix $E\nabla_2 \tau(Y, \mathbf{X}; \boldsymbol{\theta}_0)$ is strictly negative definite.

Proposition 2.1. *(Sherman, 1993) Assume that Assumptions 2.1-2.3 hold. We have, uniformly over any $o_p(1)$ neighborhood of $\boldsymbol{\theta}_0$,*

$$Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}_0) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{A}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{\sqrt{n}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{W}_n + O_p(|\boldsymbol{\theta} - \boldsymbol{\theta}_0|^3) + o_p(\frac{1}{n}), \quad (2.9)$$

where $\mathbf{W}_n = \frac{1}{\sqrt{n}} \sum_i \nabla_1 \tau(Y_i, \mathbf{X}_i; \boldsymbol{\theta}_0)$, $2\mathbf{A}(\boldsymbol{\theta}) = E\nabla_2 \tau(Y, \mathbf{X}; \boldsymbol{\theta})$ and $\mathbf{A}_0 = \mathbf{A}(\boldsymbol{\theta}_0)$. Consequently, for the MRC estimator $\hat{\boldsymbol{\theta}}_n$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = A_0^{-1}\mathbf{W}_n + o_p(1) \xrightarrow{L} N(\mathbf{0}, \mathbf{D}_0), \quad (2.10)$$

where $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{A}^{-1}(\boldsymbol{\theta})\mathbf{V}(\boldsymbol{\theta})\mathbf{A}^{-1}(\boldsymbol{\theta})$, $\mathbf{V}(\boldsymbol{\theta}) = E(\nabla_1 \tau(Y, \mathbf{X}; \boldsymbol{\theta})[\nabla_1 \tau(Y, \mathbf{X}; \boldsymbol{\theta})]')$ and $\mathbf{D}_0 = \mathbf{D}(\boldsymbol{\theta}_0)$.

Because of the standardization, the rank correlation criterion function Q_n is bounded by 1. It is not difficult to establish a uniform law of large numbers

$$\lim_n \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| = 0, \ a.s., \quad (2.11)$$

where $Q(\boldsymbol{\theta})$ is the expectation of $Q_n(\boldsymbol{\theta})$; cf. Han (1987) and Sherman (1993). Likewise, we can show that such uniform convergence also holds for $\tilde{Q}_n$, i.e.

$$\lim_n \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\tilde{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| = 0, \ a.s. \quad (2.12)$$

Note that the limit Q remains the same.

In the following theorem, it is claimed that the estimate obtained from maximizing the smoothed rank correlation function (2.4) is also asymptotically normal with the same asymptotic covariance matrix as Han's MRCE.

Theorem 2.1. *For any given positive definite matrix* $\mathbf{\Sigma}$*, let* $\widetilde{Q}_n(\boldsymbol{\theta})$ *be defined as in* (2.4) *and* $\widetilde{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} E_Z \widetilde{Q}_n(\boldsymbol{\theta} + Z/\sqrt{n})$*. Then, under Assumptions 2.1-2.3,* $\widetilde{\boldsymbol{\theta}}_n$ *is consistent,* $\widetilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ *a.s. and asymptotically normal,*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{D}_0),$$

where $\mathbf{D}_0$ *is defined as in Proposition 2.1. In addition,* $\widetilde{\boldsymbol{\theta}}_n$ *is asymptotically equivalent to* $\hat{\boldsymbol{\theta}}_n$ *in the sense that* $\widetilde{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n + o_p(n^{-1/2})$*.*

Recall that (2.7) defines the sandwich-type variance estimator by pretending that $\widetilde{Q}_n$ is a standard smooth objective function. Theorem 2.2 below shows that (2.7) is consistent.

Theorem 2.2. *Let* $\hat{\boldsymbol{\theta}}_n$ *be the MRC estimator and* $\hat{\mathbf{D}}_n$ *be defined by* (2.7)*. Then, for any fixed positive definite matrix* $\mathbf{\Sigma}$*,* $\hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n, \mathbf{\Sigma})$ *converges in probability to* $\mathbf{D}_0$*, the limiting variance-covariance matrix of* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$*.*

Remark 2.3. The self-induced smoothing uses the limiting covariance matrix $\mathbf{D}_0$ as $\mathbf{\Sigma}$. In practice, we may initially choose the identity matrix for $\mathbf{\Sigma}$, which is the same way as the initial step in Algorithm 2.1. By Theorem 2.1, we know that the one-step estimator $\hat{\mathbf{\Sigma}}_n^{(1)}$ in Algorithm 2.1 converges in probability to the true covariance. However, this one-step estimator depends on the initial choice of $\mathbf{\Sigma}$. Algorithm 2.1 is an iterative algorithm with the variance-covariance estimator converging to the fixed point of $\hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n, \mathbf{\Sigma}) = \mathbf{\Sigma}$.

Convergence of Algorithm 2.1 is ensured by the following theorem. For notational simplicity, we let $vech(\mathbf{B})$ be the vectorization of matrix $\mathbf{B}$. For any function v of $\boldsymbol{\Sigma}$,

$$\left|\frac{\partial}{\partial \boldsymbol{\Sigma}}\right| v = \sum_{i,j} \left|\frac{\partial}{\partial \Sigma_{i,j}} v\right|, \quad \frac{\partial v}{\partial \boldsymbol{\Sigma}} = (\frac{\partial v}{\partial \Sigma_{1,1}}, \frac{\partial v}{\partial \Sigma_{2,1}}, ..., \frac{\partial v}{\partial \Sigma_{d,d}})',$$

where $\boldsymbol{\Sigma}_{r,s}$ denotes the (r,s) entry of $\boldsymbol{\Sigma}$.

Theorem 2.3. *Let $\hat{\boldsymbol{\Sigma}}_n^{(k)}$ be defined as in Algorithm 2.1. Suppose that Assumptions 2.1-2.3 hold. Then there exist $\boldsymbol{\Sigma}_n^*$, $n \geq 1$, such that for any $\epsilon > 0$, there exists N, such that for all $n > N$,*

$$P(\lim_{k\to\infty} \hat{\boldsymbol{\Sigma}}_n^{(k)} = \boldsymbol{\Sigma}_n^*, \quad \|\boldsymbol{\Sigma}_n^* - \mathbf{D}_0\| < \epsilon) > 1 - \epsilon.$$

Remark 2.4. For a fixed n, $\boldsymbol{\Sigma}_n^*$ represents the fixed point matrix in the iterative algorithm. The above theorem shows that with probability approaching 1, the iterative algorithm converges to a limit, as $k \to \infty$, and the limit converges in probability to the limiting covariance matrix $\mathbf{D}_0$.

Remark 2.5. The speed of convergence of $\hat{\boldsymbol{\Sigma}}_n^{(k)}$ to $\boldsymbol{\Sigma}_n^*$ is faster than any exponential rate in the sense that $\|\hat{\boldsymbol{\Sigma}}_n^{(k)} - \boldsymbol{\Sigma}_n^*\| = o(\eta^k)$ for any $\eta > 0$. This can be seen from Step 2 of Algorithm 2.1 in Subsection 2.2.1 and (2.13) below,

$$\sup_{\|\theta-\theta_0\|=o(1), \Sigma\in\mathcal{N}(D_0)} \left|\frac{\partial}{\partial \boldsymbol{\Sigma}}\right| [\hat{\mathbf{D}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})]_{r,s} = o_p(1), \tag{2.13}$$

which will be proved in the Section 2.5. Here $\mathcal{N}(\mathbf{D}_0)$ is a small neighborhood of $\mathbf{D}_0$ and $\boldsymbol{\Sigma}$ is a positive definite matrix.

2.2.3 Proofs of the Theorems

In this section, proofs are provided for (1) asymptotic equivalence of SMRCE to MRCE, (2) consistency of the induced variance estimator and (3) convergence of

Algorithm 2.1. Some of the technical developments used in the proofs will be given in the Section 2.5.

Proof of Theorem 2.1. Without loss of generality, let us assume $\boldsymbol{\theta}_0 = \mathbf{0}$. As in Subsection 2.1.1, let $\mathbf{Z}$ be a d-variate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Define

$$\widetilde{Q}_n(\boldsymbol{\theta}) = E_{\mathbf{Z}} Q_n(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}).$$

Let $\Gamma_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}_0)$ and $\widetilde{\Gamma}_n(\boldsymbol{\theta}) = E_{\mathbf{Z}}\Gamma_n(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}) = \widetilde{Q}_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}_0)$. Define

$$\widetilde{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} \left[\widetilde{Q}_n(\boldsymbol{\theta})\right] = \text{argmax}_{\boldsymbol{\theta}} \widetilde{\Gamma}_n(\boldsymbol{\theta}).$$

Let $\boldsymbol{\Omega}_n = I[\|\mathbf{Z}\|_2 > 2d\log n]$, where $\|\mathbf{Z}\|_2 = \sqrt{\mathbf{Z}'\mathbf{Z}}$. Then $P(\boldsymbol{\Omega}_n) = o(n^{-2})$ due to the Gaussian tail of $\mathbf{Z}$. Since $|Q_n(\boldsymbol{\theta})| \leq 1$ and $|\Gamma_n(\boldsymbol{\theta})| \leq 2$,

$$|E_{\mathbf{Z}}\{\Gamma_n(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}) I[\boldsymbol{\Omega}_n]\}| \leq P(\boldsymbol{\Omega}_n) = o(n^{-2}).$$

By the Cauchy-Schwarz inequality,

$$E_{\mathbf{Z}}\{|\mathbf{Z}| I[\boldsymbol{\Omega}_n]\} = o(n^{-2}) \text{ and } E_{\mathbf{Z}}\{|\mathbf{Z}|^2 I[\boldsymbol{\Omega}_n]\} = o(n^{-2}).$$

By (2.9), uniformly over $o(1)$ neighborhoods of $\mathbf{0}$,

$$\begin{aligned} E_{\mathbf{Z}}\{\Gamma_n(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}) I[\boldsymbol{\Omega}_n^c]\} &= (1/2) E_{\mathbf{Z}}\{(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n})' \mathbf{A}_0 (\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}) I[\boldsymbol{\Omega}_n^c]\} \\ &+ (1/\sqrt{n}) E_{\mathbf{Z}}\{(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n})' \mathbf{W}_n I[\boldsymbol{\Omega}_n^c]\} + o_p(E_{\mathbf{Z}}\{|\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}|^2 I[\boldsymbol{\Omega}_n^c]\} + \frac{1}{n}). \end{aligned}$$

Note that

$$E_{\mathbf{Z}}\{|\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}|^2 I[\boldsymbol{\Omega}_n^c]\} \leq 2(E_{\mathbf{Z}}|\boldsymbol{\theta}|^2 + E_{\mathbf{Z}}|\mathbf{Z}|^2/n) = O(|\boldsymbol{\theta}|^2 + 1/n).$$

Therefore, uniformly over $o(1)$ neighborhoods of $\mathbf{0}$, we have

$$\widetilde{\Gamma}_n(\boldsymbol{\theta}) = (1/2)\boldsymbol{\theta}' \mathbf{A}_0 \boldsymbol{\theta} + (1/\sqrt{n})\boldsymbol{\theta}' \mathbf{W}_n + E(\mathbf{Z}' \mathbf{A}_0 \mathbf{Z})/2n + o_p(|\boldsymbol{\theta}|^2 + 1/n). \tag{2.14}$$

Replacing $\boldsymbol{\theta}$ in (2.14) with $\boldsymbol{\theta}_0 = \mathbf{0}$ and subtracting it from $\widetilde{\Gamma}_n(\boldsymbol{\theta})$, we have

$$\widetilde{\Gamma}_n(\boldsymbol{\theta}) - \widetilde{\Gamma}_n(\boldsymbol{\theta}_0) = \frac{1}{2}\boldsymbol{\theta}'\mathbf{A}_0\boldsymbol{\theta} + \frac{1}{\sqrt{n}}\boldsymbol{\theta}'\mathbf{W}_n + o_p(|\boldsymbol{\theta}|^2 + 1/n). \tag{2.15}$$

Combining (2.15) with Lemma 2.1 in the Section 2.5, we get,

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_0^{-1}\mathbf{W}_n + o_p(1). \tag{2.16}$$

Therefore, from (2.10) and (2.16), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n) = o_p(1).$$

Finally, strong consistency of $\widetilde{\boldsymbol{\theta}}_n$ follows the uniform almost sure convergence of $\tilde{Q}_n$ as stated in (2.12). This completes the proof. □

Proof of Theorem 2.2. For notational simplicity, let us assume throughout the proof that $\boldsymbol{\Sigma}$ is the identity matrix. The same argument with modifications to include constants for up and lower bound may be applied to deal with a general covariance matrix $\boldsymbol{\Sigma}$.

Let us first show

$$\hat{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{A}(\boldsymbol{\theta}_0). \tag{2.17}$$

By definition, $[\hat{\mathbf{A}}_n(\boldsymbol{\theta})]_{r,s} = \partial^2\tilde{Q}_n(\boldsymbol{\theta})/(\partial\theta_r\partial\theta_s)$. As defined in (2.4), $\tilde{Q}_n(\boldsymbol{\theta})$ has the following integral representation,

$$\tilde{Q}_n(\boldsymbol{\theta}) = \int Q_n(\boldsymbol{\theta} + \mathbf{z}/\sqrt{n})(2\pi)^{-\frac{d}{2}}\exp(-\frac{\|\mathbf{z}\|_2^2}{2})d\mathbf{z}.$$

By change of variable $\mathbf{t} = \boldsymbol{\theta} + \mathbf{z}/\sqrt{n}$,

$$\tilde{Q}_n(\boldsymbol{\theta}) = \int Q_n(\mathbf{t})K_n(\mathbf{t}, \boldsymbol{\theta})d\mathbf{t}, \tag{2.18}$$

where $K_n(\mathbf{t},\boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}} n^{\frac{d}{2}} \exp(-\dfrac{n\|\mathbf{t}-\boldsymbol{\theta}\|_2^2}{2})$. From (2.18),

$$\frac{\partial}{\partial\theta_r}\tilde{Q}_n(\boldsymbol{\theta}) = \int Q_n(\mathbf{t})\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t}$$

and

$$\frac{\partial^2}{\partial\theta_r\partial\theta_s}\tilde{Q}_n(\boldsymbol{\theta}) = \int Q_n(t)\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t},$$

where $\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta}) = \partial K_n(\mathbf{t},\boldsymbol{\theta})/\partial\theta_r$ and $\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta}) = \partial^2 K_n(\mathbf{t},\boldsymbol{\theta})/(\partial\theta_r\partial\theta_s)$.

In view of (2.6), to show (2.17), it suffices to prove

$$\int Q_n(\mathbf{t})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = [\mathbf{A}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1) \tag{2.19}$$

uniformly over $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| = O(n^{-1/2})$. To show (2.19), we define

$$\boldsymbol{\Omega}_{n,r} = \left\{\mathbf{t} : (t_r-\theta_r)^2 < \frac{4\log n}{n}, \sum_{i\neq r}(t_i-\theta_i)^2 < \frac{2(d-1)\log n}{n}\right\}.$$

By Lemma 2.2(i) and the boundedness of $Q_n(\mathbf{t})$, we have,

$$\int_{(\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s})^c} Q_n(\mathbf{t})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = o(n^{-1/2}),$$

where $\mathfrak{B}^c$ for set $\mathfrak{B}$ denotes its complement. Therefore, (2.19) reduces to

$$\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}} Q_n(\mathbf{t})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = [\mathbf{A}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1). \tag{2.20}$$

To show (2.20), let us establish a quadratic expansion of $Q_n(\mathbf{t})$ for $\mathbf{t}\in\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}$. Since $\|\mathbf{t}-\boldsymbol{\theta}\|_2 < \sqrt{4d\log n/n}$ for $\mathbf{t}\in\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}$ and $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 = O(n^{-1/2})$, it follows that $\|\mathbf{t}-\boldsymbol{\theta}_0\|_2 = o(1)$. Therefore, by (2.9),

$$\begin{aligned} Q_n(\mathbf{t}) = Q_n(\boldsymbol{\theta}_0) &+ \frac{1}{2}(\mathbf{t}-\boldsymbol{\theta}_0)'\mathbf{A}(\boldsymbol{\theta}_0)(\mathbf{t}-\boldsymbol{\theta}_0) \\ &+ (\mathbf{t}-\boldsymbol{\theta}_0)'\mathbf{W}_n/\sqrt{n} + O_p(|\mathbf{t}-\boldsymbol{\theta}_0|^3) + o_p(1/n). \end{aligned} \tag{2.21}$$

Therefore, the left hand side of (2.20) equals $\mathbf{I}+\mathbf{II}+\mathbf{III}+\mathbf{IV}$, where

$$\begin{aligned}
\mathbf{I} &= \int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}} \left[O_p(|\mathbf{t}-\boldsymbol{\theta}_0|^3)+o_p(1/n)\right]\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t},\\
\mathbf{II} &= Q_n(\boldsymbol{\theta}_0)\times\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}} \ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t},\\
\mathbf{III} &= \frac{\mathbf{W}_n'}{\sqrt{n}}\times\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta}_0)\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t},\\
\mathbf{IV} &= \frac{1}{2}\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta}_0)'\mathbf{A}(\boldsymbol{\theta}_0)(\mathbf{t}-\boldsymbol{\theta}_0)\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t}.
\end{aligned}$$

By the definition of $\boldsymbol{\Omega}_{n,r}$,

$$|\mathbf{I}| \le \left|O_p\left(\frac{(\log n)^{\frac{3}{2}}}{n\sqrt{n}}\right)+o_p(1/n)\right|\times\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}|\ddot{K}_{n,r,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t}.$$

By Lemma 2.2(ii), $\mathbf{I}=o_p(1)$. Furthermore, $\mathbf{II}=o(n^{-1/2})$ due to Lemma 2.2(iii). Note that

$$\begin{aligned}
\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta}_0)\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} &= \int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t}\\
&\quad+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} \qquad (2.22)\\
&= (\boldsymbol{\theta}-\boldsymbol{\theta}_0)\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t},
\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{\Omega}_{n,r}$ and $\boldsymbol{\Omega}_{n,s}$ are symmetric at $\boldsymbol{\theta}$ and $(\mathbf{t}-\boldsymbol{\theta})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})$ is an odd function of $[\mathbf{t}-\boldsymbol{\theta}]_r$ for $r=1,2,...,d$. Combining this with Lemma 2.2(i), we have $\mathbf{III}=\mathbf{o(n^{-1})}$. Again by symmetry,

$$\begin{aligned}
&\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta}_0)'\mathbf{A}(\boldsymbol{\theta}_0)(\mathbf{t}-\boldsymbol{\theta}_0)\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t}\\
&\qquad=\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}(\mathbf{t}-\boldsymbol{\theta})'\mathbf{A}(\boldsymbol{\theta}_0)(\mathbf{t}-\boldsymbol{\theta})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} \qquad (2.23)\\
&\qquad+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{A}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}}\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t}.
\end{aligned}$$

By Lemma 2.2 (i) and (iv), $\mathbf{IV} = [\mathbf{A}(\boldsymbol{\theta}_0)]_{r,s} + o(n^{-1/2})$. Combining the approximations for $\mathbf{I} - \mathbf{IV}$, we get (2.20).

Next let us prove $\hat{\mathbf{V}}_n(\hat{\theta}_n) \xrightarrow{p} \mathbf{V}(\theta_0)$ by showing, componentwise,

$$[\hat{\mathbf{V}}_n(\boldsymbol{\theta})]_{r,s} = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1) \tag{2.24}$$

uniformly over $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O(n^{-1/2})$ for $r, s = 1, ..., d$.

Define

$$q(\mathbf{u}, \tilde{\mathbf{u}}; \boldsymbol{\theta}) = I_{[y>\tilde{y}]} I_{[(\mathbf{x}-\tilde{\mathbf{x}})'\boldsymbol{\beta}>0]} + I_{[y<\tilde{y}]} I_{[(\mathbf{x}-\tilde{\mathbf{x}})'\boldsymbol{\beta}<0]},$$

where $\mathbf{u} = (y, \mathbf{x})$ and $\tilde{\mathbf{u}} = (\tilde{y}, \tilde{\mathbf{x}})$. In addition, let $\tau_n(\mathbf{u}, \boldsymbol{\theta}) = \int q(\mathbf{u}, \tilde{\mathbf{u}}; \boldsymbol{\theta}) \mathbb{F}_n(d\tilde{\mathbf{u}})$, where $\mathbb{F}_n(\cdot)$ is the empirical distribution for $\mathbf{u}_i$'s. By definition,

$$\begin{aligned} [\hat{\mathbf{V}}_n(\boldsymbol{\theta})]_{r,s} = \frac{1}{n} \sum_{i=1}^{n} & \left[\frac{\partial}{\partial \theta_r} \int \tau_n(\mathbf{u}_i, \boldsymbol{\theta} + \frac{\mathbf{z}}{\sqrt{n}}) (2\pi)^{-\frac{d}{2}} e^{-\frac{\|\mathbf{z}\|_2^2}{2}} d\mathbf{z} \right] \\ & \times \left[\frac{\partial}{\partial \theta_s} \int \tau_n(\mathbf{u}_i, \boldsymbol{\theta} + \frac{\tilde{\mathbf{z}}}{\sqrt{n}}) (2\pi)^{-\frac{d}{2}} e^{-\frac{\|\tilde{\mathbf{z}}\|_2^2}{2}} d\tilde{\mathbf{z}} \right]. \end{aligned}$$

Letting $\mathbf{t} = \boldsymbol{\theta} + \mathbf{z}/\sqrt{n}$ and $\boldsymbol{\omega} = \boldsymbol{\theta} + \tilde{\mathbf{z}}/\sqrt{n}$, we have

$$\begin{aligned} [\hat{\mathbf{V}}_n(\boldsymbol{\theta})]_{r,s} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_r} \int \tau_n(\mathbf{u}_i, \mathbf{t}) K_n(\mathbf{t}, \boldsymbol{\theta}) d\mathbf{t} \times \frac{\partial}{\partial \theta_r} \int \tau_n(\mathbf{u}_i, \boldsymbol{\omega}) K_n(\boldsymbol{\omega}, \theta) d\boldsymbol{\omega} \\ &= \frac{1}{n} \sum_{i=1}^{n} \int \tau_n(\mathbf{u}_i, \mathbf{t}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) d\mathbf{t} \times \int \tau_n(\mathbf{u}_i, \boldsymbol{\omega}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\boldsymbol{\omega} \\ &= \int G_n(\mathbf{t}, \boldsymbol{\omega}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\mathbf{t} d\boldsymbol{\omega}, \end{aligned}$$

where $G_n(\mathbf{t}, \boldsymbol{\omega}) = \frac{1}{n} \sum_{i=1}^{n} \tau_n(\mathbf{u}_i, \mathbf{t}) \tau_n(\mathbf{u}_i, \boldsymbol{\omega})$, which is bounded by 0 and 1. By Lemma 2.2 (vii),

$$\begin{aligned} [\hat{\mathbf{V}}_n(\boldsymbol{\theta})]_{r,s} &= o(n^{-\frac{1}{2}}) \\ &\quad + \int_{\boldsymbol{\Omega}_{n,r} \times \boldsymbol{\Omega}_{n,s}} G_n(\mathbf{t}, \boldsymbol{\omega}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\mathbf{t} d\boldsymbol{\omega} \end{aligned} \tag{2.25}$$

uniformly over $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| = O(n^{-\frac{1}{2}})$. Let $f(\mathbf{u},\mathbf{v},\mathbf{w};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = q(\mathbf{u},\mathbf{v};\boldsymbol{\theta}_1)\times q(\mathbf{u},\mathbf{w};\boldsymbol{\theta}_2)$ and $f^*(\mathbf{u},\mathbf{v},\mathbf{w};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, the symmetrized f. By definition,

$$\begin{aligned} G_n(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = &\frac{1}{\binom{n}{3}}\sum_{i<j<k} f^*(\mathbf{u}_i,\mathbf{u}_j,\mathbf{u}_k;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ &+\frac{1}{n}\times\frac{1}{\binom{n}{2}}\sum_{i<j} f^*(\mathbf{u}_i,\mathbf{u}_j,\mathbf{u}_j;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \triangleq U_n + \frac{1}{n}\tilde{U}_n. \end{aligned} \tag{2.26}$$

Clearly U_n is a third-order U-statistics and $\tilde{U}_n$ is a second-order U-statistics. Applying Hoeffding's decomposition (van der Vaart, 1998, section 12.3),

$$U_n = \sum_{c=0}^{3}\binom{3}{c}U_{n,c}, \tag{2.27}$$

where $U_{n,c}$ is a U-statistics of order c ($c = 0,1,2,3$) and defined as

$$U_{n,c} = \frac{1}{\binom{3}{c}}\sum_{|B|=c}\frac{1}{\binom{n}{3}}\sum_{i} P_B\left[f^*(\mathbf{u}_{i_1},\mathbf{u}_{i_2},\mathbf{u}_{i_3};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right].$$

Here, adopting the notations from van der Vaart (1998, Section 11.4), we define $P_B\left[f^*(\mathbf{u}_{i_1},\mathbf{u}_{i_2},\mathbf{u}_{i_3};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]$ as a projection of f^* such that

$$\begin{aligned} P_{\o}f^* &= Ef^*, \\ P_{\{i\}}f^* &= E[f^*|\mathbf{u}_i] - Ef^*, \\ P_{\{i,j\}}f^* &= E[f^*|\mathbf{u}_i,\mathbf{u}_j] - E[f^*|\mathbf{u}_i] - E[f^*|\mathbf{u}_j] + Ef^*, \\ P_{\{1,2,3\}}f^* &= E[f^*|\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3] - \sum_{i\neq j}E[f^*|\mathbf{u}_i,\mathbf{u}_j] + \sum_{i=1,2,3}E[f^*|\mathbf{u}_i] - Ef^*. \end{aligned}$$

We know from Hoeffding's decomposition that $U_{n,2}$ and $U_{n,3}$ are second- and third-order degenerated U-statistics with bounded kernels and thus of order $o_p(n^{-1})$ and $o_p(n^{-3/2})$; see Sherman (1994b, Corollary 8). Therefore, by Lemma 2.2(vi),

$$\int_{\boldsymbol{\Omega}_{n,r}\times\boldsymbol{\Omega}_{n,s}} U_{n,c}\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} = o_p(1), \text{ for } c = 2,3. \tag{2.28}$$

Replacing $U_{n,c}$ by $\tilde{U}_n/n$ in (2.28) also results in $o_p(1)$. Then combining this and (2.28) with (2.26) and (2.27), (2.25) reduces to

$$\begin{aligned}[\hat{\mathbf{V}}_n(\boldsymbol{\theta})]_{r,s} = 3 \times \int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}} & U_{n,1} \times \dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} \\ + \int_{\boldsymbol{\Omega}_{n,r}\cap\boldsymbol{\Omega}_{n,s}} & Ef \times \dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} + o_p(1).\end{aligned} \tag{2.29}$$

Let $f_1(\mathbf{u}_j;\mathbf{t},\boldsymbol{\omega}) = E[f(\mathbf{u},\mathbf{v},\mathbf{w};\mathbf{t},\boldsymbol{\omega})|\mathbf{u}=\mathbf{u}_j]$, $f_2(\mathbf{v}_j;\mathbf{t},\boldsymbol{\omega}) = E[f(\mathbf{u},\mathbf{v},\mathbf{w};\mathbf{t},\boldsymbol{\omega})|\mathbf{v}=\mathbf{v}_j]$ and $f_3(\mathbf{w}_j;\mathbf{t},\boldsymbol{\omega}) = E[f(\mathbf{u},\mathbf{v},\mathbf{w};\mathbf{t},\boldsymbol{\omega})|\mathbf{w}=\mathbf{w}_j]$. Define $\tilde{G}_n(\mathbf{t},\boldsymbol{\omega}) = \frac{1}{n}\sum_{j=1}^{n} f_1(\mathbf{u}_j;\mathbf{t},\boldsymbol{\omega})$. By the definitions of $f(\mathbf{u},\mathbf{v},\mathbf{w};\mathbf{t},\boldsymbol{\omega})$ and $q(\mathbf{u},\mathbf{v};\boldsymbol{\theta})$, we have $\tilde{G}_n(\mathbf{t},\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\tau(\mathbf{u}_i,\mathbf{t})\tau(\mathbf{u}_i,\boldsymbol{\omega})$. By Lemma 2.3 and applying integration by parts twice,

$$\begin{aligned}\int_{\boldsymbol{\Omega}_{n,r}\times\boldsymbol{\Omega}_{n,s}} & \tilde{G}_n(\mathbf{t},\boldsymbol{\omega})\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} = o(n^{-1/2}) \\ & + \int_{\tilde{\boldsymbol{\Omega}}_{n,r}\times\tilde{\boldsymbol{\Omega}}_{n,s}} \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\tau(\mathbf{u}_i,\boldsymbol{\theta}+\frac{\mathbf{z}}{\sqrt{n}})}{\partial\theta_r}\frac{\partial\tau(\mathbf{u}_i,\boldsymbol{\theta}+\frac{\tilde{\mathbf{z}}}{\sqrt{n}})}{\partial\theta_s}\right\}\prod_i d\Phi(z_i)d\Phi(\tilde{z}_i),\end{aligned}$$

where $\tilde{\boldsymbol{\Omega}}_{n,r} := \{\mathbf{z} : z_r^2 < 4\log n, \sum_{i\neq r} z_i^2 < 2(d-1)\log n\}$. By Lemma 2.3,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_r}\tau(\mathbf{u}_i,\boldsymbol{\theta}_1^*)\frac{\partial}{\partial\theta_s}\tau(\mathbf{u}_i,\boldsymbol{\theta}_2^*) = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1)$$

uniformly over $\{(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*) : \|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_0\|_2 = o(1), i = 1,2\}$. Therefore,

$$\int_{\boldsymbol{\Omega}_{n,r}\times\boldsymbol{\Omega}_{n,s}} \frac{1}{n}\sum_{j=1}^{n} f_1(\mathbf{u}_j;\mathbf{t},\boldsymbol{\omega}) \times \dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1).$$

Similarly, applying integration by parts and by Lemma 2.3 and 2.2(vi), we have

$$\int_{\boldsymbol{\Omega}_{n,r}\times\boldsymbol{\Omega}_{n,s}} \frac{1}{n}\sum_{j=1}^{n} f_2(\mathbf{v}_j;\mathbf{t},\boldsymbol{\omega}) \times \dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1),$$

$$\int_{\boldsymbol{\Omega}_{n,r}\times\boldsymbol{\Omega}_{n,s}} \frac{1}{n}\sum_{j=1}^{n} f_3(\mathbf{w}_j;\mathbf{t},\boldsymbol{\omega}) \times \dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1),$$

$$\int_{\mathbf{\Omega}_{n,r}\times\mathbf{\Omega}_{n,s}} E[f(\mathbf{u},\mathbf{v},\mathbf{w};\mathbf{t},\boldsymbol{\omega})]\times\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega}=[\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s}+o_p(1).$$

Hence the right hand side of (2.29) is $[\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s}+o_p(1)$, which gives (2.24). From (2.17) and (2.24), $\hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n)\xrightarrow{p}\mathbf{D}_0$. □

Proof of Theorem 2.3. From Theorem 2.2, we know that $\hat{\mathbf{\Sigma}}_n^{(1)}\xrightarrow{p}\mathbf{D}_0$ and $\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{D}_0)\xrightarrow{p}\mathbf{D}_0$. By the mean value theorem,

$$\begin{aligned}[\hat{\mathbf{\Sigma}}_n^{(2)}-\mathbf{D}_0]_{r,s}&=[\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\hat{\mathbf{\Sigma}}_n^{(1)})-\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{D}_0)]_{r,s}+[\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{D}_0)-\mathbf{D}_0]_{r,s}\\&=\left[\frac{\partial}{\partial\mathbf{\Sigma}}[\hat{\mathbf{D}}_n]_{r,s}\bigg|_{\mathbf{\Sigma}=\mathbf{\Sigma}^*}\right]'\times vech(\hat{\mathbf{\Sigma}}_n^{(1)}-\mathbf{D}_0)+[\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{D}_0)-\mathbf{D}_0]_{r,s},\end{aligned}$$

where $\|\mathbf{\Sigma}^*-\mathbf{D}_0\|\le\|\hat{\mathbf{\Sigma}}_n^{(1)}-\mathbf{D}_0\|$ and thus $\mathbf{\Sigma}^*\in\mathcal{N}(\mathbf{D}_0)$. In view of Lemma 2.4 and $\hat{\mathbf{\Sigma}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{D}_0)\xrightarrow{p}\mathbf{D}_0$, $\hat{\mathbf{\Sigma}}_n^{(2)}\xrightarrow{p}\mathbf{D}_0$. Again by the mean value theorem,

$$[\hat{\mathbf{\Sigma}}_n^{(k+1)}-\hat{\mathbf{\Sigma}}_n^{(1)}]_{r,s}=\left[\frac{\partial}{\partial\mathbf{\Sigma}}[\hat{\mathbf{D}}_n]_{r,s}\bigg|_{\mathbf{\Sigma}=\mathbf{\Sigma}^*}\right]'\times vech(\hat{\mathbf{\Sigma}}_n^{(k)}-\mathbf{D}_0),$$

where $\|\mathbf{\Sigma}^*-\mathbf{D}_0\|\le\|\hat{\mathbf{\Sigma}}_n^{(k)}-\mathbf{D}_0\|$. Then by Lemma 2.4 and mathematical induction, we know that for any $\epsilon>0$ and $\eta>0$, there exist K and N, such that for any $n>N$ and $k>K$,

$$P\left(\left|[\hat{\mathbf{\Sigma}}_n^{(k+1)}-\hat{\mathbf{\Sigma}}_n^{(k)}]_{r,s}\right|\le\eta\times\left|[\hat{\mathbf{\Sigma}}_n^{(k)}-\hat{\mathbf{\Sigma}}_n^{(k-1)}]_{r,s}\right|,\text{ for all }k>K\right)>1-\epsilon,$$

where $1\le s,r\le d$. Note that the inequality inside the above probability implies that $\hat{\mathbf{\Sigma}}_n^{(k)}$ converges as $k\to\infty$ and the limit $\mathbf{\Sigma}_n^*$ satisfies $\mathbf{\Sigma}_n^*=\hat{\mathbf{D}}_n(\hat{\boldsymbol{\theta}}_n,\mathbf{\Sigma}_n^*)$ and $\mathbf{\Sigma}_n^*\xrightarrow{p}\mathbf{D}_0$. □

2.3 Extensions

In this section, the approach is extended to the partial rank correlation (PRC) criterion function Q_n^*, defined by (3), of Khan and Tamer (2007) for censored data.

Under the usual conditional independence between failure and censoring times given covariates and additional regularity conditions, Khan and Tamer (2007) developed asymptotic properties for PRCE that are parallel to those by Sherman (1993).

The same self-induced smoothing can be applied to partial rank correlation criteria function to get

$$\begin{aligned}\tilde{Q}_n^*(\boldsymbol{\theta}) &= E_{\mathbf{Z}} Q_n^*(\boldsymbol{\theta} + \mathbf{Z}/\sqrt{n}) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I[\tilde{Y}_i > \tilde{Y}_j] \Phi\left(\sqrt{n} \mathbf{X}_{ij}' \boldsymbol{\beta}(\boldsymbol{\theta}) / \sigma_{ij}\right).\end{aligned} \tag{2.29}$$

Define its maximizer, $\tilde{\boldsymbol{\theta}}_n^*$, as the smoothed partial rank correlation estimator (SPRCE). Let

$$\hat{\mathbf{A}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{2n(n-1)} \sum_{i \neq j} \left\{ H_{ij} \times \dot{\phi}\left(\frac{\sqrt{n}\mathbf{X}_{ij}'\boldsymbol{\beta}}{\sigma_{ij}}\right) \left[\frac{\sqrt{n}\mathbf{X}_{ij}^{(1)}}{\sigma_{ij}}\right]^{\otimes 2} \right\}, \tag{2.30}$$

$$\hat{\mathbf{V}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{n^3} \sum_{i=1}^{n} \left\{ \sum_j \left[H_{ij} \times \phi\left(\frac{\sqrt{n}\mathbf{X}_{ij}'\boldsymbol{\beta}}{\sigma_{ij}}\right) \frac{\sqrt{n}\mathbf{X}_{ij}^{(1)}}{\sigma_{ij}} \right] \right\}^{\otimes 2}, \tag{2.31}$$

$$\hat{\mathbf{D}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = [\hat{\mathbf{A}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma})]^{-1} \times \hat{\mathbf{V}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \times [\hat{\mathbf{A}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma})]^{-1}, \tag{2.32}$$

where $H_{ij} = \Delta_j \times I[\tilde{Y}_i > \tilde{Y}_j] - \Delta_i \times I[\tilde{Y}_j > \tilde{Y}_i]$.

Based on $\hat{\mathbf{D}}_n^*(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, we have the following iterative algorithm to compute the SPRCE and variance estimate simultaneously.

Algorithm 2.2. *(SPRCE)*

1. *Compute the PRC estimator* $\hat{\boldsymbol{\theta}}_n^*$ *and set* $\hat{\boldsymbol{\Sigma}}^{(0)}$ *to be the identity matrix.*

2. *Update variance-covariance matrix* $\hat{\boldsymbol{\Sigma}}_n^{*(k)} = \hat{\mathbf{D}}_n^*(\hat{\boldsymbol{\theta}}_n^*, \hat{\boldsymbol{\Sigma}}_n^{*(k-1)})$. *Smooth the partial rank correlation* $Q_n^*(\boldsymbol{\theta})$ *using covariance matrix* $\hat{\boldsymbol{\Sigma}}_n^{*(k)}$. *Maximize the resulting smoothed partial rank correlation to get an estimator* $\hat{\boldsymbol{\theta}}_n^{*(k)}$.

3. Repeat step 2 until $\hat{\boldsymbol{\theta}}_n^{*(k)}$ *converge.*

In addition to Assumptions 2.1-2.3, Khan and Tamer (2007) added the following assumption for the consistency of PRCE.

Assumption 2.4. Let $\mathbf{S}_X$ be the support of $\mathbf{X}_i$, and $\mathfrak{X}_{uc}$ be the set

$$\mathfrak{X}_{uc} = \{\mathbf{x} \in \mathbf{S}_X : P(\Delta_i = 1 | \mathbf{X}_i = \mathbf{x}) > 0\}.$$

Then $P(\mathfrak{X}_{uc}) > 0$.

Similar to the rank correlation function, it can be shown that under Assumptions 2.1-2.4, (2.9) and (2.11) still hold for partial rank correlation function $Q_n^*(\boldsymbol{\theta})$. Therefore, Theorems 2.2.1-3 in Section 2.2 continue to hold when replacing the point and variance estimators for smoothed rank correlation by the corresponding ones for the smoothed partial rank correlation. Specifically, for any positive definite matrix $\boldsymbol{\Sigma}$, under Assumptions 2.1-2.4, we have

1. The SPRCE $\widetilde{\boldsymbol{\theta}}_n^*$ is asymptotically equivalent to the PRCE $\hat{\boldsymbol{\theta}}_n^*$ in the sense that $\widetilde{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}}_n^* + o_p(n^{-1/2})$, and, therefore,

   $$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n^* - \theta_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{D}_0^*),$$

   where $\mathbf{D}_0^*$ is the limiting variance-covariance matrix of $\hat{\boldsymbol{\theta}}_n^*$.

2. Variance estimator is consistent: $\hat{\mathbf{D}}_n^*(\hat{\boldsymbol{\theta}}_n^*, \boldsymbol{\Sigma}) \xrightarrow{p} \mathbf{D}_0^*$.

3. Algorithm 2.2 converges numerically in the sense that there exist $\boldsymbol{\Sigma}_n^*$, $n \geq 1$, such that for any $\epsilon > 0$, there exists N, such that for all $n > N$, $P(\lim_{k\to\infty} \hat{\boldsymbol{\Sigma}}_n^{*(k)} = \boldsymbol{\Sigma}_n^*, \quad \|\boldsymbol{\Sigma}_n^* - \mathbf{D}_0^*\| < \epsilon) > 1 - \epsilon$.

The proofs are similar to those of Theorems 2.2.1-3 in Section 2.2, and are, therefore, omitted.

Table 2.1: Regression Analysis of PBC data

	Albumin	SE
SPRCE	-4.29	1.40
PRCE	-3.50	-
Cox	-3.04	0.60

2.4 Numerical Results

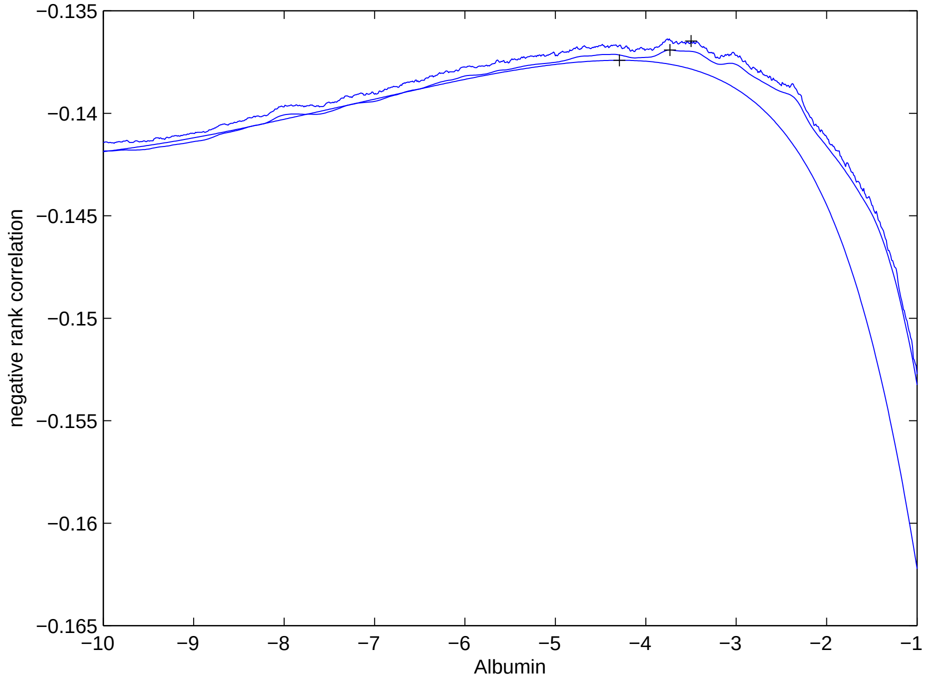
In this section, we first apply the proposed self-induced smoothing method to analyze the primary biliary cirrhosis (PBC) data (Fleming and Harrington, 1990, Appendix D) and compare the result with that using the Cox regression. Then the results from several simulation studies are reported by using the method.

2.4.1 PBC data

We apply smoothed PRCE to the survival times of the first 312 subjects with no missing covariates in the PBC data. Two covariates albumin and age50 (age divided by 50) are included. We reparameterize the transformation model (1) by setting β_{age50} as 1, and estimated $\theta_{albumin}$ by SPRCE. We also calculate PRCE for $\theta_{albumin}$ and fitted the standard Cox model. For the Cox regression, the ratio $\hat{\beta}_{albumin}/\hat{\beta}_{age50}$ is the estimate of $\theta_{albumin}$. The results are summarized in Table 2.1. Note that PRCE does not have a readily available standard error estimate. The standard error of $\hat{\beta}_{albumin}/\hat{\beta}_{age50}$ in the Cox model is estimated by the delta method. Estimates from both the SPRCE and the Cox model conclude that the ratio of $\beta_{albumin}$ to β_{age50} is significant.

To further assess the self-induced smoothing procedure, we plot the original objective function as well as the smoothed one in the first and last steps of our algorithm,

Figure 2.1: The smoothed rank correlation for PBC data



as shown in Figure 2.1. The top curve is the original objective function, the middle curve is one after the initial smoothing, and the bottom curve is the limit of the iterative algorithm (after 8 iterations). It appears that the one-step smoothed objective function is under-smoothed in terms of the level of fluctuations, and the limiting curve is quite smooth.

2.4.2 Simulation studies

We conduct simulation studies for a number of cases. In the first case (Design I), we generate $\mathbf{X}$ from a bivariate normal distribution with mean $[-10, 20]'$ and a covariance matrix diag$\{3^2, 2^2\}$. Then set $\boldsymbol{\beta}_0^T = (\theta, 1) = [1.6, 1]$ and generate ϵ from the proba-

Table 2.2: The proportional hazard model without censoring

$n = 500$	Est	Mean	Bias	RMSE	SE	coverage
θ	SMRCE	1.601	1.2×10^{-3}	0.0298	0.0316	92.3%
	MRCE	1.601	0.8×10^{-3}	0.0340	-	-
	Cox	1.599	-0.9×10^{-3}	0.0200	-	-

$n = 1000$	Est	Mean	Bias	RMSE	SE	coverage
θ	SMRCE	1.601	1.0×10^{-3}	0.0193	0.0212	93.9%
	MRCE	1.600	0.2×10^{-3}	0.0225	-	-
	Cox	1.600	0.1×10^{-3}	0.0141	-	-

$n = 2000$	Est	Mean	Bias	RMSE	SE	coverage
θ	SMRCE	1.600	0.2×10^{-3}	0.0136	0.0144	94.9%
	MRCE	1.600	-0.1×10^{-3}	0.0158	-	-
	Cox	1.600	0.1×10^{-3}	0.0100	-	-

Table 2.3: The proportional hazard model with censoring

$n = 600$	Est	Mean	Bias	RMSE	SE	coverage
θ	SPRCE	1.604	3.7×10^{-3}	0.0282	0.0300	93.2%
	PRCE	1.603	2.9×10^{-3}	0.0327	-	-
	Cox	1.601	1.0×10^{-3}	0.0204	-	-

$n = 1200$	Est	Mean	Bias	RMSE	SE	coverage
θ	SPRCE	1.601	1.1×10^{-3}	0.0190	0.0201	93.9%
	PRCE	1.601	0.8×10^{-3}	0.0217	-	-
	Cox	1.600	-0.2×10^{-3}	0.0139	-	-

$n = 2400$	Est	Mean	Bias	RMSE	SE	coverage
θ	SPRCE	1.600	0.4×10^{-3}	0.0127	0.0136	95.4%
	PRCE	1.600	0.1×10^{-3}	0.0148	-	-
	Cox	1.600	-0.2×10^{-3}	0.0097	-	-

Table 2.4: The linear model with gaussian noise

$n = 250$	Est	Mean	Bias	RMSE	SE	coverage
θ_1	SMRCE	1.615	1.5×10^{-2}	0.0747	0.0756	91.7%
	MRCE	1.612	1.2×10^{-2}	0.0730	-	-
	LS	1.601	0.7×10^{-3}	0.0296	-	-
θ_2	SMRCE	.5042	0.4×10^{-2}	0.0427	0.0443	93.6%
	MRCE	.5058	0.5×10^{-2}	0.0423	-	-
	LS	.5006	0.6×10^{-3}	0.0354	-	-

$n = 500$	Est	Mean	Bias	RMSE	SE	coverage
θ_1	SMRCE	1.605	4.9×10^{-3}	0.0515	0.0513	92.7%
	MRCE	1.607	6.7×10^{-3}	0.0523	-	-
	LS	1.601	0.7×10^{-3}	0.021	-	-
θ_2	SMRCE	.5023	2.3×10^{-3}	0.0296	0.0302	94.6%
	MRCE	.5042	4.2×10^{-3}	0.0316	-	-
	LS	.5006	0.6×10^{-3}	0.0254	-	-

$n = 1000$	Est	Mean	Bias	RMSE	SE	coverage
θ_1	SMRCE	1.603	3.6×10^{-3}	0.0361	0.0348	92.4%
	MRCE	1.603	3.4×10^{-3}	0.0382	-	-
	LS	1.601	0.5×10^{-3}	0.0144	-	-
θ_2	SMRCE	.5009	0.9×10^{-3}	0.0203	0.0207	94.8%
	MRCE	.5018	1.8×10^{-3}	0.0214	-	-
	LS	.5004	0.4×10^{-3}	0.0176	-	-

bility density function $f(w) = 2\exp(2w - \exp(2w))$. Set the transformation $H(x)$ as $H^{-1}(y) = log(y^2)$. This is indeed a Weibull proportional hazard model. The sample sizes are $n = 500, 1000, 2000$ and the number of replications is 500. The SMRCE, MRCE and Cox model are used to estimate θ, and the standard error of SMRCE is computed by Algorithm 2.1. The mean(Mean), bias(Bias) and root mean square error(RMSE) for each method as well as mean of standard error(SE) and coverage of 95% confidence interval for the SMRCE are reported in Table 2.2.

The second case (Design II) is similar to the first one except that Y is censored by a random variable C, which is independent of $\mathbf{X}$ and normally distributed with mean $\mu = 9.2$ and variance $\sigma^2 = 0.5^2$. The sample sizes are $n = 600, 1200, 2400$ and the number of replications is 500. This design is similar to that in Gørgens and Horowitz (1999). The SPRCE, PRCE and Cox model are used to estimate θ, and the standard error of SPRCE is computed by Algorithm 2.2. The resulting estimates are summarized in Table 2.3 where we also report bias(Bias), root mean square error(RMSE), mean of standard error(SE), and coverage of 95% confidence interval.

In the third case (Design III), we generate $\mathbf{X} = [X_1, X_2, X_3]'$ by two steps. Fist of all, generate $[X_1, X_3]'$ from a bivariate normal distribution with mean $[-2, 2]'$ and an identity covariance matrix. Then generate X_3 as 0 or 2 with equal probability. Set $\boldsymbol{\beta}_0^T = (\theta_1, \theta_2, 1) = [1.6, 0.5, 1]$ and generate ϵ from a normal distribution with $\mu = 0$ and $\sigma^2 = 0.5^2$. Set the transformation $H(x) = x$. The sample sizes are $n = 250, 500, 1000$ and the number of replications is 500. The SMRCE, MRCE and least squared method are applied to estimate θ_1 and θ_2, and the standard error of SMRCE is computed by Algorithm 2.1. Table 2.4 reports the mean (Mean), bias (Bias) and root mean square error (RMSE) for each method as well as mean of standard error (SE) and coverage of 95% confidence interval for the SMRCE.

From Tables 2.2, 2.3 and 2.4, we find that (1) the root mean squared error is close to the mean standard error for the SMRCE (SPRCE); (2) as the sample size increases, the bias reduces and the coverage of 95% confidence interval converges to the nominal level. These show that the proposed variance estimator is accurate and Algorithms 2.1 and 2.2 work well.

2.5 Other Proofs

2.5.1 Lemmas and corollaries

Lemma 2.1 below is due to Sherman (1993, Theorem 2).

Lemma 2.1. *Denote $\Gamma_n(\boldsymbol{\theta})$ as general objective functions which are centered and satisfies the same regularity conditions as in Sherman (1993). Suppose $\boldsymbol{\theta}_n := argmax_{\boldsymbol{\Theta}} \Gamma_n(\boldsymbol{\theta})$ is consistent for $\boldsymbol{\theta}_0$, an interior point of $\boldsymbol{\Theta}$. Suppose also that uniformly over $o_p(1)$ neighborhoods of $\boldsymbol{\theta}_0$,*

$\Gamma_n(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{W}_n + o_p(|\boldsymbol{\theta}-\boldsymbol{\theta}_0|^2) + o_p(1/n)$ *where* $\mathbf{A}$ *is a negative definite matrix, and $\mathbf{W}_n$ converges in distribution to a $N(\mathbf{0}, \mathbf{V})$ random vector. Then*

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = -\mathbf{A}^{-1}\mathbf{W}_n + o_p(1) \xrightarrow{L} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{V}\mathbf{A}^{-1}).$$

Recall in Theorem 2.2, define $K_n(\mathbf{t}, \boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}} n^{\frac{d}{2}} \exp(-\frac{n\|\mathbf{t}-\boldsymbol{\theta}\|_2^2}{2})$ and its first and second partial derivatives with respect to $\boldsymbol{\theta}$ as

$$\begin{aligned}
\dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) &= (\frac{2\pi}{n})^{-\frac{d}{2}} n(t_r - \theta_r) e^{-\frac{n\|\mathbf{t}-\theta\|_2^2}{2}}, \\
\ddot{K}_{n,r,r}(\mathbf{t}, \boldsymbol{\theta}) &= (\frac{2\pi}{n})^{-\frac{d}{2}} n(n(t_r - \theta_r)^2 - 1) e^{-\frac{n\|\mathbf{t}-\theta\|_2^2}{2}}, \\
\ddot{K}_{n,r,s}(\mathbf{t}, \boldsymbol{\theta}) &= (\frac{2\pi}{n})^{-\frac{d}{2}} n^2(t_r - \theta_r)(t_s - \theta_s) e^{-\frac{n\|\mathbf{t}-\theta\|_2^2}{2}}.
\end{aligned}$$

Also recall

$$\mathbf{\Omega}_{n,r} = \{\mathbf{t} : (t_r - \theta_r)^2 < 4\log n/n, \sum_{i\neq r}(t_i - \theta_i)^2 < 2(d-1)\log n/n\}.$$

Then we have the following lemma.

Lemma 2.2. *Uniformly over* $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 = O(n^{-\frac{1}{2}})$,

(i) $\displaystyle\int_{(\mathbf{\Omega}_{n,r}\cap\mathbf{\Omega}_{n,s})^c} F(\mathbf{t})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = o(\frac{1}{\sqrt{n}}), \forall F(\mathbf{t})$ *s.t.* $0 \le F(\mathbf{t}) \le 1.$

(ii) $\displaystyle\int_{\mathbf{\Omega}_{n,r}\cap\mathbf{\Omega}_{n,s}} \frac{1}{n}|\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = O(1).$

(iii) $\displaystyle\int_{\mathbf{\Omega}_{n,r}\cap\mathbf{\Omega}_{n,s}} \ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = o(n^{-1/2}).$

(iv) $\displaystyle\int_{\mathbf{\Omega}_{n,r}\cap\mathbf{\Omega}_{n,s}} \frac{1}{2}(\mathbf{t}-\boldsymbol{\theta})'\mathbf{A}(\mathbf{t}-\boldsymbol{\theta})\ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = [A]_{r,s} + o(n^{-1}).$

(v) $\displaystyle\int_{\mathbf{\Omega}^c_{n,r}} |\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = O(n^{-3/2}).$

(vi) $\displaystyle\int_{\mathbf{\Omega}_{n,r}} \frac{1}{\sqrt{n}}|\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = O(1).$

(vii) *For any given* $0 \le G(\mathbf{t},\boldsymbol{\omega}) \le 1$,

$$\int G(\mathbf{t},\boldsymbol{\omega})\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} \\ = \int_{\mathbf{\Omega}_{n,r}\times\mathbf{\Omega}_{n,s}} G(\mathbf{t},\boldsymbol{\omega})\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})\dot{K}_{n,s}(\boldsymbol{\omega},\boldsymbol{\theta})d\mathbf{t}d\boldsymbol{\omega} + o(n^{-\frac{1}{2}}).$$

Proof. Let $\tilde{\mathbf{\Omega}}_{n,r} = \left\{\mathbf{t} : t_r^2 < 4\log n/n, \sum_{i\neq r} t_i^2 < 2(d-1)\log n/n\right\}$, and divide its complement into $\tilde{\mathbf{\Omega}}^{(1)}_{n,r} := \left\{\mathbf{t} : t_r^2 > 4\log n/n\right\}$ and $\tilde{\mathbf{\Omega}}^{(2)}_{n,r} := \{\mathbf{t} : t_r^2 < 4\log n/n,$ $\sum_{i\neq r} t_i^2 \ge 2(d-1)\log n/n\right\}$. Let us prove (i)-(iv) for $s = r$. For $s \neq r$, the proofs are similar and omitted.

For (i), note that

$$\int_{\mathbf{\Omega}_{n,r}^c} F(\mathbf{t})\ddot{K}_{n,r,r}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = \int_{\tilde{\mathbf{\Omega}}_{n,r}^c} F(\mathbf{t}+\boldsymbol{\theta})\ddot{K}_{n,r,r}(\mathbf{t},\mathbf{0})d\mathbf{t}.$$

Since $0 \leq F(\mathbf{t}) \leq 1$ and $(nt_r^2-1)I[\tilde{\mathbf{\Omega}}_{n,r}^{(1)}] \geq 0$,

$$\begin{aligned}
\int_{\tilde{\mathbf{\Omega}}_{n,r}^{(1)}} F(\mathbf{t}+\boldsymbol{\theta})\ddot{K}_{n,r,r}(\mathbf{t},\mathbf{0})d\mathbf{t} = &\int_{\tilde{\mathbf{\Omega}}_{n,r}^{(1)}} F(\mathbf{t}+\boldsymbol{\theta})(\frac{2\pi}{n})^{-\frac{d}{2}} n(nt_r^2-1)e^{-\frac{n\|\mathbf{t}\|_2^2}{2}}d\mathbf{t} \\
\leq &(\frac{2\pi}{n})^{-\frac{d}{2}} \int_{\tilde{\mathbf{\Omega}}_{n,r}^{(1)}} n(nt_r^2-1)e^{-\frac{n\|\mathbf{t}\|_2^2}{2}}d\mathbf{t} \\
= &(2\pi)^{-\frac{1}{2}} \int_{\tilde{\mathbf{\Omega}}_{n,r}^{(1)}} n^{\frac{1}{2}} d(nt_r e^{-\frac{nt_r^2}{2}}) \prod_{i\neq r} d\Phi(\sqrt{n}t_i) = o(\frac{1}{\sqrt{n}}),
\end{aligned}$$

where the last equality follows from $0 \leq \int \prod_{i\neq r} d\Phi(\sqrt{n}t_i) \leq 1$. Similarly,

$$\begin{aligned}
&\int_{\tilde{\mathbf{\Omega}}_{n,r}^{(2)}} F(\mathbf{t}+\boldsymbol{\theta})\ddot{K}_{n,r,r}(\mathbf{t},\mathbf{0})d\mathbf{t} = \int_{\tilde{\mathbf{\Omega}}_{n,r}^{(2)}} F(\mathbf{t}+\boldsymbol{\theta})(\frac{2\pi}{n})^{-\frac{d}{2}}(n^2t_r^2-n)e^{-\frac{n\|\mathbf{t}\|_2^2}{2}}d\mathbf{t} \\
\leq &(2\pi)^{-\frac{1}{2}} \int_{\tilde{\mathbf{\Omega}}_{n,r}^{(2)}} n|nt_r^2-1|e^{-\frac{nt_r^2}{2}} d\sqrt{n}t_r \prod_{i\neq r} d\Phi(\sqrt{n}t_i) \leq 8\frac{n(\log n)^{3/2}}{n^2} = o(\frac{1}{\sqrt{n}}).
\end{aligned}$$

For (ii), by definition,

$$\begin{aligned}
\frac{1}{n}\int_{\mathbf{\Omega}_{n,r}} |\ddot{K}_{n,r,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = &(2\pi)^{-\frac{1}{2}} \int_{\mathbf{\Omega}_{n,r}} \sqrt{n}|nt_r^2-1|e^{-\frac{nt_r^2}{2}}dt_r \prod_{i\neq r} d\Phi(\sqrt{n}t_i) \\
\leq &\frac{1}{\pi\sqrt{n}}\left(\left[nt_r e^{-\frac{nt_r^2}{2}}\right]\Bigg|_{t_r=1/\sqrt{n}}^{2\sqrt{\frac{\log n}{n}}} + \left[nt_r e^{-\frac{nt_r^2}{2}}\right]\Bigg|_{t_r=1/\sqrt{n}}^{0}\right) = O(1),
\end{aligned}$$

where the inequality follows from $0 \leq \int \prod_{i\neq r} d\Phi(\sqrt{n}t_i) \leq 1$.

For (iii), by definition, $\int_{\mathbf{\Omega}_{n,r}} \ddot{K}_{n,r,r}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = \int_{\tilde{\mathbf{\Omega}}_{n,r}} (\frac{2\pi}{n})^{-\frac{d}{2}}(n^2t_r^2-n)e^{-\frac{n\|t\|_2^2}{2}}d\mathbf{t}$ $= 2n^{\frac{3}{2}}t_r e^{-\frac{nt_r^2}{2}}\Big|_{t_r=0}^{2\sqrt{\frac{\log n}{n}}} \times \int_{\tilde{\mathbf{\Omega}}_{n,r}} \prod_{i\neq r} d\Phi(\sqrt{n}t_i) = o(\frac{1}{\sqrt{n}})$, where the last equality follows from $0 \leq \int \prod_{i \neq r} d\Phi(\sqrt{n}t_i) \leq 1$.

For (iv), by definition and applying integration by parts twice,

$$
\begin{aligned}
&\int_{\mathbf{\Omega}_{n,r}} \frac{1}{2}(\mathbf{t}-\boldsymbol{\theta})'\mathbf{A}(\mathbf{t}-\boldsymbol{\theta})\ddot{K}_{n,r,r}(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} \\
=&\int_{\tilde{\mathbf{\Omega}}_{n,r}} \frac{1}{2}\mathbf{t}'\mathbf{A}\mathbf{t}(2\pi)^{-\frac{1}{2}}\sqrt{n}d(-nt_r e^{-\frac{nt_r^2}{2}})\prod_{i \neq r} d\Phi(\sqrt{n}t_i) \\
=&o(n^{-1}) + \int_{\tilde{\mathbf{\Omega}}_{n,r}} \mathbf{t}'\mathbf{A}\mathbf{e_r}(2\pi)^{-\frac{1}{2}}\sqrt{n}d(-e^{-\frac{nt_r^2}{2}})\prod_{i \neq r} d\Phi(\sqrt{n}t_i) \\
=&\mathbf{A}_{r,r} + o(n^{-1}),
\end{aligned}
$$

where $\mathbf{e_r}' = (0, ..., 0, 1, 0, ..., 0)$ with r^{th} entry being 1, and the last equality follows from the Gaussian tail probability.

For (v), we know, by definition, $\int_{\mathbf{\Omega}_{n,r}^C} |\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = \int_{\tilde{\mathbf{\Omega}}_{n,r}^C} |\dot{K}_{n,r}(\mathbf{t},\mathbf{0})|d\mathbf{t}$. By symmetry,

$$
\begin{aligned}
\int_{\tilde{\mathbf{\Omega}}_{n,r}^{(1)}} |\dot{K}_{n,r}(\mathbf{t},\mathbf{0})|d\mathbf{t} &= \frac{2}{\sqrt{2\pi}} \int_{t_r \geq 0, \tilde{\mathbf{\Omega}}_{n,r}^{(1)}} n^{\frac{1}{2}} d(e^{-\frac{nt_r^2}{2}}) \prod_{i \neq r} d\Phi(\sqrt{n}t_i) \\
&\leq \frac{\sqrt{n}}{n^2} \times 1 = O(n^{-\frac{3}{2}}),
\end{aligned}
$$

where the inequality follows from $0 \leq \int \prod_{i \neq r} d\Phi(\sqrt{n}t_i) \leq 1$. Similarly,

$$
\int_{\tilde{\mathbf{\Omega}}_{n,r}^{(2)}} |\ddot{K}_{n,r,r}(\mathbf{t},\mathbf{0})|d\mathbf{t} = \frac{2\sqrt{n}}{\sqrt{2\pi}} \int_{t_r \geq 0, \tilde{\mathbf{\Omega}}_{n,r}^{(2)}} d(e^{-\frac{nt_r^2}{2}}) \prod_{i \neq r} d\Phi(\sqrt{n}t_i) \leq \frac{\sqrt{n}}{n^2} = O(n^{-\frac{3}{2}}).
$$

For (vi), by definition,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} \int_{\mathbf{\Omega}_{n,r}} |\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta})|d\mathbf{t} = \int_{\tilde{\mathbf{\Omega}}_{n,r}} (2\pi)^{-\frac{d}{2}} n^{\frac{d-1}{2}} n|t_r| e^{-\frac{n\|t\|_2^2}{2}} d\mathbf{t} \\
=&(2\pi)^{-\frac{1}{2}} 2 \int_{t_r \geq 0, \tilde{\mathbf{\Omega}}_{n,r}} d(e^{-\frac{nt_r^2}{2}}) \prod_{i \neq r} d\Phi(\sqrt{n}t_i) = O(1),
\end{aligned}
$$

where the second equality is due to symmetry and the third equality follows from $0 \le \int \prod_{i \ne r} d\Phi(\sqrt{n} t_i) \le 1$.

To prove (vii), without loss of generality, let us assume $0 \le G(\mathbf{t}, \boldsymbol{\omega}) \le 1$. Denote $\boldsymbol{\Omega}^a_{n,r}$ as $\boldsymbol{\Omega}_{n,r}$ and $\boldsymbol{\Omega}^b_{n,r}$ its complement. Then,

$$\begin{aligned}
&\int_{\boldsymbol{\Omega}^k_{n,r} \times \boldsymbol{\Omega}^l_{n,s}} G(\mathbf{t}, \boldsymbol{\omega}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\mathbf{t} d\boldsymbol{\omega} \\
&\qquad \le \int_{\boldsymbol{\Omega}^k_{n,r}} |\dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta})| d\mathbf{t} \times \int_{\boldsymbol{\Omega}^l_{n,s}} |\dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta})| d\boldsymbol{\omega} \\
&\qquad = \int_{\tilde{\boldsymbol{\Omega}}^k_{n,r}} |\dot{K}_{n,r}(\mathbf{t}, \mathbf{0})| d\mathbf{t} \times \int_{\tilde{\boldsymbol{\Omega}}^l_{n,s}} |\dot{K}_{n,s}(\boldsymbol{\omega}, \mathbf{0})| d\boldsymbol{\omega}
\end{aligned}$$

where k and l are chosen from $\{a, b\}$. Then by (v) and (vi),

$$\begin{aligned}
&\int G(\mathbf{t}, \boldsymbol{\omega}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\mathbf{t} d\boldsymbol{\omega} \\
&\qquad = \int_{\boldsymbol{\Omega}_{n,r} \times \boldsymbol{\Omega}_{n,s}} G(\mathbf{t}, \boldsymbol{\omega}) \dot{K}_{n,r}(\mathbf{t}, \boldsymbol{\theta}) \dot{K}_{n,s}(\boldsymbol{\omega}, \boldsymbol{\theta}) d\mathbf{t} d\boldsymbol{\omega} + o(n^{-\frac{1}{2}}).
\end{aligned}$$

□

Lemma 2.3. *Uniformly over* $(\mathbf{t}, \boldsymbol{\omega})$ *such that* $\|\mathbf{t} - \boldsymbol{\theta}_0\| = o(1)$ *and* $\|\boldsymbol{\omega} - \boldsymbol{\theta}_0\| = o(1)$, *we have*

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial \tau(\mathbf{u}_i, \mathbf{t})}{\partial \theta_r} \tau(\mathbf{u}_i, \boldsymbol{\omega}) \right] = E\left[\frac{\partial \tau(\mathbf{u}, \boldsymbol{\theta}_0)}{\partial \theta_r} \tau(\mathbf{u}, \boldsymbol{\theta}_0) \right] + o_p(1), \tag{2.33}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial \tau(\mathbf{u}_i, \mathbf{t})}{\partial \theta_r} \frac{\partial \tau(\mathbf{u}_i, \boldsymbol{\omega})}{\partial \theta_s} \right] = E\left[\frac{\partial \tau(\mathbf{u}, \boldsymbol{\theta}_0)}{\partial \theta_r} \frac{\partial \tau(\mathbf{u}, \boldsymbol{\theta}_0)}{\partial \theta_s} \right] + o_p(1). \tag{2.34}$$

Proof. The main steps of the proof are scratched below.

First of all, observe that

$$\begin{aligned}\Bigg|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\tau(\mathbf{u}_i,\mathbf{t})}{\partial\theta_r}\tau(\mathbf{u}_i,\boldsymbol{\omega})\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\tau(\mathbf{u}_i,\boldsymbol{\theta}_0)}{\partial\theta_r}\tau(\mathbf{u}_i,\boldsymbol{\omega})\right]\Bigg| \\ \le \frac{1}{n}\sum_{i=1}^{n}\left[\left|\frac{\partial\tau(\mathbf{u}_i,\mathbf{t})}{\partial\theta_r} - \frac{\partial\tau(\mathbf{u}_i,\boldsymbol{\theta}_0)}{\partial\theta_r}\right| \times |\tau(\mathbf{u}_i,\boldsymbol{\omega})|\right] \\ \le \frac{1}{n}\sum_{i=1}^{n} M_2(\mathbf{u}_i) \times |\mathbf{t}-\boldsymbol{\theta}_0|,\end{aligned}$$

where $M_2(\mathbf{u})$ is an integrable function. The last inequality is due to Assumption 2.3 and $|\tau(\mathbf{u},\boldsymbol{\theta})| \le 1$.

Since $M_2(\mathbf{u})$ is integrable, by the law of large numbers, the left hand side of above inequality is thus $o_p(1)$. By a similar argument, we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\tau(\mathbf{u}_i,\mathbf{t})}{\partial\theta_r}\tau(\mathbf{u}_i,\boldsymbol{\omega})\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\tau(\mathbf{u}_i,\boldsymbol{\theta}_0)}{\partial\theta_r}\tau(\mathbf{u}_i,\boldsymbol{\theta}_0)\right] + o_p(1).$$

By the law of large numbers, we get (2.33). The proof of (2.34) is similar. □

Lemma 2.4. *Let $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{V}}_n$ be the same as those in* (2.7)*. Then, for $1 \le r, s \le d$, we have*

$$\sup_{\|\theta-\theta_0\|=o(1),\Sigma\in\mathcal{N}(D_0)}\left|\frac{\partial}{\partial\boldsymbol{\Sigma}}\right|[\hat{\mathbf{A}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})]_{r,s} = o_p(1),$$

$$\sup_{\|\theta-\theta_0\|=o(1),\Sigma\in\mathcal{N}(D_0)}\left|\frac{\partial}{\partial\boldsymbol{\Sigma}}\right|[\hat{\mathbf{V}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})]_{r,s} = o_p(1),$$

where $\mathcal{N}(\mathbf{D}_0)$ is a small neighborhood of $\mathbf{D}_0$ and $\boldsymbol{\Sigma}$ is a positive definite matrix.

Proof. Let us now extend the definition of kernels in Lemma 2.2 for any covariance matrix $\boldsymbol{\Sigma}$ as follows,

$$K_n(\mathbf{t},\boldsymbol{\theta},\boldsymbol{\Sigma}) := (\frac{2\pi}{n})^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{n}{2}(\mathbf{t}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\theta})),$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Then the first and second derivatives of K_n with respect to $\boldsymbol{\theta}$ become

$$\begin{aligned}\dot{K}_{n,r}(\mathbf{t},\boldsymbol{\theta},\boldsymbol{\Sigma}) &:= (\frac{2\pi}{n})^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}n\mathbf{e_r}'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\theta})e^{-\frac{n}{2}(\mathbf{t}-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\theta)},\\ \ddot{K}_{n,r,s}(\mathbf{t},\boldsymbol{\theta},\boldsymbol{\Sigma}) &:= (\frac{2\pi}{n})^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}n\mathbf{e_r}'\left[n\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\theta})(\mathbf{t}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\right]\mathbf{e_s}\\ &\qquad\qquad \times e^{-\frac{n}{2}(\mathbf{t}-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\theta)}.\end{aligned}$$

Partition $\mathbb{R}^d$ into $\boldsymbol{\Omega}_{n,r}$ and its complement $\boldsymbol{\Omega}_{n,r}^c$, where $\boldsymbol{\Omega}_{n,r} := \Big\{\mathbf{t} : (\mathbf{t}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\theta}) < 6d\log n/n\Big\}$. Furthermore, define $\tilde{\boldsymbol{\Omega}}_{n,r} := \Big\{\mathbf{t} : \mathbf{t}'\boldsymbol{\Sigma}^{-1}\mathbf{t} < 6d\log n/n\Big\}$.

Note that $(\mathbf{t}-\boldsymbol{\theta})(\mathbf{t}-\boldsymbol{\theta})'e^{-\frac{1}{2}(\mathbf{t}-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\theta)}$ is bounded for $\boldsymbol{\Sigma}\in\mathcal{N}(D_0)$. Similar to the proofs of Theorem 2 and Lemma 2.2, we can get

$$\frac{\partial[\hat{\mathbf{A}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})]_{r,s}}{\partial\boldsymbol{\Sigma}} = [\mathbf{A}(\boldsymbol{\theta}_0)]_{r,s}\int_{\tilde{\boldsymbol{\Omega}}_{n,r}}\frac{\partial K_n(\mathbf{t},\mathbf{0},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}d\mathbf{t}+o_p(1),$$

$$\frac{\partial[\hat{\mathbf{V}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})]_{r,s}}{\partial\boldsymbol{\Sigma}} = [\mathbf{V}(\boldsymbol{\theta}_0)]_{r,s}\int_{\tilde{\boldsymbol{\Omega}}_{n,r}}\frac{\partial K_n(\mathbf{t},\mathbf{0},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}d\mathbf{t}+o_p(1),$$

uniformly over $(\boldsymbol{\theta},\boldsymbol{\Sigma})$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| = o(1)$ and $\|\boldsymbol{\Sigma}-\mathbf{D}_0\| = o(1)$.

Likewise, we have $\int_{\tilde{\boldsymbol{\Omega}}_{n,r}^c}\frac{\partial K_n(\mathbf{t},\mathbf{0},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}d\mathbf{t} = o(1)$, which, combined with $\int\frac{\partial K_n(\mathbf{t},\mathbf{0},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}d\mathbf{t} = 0$, implies $\int_{\tilde{\boldsymbol{\Omega}}_{n,r}}\frac{\partial K_n(\mathbf{t},\mathbf{0},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}d\mathbf{t} = o(1)$. This completes the proof. □

Corollary 2.1. *For $1 \le r, s \le d$, we have*

$$\sup_{\|\theta-\theta_0\|=o(1),\Sigma\in\mathcal{N}(D_0)}\left|\frac{\partial}{\partial\boldsymbol{\Sigma}}\right|[\hat{\mathbf{A}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})^{-1}]_{r,s} = o_p(1).$$

Proof. First, by Theorem 2, Lemma 2.4 and the mean value theorem, we can show that $[\hat{\mathbf{A}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma})]_{r,s} = [\hat{\mathbf{A}}_n(\boldsymbol{\theta},\boldsymbol{\Sigma}) - \hat{\mathbf{A}}_n(\boldsymbol{\theta},\mathbf{D}_0) + \hat{\mathbf{A}}_n(\boldsymbol{\theta},\mathbf{D}_0)]_{r,s} = [\mathbf{A}(\boldsymbol{\theta}_0)]_{r,s} + o_p(1)$. By matrix differentiation, $d\mathbf{A}^{-1} = -\mathbf{A}^{-1}(d\mathbf{A})\mathbf{A}^{-1}$. Thus $\hat{\mathbf{A}}_n^{-1} - \mathbf{A}_0^{-1} = -\mathbf{A}_0^{-1}(\hat{\mathbf{A}}_n -$

$\mathbf{A}_0)\mathbf{A}_0^{-1} + o(\|\hat{\mathbf{A}}_n - \mathbf{A}_0\|_1)$, where $\mathbf{A}_0 = \mathbf{A}(\boldsymbol{\theta}_0)$. The rest of the proof is straightforward and thus omitted. □

Lemma 2.5. *For $1 \le r, s \le d$, we have*

$$\sup_{\|\theta-\theta_0\|=o(1), \Sigma \in \mathcal{N}(D_0)} \left| \frac{\partial}{\partial \boldsymbol{\Sigma}} \right| [\hat{\mathbf{D}}_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})]_{r,s} = o_p(1).$$

Proof. The result follows immediate from Lemma 2.4 and Corollary 2.1. □

2.5.2 A sufficient condition for Assumption 2.3

Suppose f is the joint density for $(\mathbf{X}, Y)$ and $f(\cdot|\mathbf{r}, s)$ is the conditional density of $X^{(2)}$ given $\mathbf{X}^{(1)} = \mathbf{r}$ and $Y = s$. Suppose $g(\cdot|s, \boldsymbol{\theta})$ is the conditional density of $\mathbf{X}'\boldsymbol{\beta}(\boldsymbol{\theta})$ given $Y = s$ and $g(\cdot|\mathbf{r}, s, \boldsymbol{\theta})$ is the conditional density of $\mathbf{X}'\boldsymbol{\beta}(\boldsymbol{\theta})$ given $\mathbf{X}^{(1)} = \mathbf{r}$ and $Y = s$. By change of variable, $g(\mathbf{t}|\mathbf{r}, s, \boldsymbol{\theta}) = f(\mathbf{t} - \mathbf{r}'\boldsymbol{\theta}|\mathbf{r}, s)$. Therefore,

$$g(\mathbf{t}|s, \boldsymbol{\theta}) = \textstyle\int g(\mathbf{t}|\mathbf{r}, s, \boldsymbol{\theta}) G_{\mathbf{X}^{(1)}|s}(d\mathbf{r}) = \int f(\mathbf{t} - \mathbf{r}'\boldsymbol{\theta}|\mathbf{r}, s) G_{\mathbf{X}^{(1)}|s}(d\mathbf{r}),$$

where $G_{\mathbf{X}^{(1)}|s}$ is the conditional distribution of $\mathbf{X}^{(1)}$ given $Y = s$. We also observe that,

$$\tau(z, \boldsymbol{\theta}) = \textstyle\int_{-\infty}^{x'\beta(\theta)} \int_{-\infty}^{y} g(\mathbf{t}|s, \boldsymbol{\theta}) G_Y(ds) d\mathbf{t} + \int_{x'\beta(\theta)}^{\infty} \int_{y}^{\infty} g(\mathbf{t}|s, \boldsymbol{\theta}) G_Y(ds) d\mathbf{t},$$

where G_Y is the marginal distribution of Y. Therefore if the conditional density $f_{\mathbf{X}^{(2)}|\mathbf{X}^{(1)},Y}(\cdot|\mathbf{r}, s)$ has bounded derivatives up to order three for each $(\mathbf{r}, s)$ in the support of space $\mathbf{X}^{(1)} \otimes Y$, it is not difficult to show that Assumption 2.3 is satisfied. The sufficient condition can be easily verified in certain common situations such as when the conditional density $f_{\mathbf{X}^{(2)}|\mathbf{X}^{(1)},Y}$ is normal.

2.6 Discussion

This chapter provides a simple yet general recipe for smoothing the discontinuous rank correlation criteria function. The smoothing is self-induced in the sense that the implied bandwidth is essentially the asymptotic standard deviation of the regression parameter estimator. It is shown that such smoothing does not introduce any significant bias in that the resulting estimator is asymptotically equivalent to the original maximum rank correlation estimator, which is asymptotically normal. The smoothed rank correlation can be used as if it were a regular smooth criterion function in the usual M-estimation problem, in the sense that the standard sandwich-type plug-in variance-covariance estimator is consistent. Simulation and real data analysis provide additional evidence that the proposed method gives the right amount of smoothing.

Chapter 3

Variable Selection

3.1 Introduction

Variable selection has become increasingly important in the regression analysis due to the growing size of data. It is common nowadays that hundreds or even thousands of explanatory variables are available for many real-life regression problems. For example, to build prediction models for 1-minute stock price data, we may use the returns on stocks as well as their lagged returns of all possible lag lengths. The large number of stocks interacting with various lag lengths results in a tremendous dataset. However, only a small portion of those variables is useful for price prediction.

In an attempt to automatically select most important variables and construct sparse prediction models, statisticians modify the least squares method by adding different kinds of penalty terms to the objective functions. This methodology is the so-called penalized least squares approach. Examples we are familiar with include the bridge regression proposed by Frank and Friedman (1993), the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996, 1997), the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) and the elastic net penalty proposed by Zou and Hastie (2005). The penalized least squares idea has been extended naturally to other likelihood-based models including the generalized linear models (Friedman, Hastie and Tibshirani, 2010) and Cox's proportional hazard model (Fan and Li, 2002).

The major difficulties of applying these regularization methods directly to selecting variables for model (1.1) are the unknown error distribution and the discreteness of the rank correlation function (2.1). We apply the self-induced smoothing method to Han's rank correlation function and add the SCAD penalty function to the smoothed rank correlation function. Through the regularized and smoothed rank correlation function, we derive a path algorithm for the variable selection problem. We use a rank correlation information criteria to select the thresholding parameter of the SCAD penalty. The resulting regularized estimator is proved to be $\sqrt{n}$-consistent and achieve the asymptotic normality under certain regularity conditions.

The rest of this chapter is organized as follows. In Section 3.2, the new algorithm is described and related large sample properties are developed. Section 3.3 contains the proofs for the theoretical results. Section 3.4 covers extensive simulation studies to compare the new method with existing ones. Section 3.5 contains the discussion and concluding remark.

3.2 Main Results

The rank correlation function $Q_n(\boldsymbol{\beta})$ is discrete in $\boldsymbol{\beta}$ due to the indicator function $I(\mathbf{X}_i'\boldsymbol{\beta} > \mathbf{X}_j'\boldsymbol{\beta})$. Therefore the usual optimization algorithms and the penalization approach cannot be applied directly to the rank correlation function. Instead, the penalty terms are introduced to the smoothed rank correlation function $\tilde{Q}_n$.

Define the regularized and smoothed rank correlation function

$$\tilde{Q}_n^{Re}(\boldsymbol{\theta}) = \tilde{Q}_n(\boldsymbol{\theta}) - \sum_{j=1}^{d} p_{\lambda_n}(|\theta_j|), \tag{3.1}$$

where $\tilde{Q}_n(\boldsymbol{\theta})$ is the smoothed rank correlation function given in Subsection 2.2.1, and p_{λ_n} is a non-concave penalty function. Define the regularized SMRCE (RSMRCE) as

$$\tilde{\boldsymbol{\theta}}_n^{Re} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{Q}_n^{Re}(\boldsymbol{\theta}). \tag{3.2}$$

For the LASSO penalty $p_\lambda(\theta) = \lambda|\theta|$, we know that $\dot{p}_\lambda(\theta) = \lambda$ for $|\theta| > 0$ and the thresholding rule associated with the LASSO penalty is

$$\hat{\theta} = \operatorname{sgn}(z)(|z| - \lambda)_+, \tag{3.3}$$

where z is the estimate from optimizing the original objective function without regularization. For the SCAD penalty

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta|, & \text{when } \lambda \geq |\theta|, \\ \lambda^2 + \frac{(\alpha-1)\lambda^2}{2} - \frac{(\alpha\lambda-|\theta|)^2}{2(\alpha-1)}, & \text{when } \lambda < |\theta| \leq \alpha\lambda, \\ \lambda^2 + \frac{(\alpha-1)\lambda^2}{2}, & \text{when } \alpha\lambda < |\theta|, \end{cases} \tag{3.4}$$

we know that

$$p'_\lambda(\theta) = \begin{cases} \lambda, & \text{when } \lambda \geq |\theta|, \\ \frac{\alpha\lambda-|\theta|}{\alpha-1}, & \text{when } \lambda < |\theta| \leq \alpha\lambda, \\ 0, & \text{when } \alpha\lambda < |\theta| \end{cases} \tag{3.5}$$

and the thresholding rule associated with the SCAD penalty is

$$\hat{\theta} = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda, \\ [(a-1)z - \operatorname{sgn}(z)a\lambda]/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda, \\ z, & \text{when } |z| > a\lambda, \end{cases} \tag{3.6}$$

where z is the estimate without regularization.

The thresholding parameter λ_n is selected by maximizing the following rank correlation information criteria (RCIC)

$$\mathrm{RCIC}(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) - \frac{1}{2n}\sum_{j=1}^{d} I(|\theta_j| > 0), \tag{3.7}$$

where Q_n is the rank correlation function as defined in (2.1). Here $\sum_{j=1}^{d} I(|\theta_j| > 0)$ is the number of non-zero coefficients or the l_0-norm of $\boldsymbol{\theta}$. Therefore the RCIC can be viewed as an AIC-type information criteria based on the rank correlation.

Algorithm 3.1. *(A coordinate-descent path algorithm for RSMRCE plus the SCAD penalty)*

1. *Fix the smoothing matrix in the self-induced smoothing method for Q_n as $I_{d\times d}$.*
2. *Let the initial guess $\tilde{\boldsymbol{\theta}}_n^{Re,(0)}$ be* 0*.*
3. *For stage k, choose $\lambda_n^{(k)}$ as the l_∞-norm of the derivative of $\tilde{Q}_n$ evaluated at estimate $\tilde{\boldsymbol{\theta}}_n^{Re,(k-1)}$.*
4. *Update $\tilde{\boldsymbol{\theta}}_n^{Re,(k)}$ coordinate-wisely by maximizing $\tilde{Q}_n(\theta_j)$, the smoothed rank correlation function of θ_j where other components θ_l's are fixed as the previous estimates, and thresholded that maximizer by the rule in* (3.6)*.*
5. *Within k^{th} stage, repeat step 4 until $\tilde{\boldsymbol{\theta}}_n^{Re,(k)}$ converge.*
6. *Go to next stage* $(k+1)$ *and repeat steps 3, 4 and 5 until k reaches the maximum number of pathes.*
7. *Choose the optimal k, or equivalently $\lambda_n^{(k)}$, by maximizing the RCIC.*

The finite sample performance of the above path algorithm has been assessed by extensive simulation studies in Section 3.4.

3.2.1 Large-sample properties

This subsection is devoted to the large sample theories. The main results are: 1. the regularized and smoothed MRC estimator (RSMRCE) is $\sqrt{n}-$consistent by choosing an appropriate non-concave penalty; 2. the regularized SMRCE is sparse and 3. the non-zero part of RSMRCE achieves asymptotic normality.

Let $\boldsymbol{\theta}_0 = (\theta_{10}, ..., \theta_{d0})' = (\boldsymbol{\theta}'_{10}, \boldsymbol{\theta}'_{20})'$, where $\boldsymbol{\theta}_{10} = (\theta_{10}, ..., \theta_{d_0 0})'$. Without loss of generality, we assume that $\boldsymbol{\theta}_{20} = 0$. Therefore, only d_0 coefficients are non-zero. Define

$$c_n = \max\left\{\dot{p}_{\lambda_n}(|\theta_{j0}|) : \theta_{j0} \neq 0\right\}. \tag{3.8}$$

Under the Assumptions 2.1-2.3, we have the following asymptotic properties for the RSMRCE with a generalized nonconcave penalty p_{λ_n}. The detailed proofs are given in the next section.

Theorem 3.1. *If* $\max_j\{|\ddot{p}_{\lambda_n}(|\theta_{j0}|)| : \theta_{j0} \neq 0\} \to 0$, *then the regularized SMRCE*

$$\hat{\boldsymbol{\theta}}_n^{Re} = \arg\max_{\boldsymbol{\Theta}} \left[\tilde{Q}_n(\boldsymbol{\theta}) - \sum_{j=1}^{d} p_{\lambda_n}(|\theta_j|)\right] \tag{3.9}$$

satisfies

$$\hat{\boldsymbol{\theta}}_n^{Re} = \boldsymbol{\theta}_0 + O_p(n^{-1/2} + c_n), \tag{3.10}$$

where c_n is given by (3.8) and $\ddot{p}_{\lambda_n}$ is the second derivative of the nonconcave penalty p_{λ_n}.

For the SCAD penalty functions, if $\lambda_n < \max_j |\theta_{j0}|/a$, $c_n = 0$ and $\max_j\{|\ddot{p}_{\lambda_n}(|\theta_{j0}|)| : \theta_{j0} \neq 0\} = 0$. Hence, the resulting regularized SMRCE is $\sqrt{n}$-consistent. For the LASSO penalty, $\dot{p}_{\lambda_n}(|\theta|)$ is λ_n except for $\theta = 0$. Therefore in order to obtain $\sqrt{n}$-consistency for the estimator regularized by LASSO penalty, we need the order of λ_n to be $O(n^{1/2})$, which does not meet the requirement for sparsity in the following theorem.

Under the same Assumptions 2.1-2.3, we have the following lemma showing the sparsity of $\hat{\boldsymbol{\theta}}_n^{Re}$.

Lemma 3.1. *Suppose that the nonconcave penalty p_{λ_n} satisfies*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} \{\dot{p}_{\lambda_n}/\lambda_n\} > 0. \tag{3.11}$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\theta}_1$ such that $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_p(n^{-1/2})$, for any constant C,

$$\tilde{Q}_n^{Re}((\boldsymbol{\theta}_1, 0)') = \arg\max_{\|\boldsymbol{\theta}_2\| \le Cn^{-1/2}} \left[\tilde{Q}_n^{Re}((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)')\right]. \tag{3.12}$$

Define the following diagonal matrix

$$\boldsymbol{\Xi} = \operatorname{diag}\left\{\ddot{p}_{\lambda_n}(|\theta_{10}|), ..., \ddot{p}_{\lambda_n}(|\theta_{d_0 0}|)\right\}. \tag{3.13}$$

Denote

$$\mathbf{b} = \begin{pmatrix} \dot{p}_{\lambda_n}(|\theta_{10}|)\,\mathrm{sgn}(\theta_{10}) \\ \dot{p}_{\lambda_n}(|\theta_{20}|)\,\mathrm{sgn}(\theta_{20}) \\ \vdots \\ \vdots \\ \dot{p}_{\lambda_n}(|\theta_{d_0 0}|)\,\mathrm{sgn}(\theta_{d_0 0}) \end{pmatrix} \tag{3.14}$$

Under the Assumptions 2.1-2.3, we have the following oracle properties.

Theorem 3.2. *Suppose that the nonconcave penalty p_{λ_n} satisfies condition* (3.11)*. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}_n^{Re} = (\hat{\boldsymbol{\theta}}_{n,1}^{Re}, \hat{\boldsymbol{\theta}}_{n,2}^{Re})'$ satisfies:*

(a) Sparsity: $\hat{\boldsymbol{\theta}}_{n,2}^{Re} = 0$.

(b) Asymptotic normality:

$$\sqrt{n}(\mathbf{A}_1(\boldsymbol{\theta}_{10}) - \boldsymbol{\Xi})\left\{\hat{\boldsymbol{\theta}}_{n,1}^{Re} - \boldsymbol{\theta}_{10} + (\mathbf{A}_1(\boldsymbol{\theta}_{10}) - \boldsymbol{\Xi})^{-1}\mathbf{b}\right\} \xrightarrow{L} N(0, \mathbf{V}_1(\boldsymbol{\theta}_{10})), \tag{3.15}$$

where $\mathbf{A}_1(\boldsymbol{\theta})$ is the first d_1-by-d_1 block matrix of $\mathbf{A}(\boldsymbol{\theta})$ which is defined in Proposition 2.1, and $-\mathbf{A}_1(\boldsymbol{\theta})$ is positive definite; $\mathbf{V}_1(\boldsymbol{\theta})$ is the first d_1-by-d_1 block matrix of $\mathbf{V}(\boldsymbol{\theta})$ which is defined in proposition 2.1.

As a consequence, the asymptotic covariance matrix of $\sqrt{n}\hat{\boldsymbol{\theta}}_{n,1}^{Re}$ is

$$[\mathbf{A}_1(\boldsymbol{\theta}_{10}) - \Xi]^{-1}\,\mathbf{V}_1(\boldsymbol{\theta}_{10})\,[\mathbf{A}_1(\boldsymbol{\theta}_{10}) - \Xi]^{-1}\,. \tag{3.16}$$

By Theorem 2.2,

$$\left[\hat{\mathbf{A}}_1(\boldsymbol{\theta}_{10}) - \Xi\right]^{-1}\hat{\mathbf{V}}_1(\boldsymbol{\theta}_{10})\left[\hat{\mathbf{A}}_1(\boldsymbol{\theta}_{10}) - \Xi\right]^{-1} \tag{3.17}$$

converges in probability to the asymptotic variance-covariance matrix (3.16). Here $\hat{\mathbf{A}}_1(\boldsymbol{\theta})$ is the first d_1-by-d_1 block matrix of $\hat{\mathbf{A}}(\boldsymbol{\theta})$ which is defined in (2.6); and $\hat{\mathbf{V}}_1(\boldsymbol{\theta})$ is the first d_1-by-d_1 block matrix of $\hat{\mathbf{V}}(\boldsymbol{\theta})$ which is defined in (2.5).

3.3 Proofs of the Theorems

In this section, proofs are provided for (1) the consistency of the regularized SMRCE, (2) the sparsity of the RSMRCE under certain conditions of the penalty function, and (3) the asymptotic normality of the RSMRCE for the non-zero coefficients under the same conditions.

Proof of Theorem 3.1. Define $\alpha_n = n^{-1/2} + c_n$. It suffices to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} \tilde{Q}_n^{Re}(\boldsymbol{\theta}_0 + \alpha_n\mathbf{u}) < \tilde{Q}_n^{Re}(\boldsymbol{\theta}_0)\right\} \geq 1 - \epsilon. \tag{3.18}$$

Since $p_{\lambda_n}(0)=0$, we have

$$\begin{aligned}\tilde{Q}_n^{Re}(\boldsymbol{\theta}_0+\alpha_n\mathbf{u})-\tilde{Q}_n^{Re}(\boldsymbol{\theta}_0)\leq&\tilde{Q}_n(\boldsymbol{\theta}_0+\alpha_n\mathbf{u})-\tilde{Q}_n(\boldsymbol{\theta}_0)\\&-\sum_{j=1}^{d_0}\left\{p_{\lambda_n}(|\theta_{j0}+\alpha_n u_j|)-p_{\lambda_n}(|\theta_{j0}|)\right\},\end{aligned}$$

where $\mathbf{u}=(u_1,...,u_d)'$ and $\boldsymbol{\theta}_0=(\theta_{10},...,\theta_{d0})$. Then by the quadratic expansion of $\tilde{Q}_n$ in Theorem 2.1, we have

$$\begin{aligned}\tilde{Q}_n^{Re}(\boldsymbol{\theta}_0+\alpha_n\mathbf{u})-\tilde{Q}_n^{Re}(\boldsymbol{\theta}_0)\leq\alpha_n\mathbf{u}'\left(\frac{1}{\sqrt{n}}W_n\right)&-\frac{1}{2}\mathbf{u}'\mathbf{A}_0\mathbf{u}\alpha_n^2+o_p(\alpha_n^2+1/n)\\&-\sum_{j=1}^{d_0}\left\{\alpha_n\dot{p}_{\lambda_n}(|\theta_{j0}|)\operatorname{sgn}(\theta_{j0})u_j\right.\qquad(3.19)\\&\left.+\alpha_n^2\ddot{p}_{\lambda_n}(|\theta_{j0}|)u_j^2(1+o(1))\right\}.\end{aligned}$$

It is obvious that the forth term is bounded by

$$d_0\alpha_n c_n C+\alpha_n^2\max\left\{|\ddot{p}_{\lambda_n}(|\theta_{j0}|)|:\theta_{j0}\neq 0\right\}C^2. \tag{3.20}$$

Since $\max\left\{|\ddot{p}_{\lambda_n}(|\theta_{j0}|)|:\theta_{j0}\neq 0\right\}\to 0$ and $W_n=O_p(1)$, the right hand side of (3.19) is dominated by $-\frac{1}{2}\mathbf{u}'\mathbf{A}_0\mathbf{u}\alpha_n^2$ for a sufficiently large C. This completes the proof. □

Proof of Lemma 3.1. It suffices to show that with probability tending to 1 as $n\to\infty$, for any $\boldsymbol{\theta}_1$ satisfying $\boldsymbol{\theta}_1-\boldsymbol{\theta}_{10}=O_p(n^{-1/2})$ and for small $\epsilon_n=Cn^{-1/2}$ and index $j=d_0+1,...,d$,

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\tilde{Q}_n^{Re}(\boldsymbol{\theta})&<0\quad\text{for }0<\theta_j<\epsilon_n,\\&>0\quad\text{for }-\epsilon_n<\theta_j<0.\end{aligned}\tag{3.21}$$

By definition,

$$\frac{\partial}{\partial\theta_j}\tilde{Q}_n^{Re}(\boldsymbol{\theta})=\frac{\partial}{\partial\theta_j}\tilde{Q}_n(\boldsymbol{\theta})-\dot{p}_{\lambda_n}(|\theta_j|)\operatorname{sgn}(\theta_j). \tag{3.22}$$

From the proof of Theorem 2.2, we know that $\partial \tilde{Q}_n(\boldsymbol{\theta})/\partial\theta_j = O_p(n^{-1/2})$. Therefore

$$\frac{\partial}{\partial\theta_j}\tilde{Q}_n^{Re}(\boldsymbol{\theta}) = \lambda_n\left[-\lambda_n^{-1}\dot{p}_{\lambda_n}(|\theta_j|)\,\mathrm{sgn}(\theta_j) + O_p(n^{-1/2}/\lambda_n)\right]. \tag{3.23}$$

Since

$$\liminf_{n\to\infty}\liminf_{\theta\to 0+}\{\dot{p}_{\lambda_n}/\lambda_n\} > 0 \quad \text{and} \quad \sqrt{n}\lambda_n \to \infty,$$

the sign of $\partial\tilde{Q}_n^{Re}(\boldsymbol{\theta})/\partial\theta_j$ is the opposite of that of θ_j for any $|\theta_j| < \epsilon_n$. This completes the proof. □

Proof of Theorem 3.2. The sparsity (a) follows from Lemma 3.1. Then we want to prove the asymptotic normality (b) for $\hat{\boldsymbol{\theta}}_{n,1}^{Re}$.

By the definition of $\hat{\boldsymbol{\theta}}_n^{Re}$ and due to its sparsity (a), for $j = 1, ..., d_0$,

$$\begin{aligned} 0 =& \left.\frac{\partial\tilde{Q}_n^{Re}(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_{n,1}^{Re},0)'} \\ =& \left.\frac{\partial\tilde{Q}_n(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=(\hat{\boldsymbol{\theta}}_{n,1}^{Re},0)'} - \dot{p}_{\lambda_n}\left(|\hat{\theta}_j^{Re}|\right)\mathrm{sgn}\left(\hat{\theta}_j^{Re}\right). \end{aligned} \tag{3.24}$$

By Taylor's expansion,

$$\begin{aligned} 0 =& \frac{\partial\tilde{Q}_n(\boldsymbol{\theta}_0)}{\partial\theta_j} + \sum_{k=1}^{d_0}\left[\left\{\frac{\partial^2}{\partial\theta_j\partial\theta_k}\tilde{Q}_n(\boldsymbol{\theta}_0) + o_p(1)\right\}\left(\hat{\theta}_k^{Re} - \theta_{k0}\right)\right] \\ & - \left(\dot{p}_{\lambda_n}(|\theta_{j0}|)\,\mathrm{sgn}(\theta_{j0}) + \{\ddot{p}_{\lambda_n}(|\theta_{j0}|) + o_p(1)\}\left(\hat{\theta}_j^{Re} - \theta_{j0}\right)\right). \end{aligned} \tag{3.25}$$

Then similar to the proof of Theorem 2.2, it can be shown that

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\tilde{Q}_n(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{A}_{jk}(\boldsymbol{\theta}_0), \tag{3.26}$$

and

$$\sqrt{n}\frac{\partial}{\partial\theta_j}\tilde{Q}_n(\boldsymbol{\theta}_0) \xrightarrow{L} N(0, \mathbf{V}_{jj}(\boldsymbol{\theta}_0)). \tag{3.27}$$

Therefore, by Slutsky's theorem, we get the asymptotic normality of $\hat{\theta}_1^{Re}$. □

3.4 Numerical Results

In this section, we apply the regularized method to the smoothed maximum rank correlation estimator. The results from several large-p and small-n examples are reported and compared for the finite sample performance. Since the regularized and smoothed maximum rank correlation estimator (RSMRCE) does not rely on the error distribution and the parametric form of the transformation function, the results are expected to be robust for the RSMRCE. Without loss of generality, we consider the linear model with different error distributions and different dependent levels for the design matrix X. We choose SCAD penalty with $a = 3.7$ (Fan and Li, 2001) to regularize the smoothed MRC estimator. The thresholding parameter λ_n is selected by a rank correlation information criteria (RCIC). The RSMRCE with SCAD penalty is compared with LASSO-BIC method where the thresholding parameter for the l_1 penalty is selected by Bayesian information criteria (BIC). The performance of these two methods are assessed in terms of median absolute deviation (MAD), the average true positive (TP) rate which is the average number of correctly selected non-zeros, the average true negative (TN) rate which is the average number of correctly selected zeros, and the average false discovery rate (FDR). The average TP rate is also known as the sensitivity and the average TN rate the specificity.

All simulations and computations are conducted in MATLAB.

Simulation 3.1. (A linear model with gaussian noise). In this example we consider the linear regression model with gaussian noise. We choose the sample size $n = 200$ and the number of predictors $d = 1000$. Therefore, this is the usual large-p and small-n setup for variable selection problems. We simulate 500 datasets from the following linear model

$$Y = \mathbf{X}'\boldsymbol{\beta} + \sigma_\epsilon \epsilon, \tag{3.28}$$

where $\boldsymbol{\beta} = (1, \boldsymbol{\theta}')' = (1, 3, 1, 1, 1, 1, 1.5, 2, 2, 2, 0, ..., 0)'$, ϵ follows standard normal distribution, $\sigma_\epsilon = 5$, and X follows multivariate normal distribution with mean $\mu_X = (20, -10, 10, 10, ..., 10, 10)'$ and covariance matrix $25 \times \boldsymbol{\Sigma}_X$ where $\boldsymbol{\Sigma}_X$'s diagonal elements are all 1's and off-diagonal elements are all equal to ρ. Here ρ is the pair-wise correlation which is chosen as $\rho = 0.0, 0.2, 0.4, 0.6, 0.8$ to introduce different levels of dependence to the design matrix.

The simulation results are summarized in Tables 3.1, 3.2 and 3.3. From Table 3.1, we observe that the thresholding parameter λ_n selected by BIC for LASSO method results in a large false discovery rate, which is almost 10 times bigger than those obtained by the regularized SMRCE plus RCIC. The sensitivity and specificity rates for LASSO-BIC are slightly better than those for RSMRCE-RCIC when $\rho = 0.0, 0.2, 0.4, 0.6$, while LASSO-BIC method sacrifices a lot in controlling the F-DR. As the positive correlation parameter ρ increases, the FDR increases and the sensitivity rate as well as the specificity rate decrease for most of the cases. From Tables 3.2 and 3.3, we know that the bias introduced by the l_1 penalty to LASSO estimate is similar for different non-zero coefficients while the bias introduced by the SCAD penalty varies. This is consistent with Theorem 3.2 and the fact that both the first and second derivatives of the LASSO penalty remain constant for different non-zero coefficients. As the positive correlation parameter ρ increases, the MAD as well as the absolute value of bias increases for most of the cases.

Simulation 3.2. (A linear model with a mixture error distribution)In this example we consider the linear regression model where the error distribution is a mixture. We choose the sample size $n = 200$ and the number of predictors $d = 1000$. Therefore, this is the usual large-p and small-n setup for variable selection problems. We simulate

Table 3.1: The simulations for a linear model with gaussian noise

	Method	FDR	Sensitivity	Specificity
$\rho = 0.0$	RSMRCE	0.0238	0.9858	0.8962
	LASSO	0.4597	1.0000	0.9091
$\rho = 0.2$	RSMRCE	0.0089	1.0000	0.9091
	LASSO	0.6515	1.0000	0.9091
$\rho = 0.4$	RSMRCE	0.0463	0.9996	0.9087
	LASSO	0.7009	1.0000	0.9091
$\rho = 0.6$	RSMRCE	0.0972	0.9771	0.8883
	LASSO	0.7230	1.0000	0.9091
$\rho = 0.8$	RSMRCE	0.2188	0.7427	0.6753
	LASSO	0.7482	1.0000	0.9091

Table 3.2: MAD of estimated coefficients for a linear model with gaussian noise

	True value	Method	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
θ_2	3	RSMRCE	0.2964	0.2951	0.3708	0.4228	0.6806
		LASSO	0.2252	0.1974	0.2223	0.2555	0.3590
θ_3	1	RSMRCE	0.1189	0.1213	0.1393	0.1780	0.2658
		LASSO	0.2265	0.2021	0.2177	0.2513	0.3255
θ_4	1	RSMRCE	0.1212	0.1166	0.1551	0.2029	0.2923
		LASSO	0.2280	0.2057	0.2200	0.2631	0.3375
θ_5	1	RSMRCE	0.1240	0.1147	0.1475	0.1882	0.2701
		LASSO	0.2270	0.2017	0.2269	0.2521	0.3439
θ_6	1	RSMRCE	0.1105	0.1150	0.1424	0.1893	0.2735
		LASSO	0.2243	0.2021	0.2098	0.2483	0.3366
θ_7	1.5	RSMRCE	0.1429	0.1541	0.1874	0.2423	0.3578
		LASSO	0.2215	0.2014	0.2247	0.2503	0.3153
θ_8	2	RSMRCE	0.2050	0.1941	0.2510	0.2882	0.4613
		LASSO	0.2245	0.2005	0.2294	0.2503	0.3560
θ_9	2	RSMRCE	0.1992	0.1942	0.2426	0.2860	0.4690
		LASSO	0.2295	0.1990	0.2124	0.2521	0.3490
θ_{10}	2	RSMRCE	0.1964	0.1989	0.2416	0.2894	0.4528
		LASSO	0.2319	0.1995	0.2134	0.2606	0.3589

Table 3.3: BIAS of estimated coefficients for a linear model with gaussian noise

	True value	Method	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
θ_2	3	RSMRCE	0.0610	-0.2337	-0.3177	-0.3756	-0.6523
		LASSO	-0.2273	-0.1959	-0.2217	-0.2550	-0.3493
θ_3	1	RSMRCE	0.0593	-0.0806	-0.1120	-0.1621	-0.2606
		LASSO	-0.2238	-0.2035	-0.2217	-0.2550	-0.3386
θ_4	1	RSMRCE	0.0598	-0.0800	-0.1176	-0.1881	-0.2524
		LASSO	-0.2315	-0.2084	-0.2223	-0.2595	-0.3317
θ_5	1	RSMRCE	0.0573	-0.0841	-0.1143	-0.1653	-0.2845
		LASSO	-0.2277	-0.2035	-0.2257	-0.2510	-0.3441
θ_6	1	RSMRCE	0.0575	-0.0800	-0.1018	-0.1831	-0.3035
		LASSO	-0.2250	-0.2033	-0.2191	-0.2569	-0.3369
θ_7	1.5	RSMRCE	0.0941	-0.1174	-0.1581	-0.1949	-0.3629
		LASSO	-0.2257	-0.2009	-0.2249	-0.2519	-0.3218
θ_8	2	RSMRCE	0.1306	-0.1514	-0.2177	-0.2327	-0.4250
		LASSO	-0.2280	-0.1993	-0.2277	-0.2501	-0.3565
θ_9	2	RSMRCE	0.1073	-0.1535	-0.2046	-0.2501	-0.4182
		LASSO	-0.2298	-0.2000	-0.2167	-0.2522	-0.3407
θ_{10}	2	RSMRCE	0.1104	-0.1538	-0.2076	-0.2445	-0.4192
		LASSO	-0.2313	-0.1980	-0.2146	-0.2551	-0.3512

500 datasets from the following linear model

$$Y = \mathbf{X}'\boldsymbol{\beta} + \sigma_\epsilon \epsilon, \tag{3.29}$$

where $\boldsymbol{\beta} = (1, \boldsymbol{\theta}')' = (1, 3, 1, 1, 1, 1, 1.5, 2, 2, 2, 0, ..., 0)'$, the distribution of ϵ is a mixture: $0.8N(0, 1) + 0.2t_1$ where t_1 is the Student's t distribution with degree of freedom equal 1, $\sigma_\epsilon = 3$, and X follows multivariate normal distribution with mean $\mu_X = (20, -10, 10, 10, ..., 10, 10)'$ and covariance matrix $25 \times \boldsymbol{\Sigma}_X$ where $\boldsymbol{\Sigma}_X$'s diagonal elements are all 1's and off-diagonal elements are all equal to ρ. Here ρ is the pair-wise correlation which is chosen as $\rho = 0.0, 0.2, 0.4, 0.6, 0.8$ to introduce different levels of dependence to the design matrix. Note that the Student's t distribution with 1 degree of freedom is exactly the Cauchy distribution. Therefore, the error term in this simulation design introduces about 20% outliers, which is used to test the robustness of the regularized SMRCE as well as that of the LASSO method.

The simulation results are summarized in Tables 3.4, 3.5 and 3.6. From Table 3.4, we observe that the thresholding parameter λ_n selected by BIC for the LASSO method results in a large false discovery rate, which is almost 10 times bigger than those obtained by the regularized SMRCE plus RCIC. The sensitivity and specificity rates for the LASSO-BIC method are now worse than those for the RSMRCE+RCIC method when $\rho = 0.0, 0.2, 0.4, 0.6, 0.8$, while the LASSO-BIC method sacrifices a lot in controlling the FDR. This is due to the mis-specification of the error distribution by LASSO method. For $\rho = 0.0, 0.2, 0.4, 0.6$, the sensitivity and specificity rates for the RSMRCE-RCIC method are similar to the corresponding values from Simulation 3.1. This shows that the RSMRCE-RCIC method is robust when the dependency in $\mathbf{X}$ is not so strong. As the positive correlation parameter ρ increases, the FDR increases and the sensitivity rate as well as the specificity rate decrease for most of the cases. From Tables 3.5 and 3.6, we know that the bias introduced by the l_1 penalty in

LASSO is similar for different non-zero coefficients while the bias introduced by the SCAD penalty varies. This is also consistent with Theorem 3.2 and the fact that both and first and second derivatives of the LASSO penalty remain constant for different non-zero coefficients. As the positive correlation parameter ρ increases, the MADs as well as the absolute values of bias increase for most of the cases. The MADs as well as the the absolute values of bias for the RSMRCE-RCIC method are smaller than the corresponding values from Simulation 3.1. This is due to a smaller σ_ϵ. Even though σ_ϵ is smaller for Simulation 3.2, the MADs as well as the the absolute values of bias for the RSMRCE-RCIC method are bigger than the corresponding values from Simulation 3.1. This is because the LASSO-BIC method relies on the normality of the error distribution, which becomes a severe problem when there are many outliers in data.

Remark 3.1. In theory, the SCAD penalty does not introduce bias to the regularized SMRCE as well as other regularized estimators by choosing a proper thresholding parameter λ_n. However, we found that the derivative of the SCAD penalty function is zero only for $|\theta_{i0}| > a\lambda$ where $a > 2$. Therefore, for those θ_{i0}'s with small absolute values, the estimator penalized by the SCAD penalty may have estimation bias due to a unified thresholding parameter. This is the reason why the bias in Tables 3.3 and 3.6 are big when using the SCAD penalty. A feasible remedy is to make the thresholding parameter λ_n adaptive in the sense that we make $\lambda_n(\theta_i)$ proportional to $|\hat{\theta}_i|$.

3.5 Discussion

This chapter develops a variable selection method for the general monotone transformation model. The variable selection method consists of the regularized SMRCE

Table 3.4: The simulations for a linear model with mixed noise

	Method	FDR	Sensitivity	Specificity
$\rho = 0.0$	RSMRCE	0.0019	0.9858	0.8962
	LASSO	0.2569	0.9376	0.8524
$\rho = 0.2$	RSMRCE	0.0061	1.0000	0.9091
	LASSO	0.5033	0.9367	0.8516
$\rho = 0.4$	RSMRCE	0.0434	1.0000	0.9091
	LASSO	0.6182	0.9433	0.8576
$\rho = 0.6$	RSMRCE	0.0943	0.9996	0.9087
	LASSO	0.6969	0.8947	0.8134
$\rho = 0.8$	RSMRCE	0.1649	0.9787	0.8897
	LASSO	0.7363	0.8578	0.7799

Table 3.5: MAD of estimated coefficients for a linear model with mixed noise

	True value	Method	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
θ_2	3	RSMRCE	0.2063	0.1896	0.2404	0.3192	0.3843
		LASSO	0.3376	0.2887	0.2889	0.3355	0.4208
θ_3	1	RSMRCE	0.0765	0.0742	0.0825	0.1162	0.1573
		LASSO	0.3307	0.2940	0.2880	0.3430	0.4061
θ_4	1	RSMRCE	0.0750	0.0712	0.0829	0.1162	0.1525
		LASSO	0.3170	0.2989	0.3035	0.3430	0.4066
θ_5	1	RSMRCE	0.0791	0.0692	0.0862	0.1225	0.1573
		LASSO	0.3154	0.3063	0.2904	0.3389	0.3979
θ_6	1	RSMRCE	0.0772	0.0710	0.0854	0.1108	0.1528
		LASSO	0.3301	0.2991	0.2819	0.3521	0.4332
θ_7	1.5	RSMRCE	0.1023	0.1063	0.1229	0.1614	0.2097
		LASSO	0.3180	0.3027	0.2865	0.3403	0.4041
θ_8	2	RSMRCE	0.1356	0.1336	0.1598	0.2107	0.2655
		LASSO	0.3366	0.2874	0.3013	0.3438	0.4063
θ_9	2	RSMRCE	0.1297	0.1372	0.1516	0.2123	0.2616
		LASSO	0.3339	0.3081	0.2793	0.3314	0.3906
θ_{10}	2	RSMRCE	0.1353	0.1292	0.1516	0.2132	0.2472
		LASSO	0.3425	0.2787	0.2954	0.3423	0.3850

Table 3.6: BIAS of estimated coefficients for a linear model with mixed noise

	True value	Method	$\rho = 0.0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
θ_2	3	RSMRCE	-0.0819	-0.1492	-0.2019	-0.2897	-0.3313
		LASSO	-0.4539	-0.4512	-0.4006	-0.5112	-0.7039
θ_3	1	RSMRCE	0.0490	-0.0494	-0.0682	-0.1021	-0.1466
		LASSO	-0.4091	-0.3661	-0.3671	-0.4022	-0.4866
θ_4	1	RSMRCE	0.0519	-0.0495	-0.0671	-0.1020	-0.1619
		LASSO	-0.4552	-0.3897	-0.3681	-0.4045	-0.5008
θ_5	1	RSMRCE	0.0516	-0.0508	-0.0736	-0.1052	-0.1697
		LASSO	-0.4209	-0.3423	-0.3673	-0.4334	-0.4737
θ_6	1	RSMRCE	0.0485	-0.0531	-0.0673	-0.0983	-0.1611
		LASSO	-0.3858	-0.3914	-0.3269	-0.4221	-0.5013
θ_7	1.5	RSMRCE	0.0786	-0.0766	-0.1636	-0.1404	-0.1652
		LASSO	-0.4176	-0.3833	-0.3947	-0.5001	-0.5645
θ_8	2	RSMRCE	0.0951	-0.1007	-0.1389	-0.1919	-0.2288
		LASSO	-0.5056	-0.3844	-0.4137	-0.5645	-0.5916
θ_9	2	RSMRCE	0.1037	-0.1014	-0.1332	-0.1932	-0.2201
		LASSO	-0.4619	-0.4104	-0.4095	-0.5523	-0.6119
θ_{10}	2	RSMRCE	0.1065	-0.0976	-0.1289	-0.1913	-0.2139
		LASSO	-0.4777	-0.4165	-0.4215	-0.5087	-0.6203

and the rank correlation information criteria. For the regularized SMRCE, we add suitable penalty functions such as SCAD to the smoothed rank correlation function which is defined in Chapter 2. The rank correlation information criteria is a modified rank correlation function which is adjusted for the dimensional complexity for the selected predictors. It is shown that such regularized SMRCE achieves desired sparsity. If the thresholding parameter λ_n are selected properly, the RSMRCE does not introduce any significant bias in the sense that the regularized estimator is consistent and asymptotically normal. Since the method based on the smoothed rank correlation function is distribution-free, the RSMRCE+RCIC method is more robust than other parametric variable selection algorithms. Simulation studies provide additional evidence that the proposed method are better than those existing methods such as LASSO-BIC.

Chapter 4

Estimation of the Monotone Transformation Function

4.1 Introduction

In this chapter we consider estimating the monotone transformation function H of model 1.1. Without loss of generality, we assume that H is strictly increasing. Instead of working on the original model 1.1, we focus on the equivalent model

$$\Lambda_0(Y) = \mathbf{X}'\boldsymbol{\beta} + \epsilon, \tag{4.1}$$

where $\Lambda_0(\cdot) = H^{-1}(\cdot)$ is the inverse of H and thus is strictly increasing as well. Chen (2002) proposed a rank-based estimator for Λ_0 assuming that there is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\beta}(\boldsymbol{\theta})$ (as in Sherman (1998)). Simulation studies by Chen (2002) showed that the rank-based method has good finite sample performance. Furthermore, Chen (2002) proved that under mild regularity conditions the rank-based estimator is uniformly consistent over a closed interval of Y and it also converges weakly to a Gaussian process with mean 0 and a bounded covariance function.

Although Chen's rank-based estimator for the transformation function Λ_0 achieves desired large sample properties, it is very difficult to estimate the covariance function for the limiting Gaussian process because the rank criteria function is discrete. Chen (2002) proposed a finite difference approach to approximate the covariance function but that approach involves with bandwidth selection. Therefore the discreteness of the rank correlation function is a major drawback for the rank-based estimator and makes the statistical inference on the estimate very difficult.

To address the issue of discreteness, the self-induced smoothing technique developed in Chapter 2 is applied to Chen's rank-based estimate. The resulting smoothed rank-based estimate is uniformly consistent over the same interval of Y and it converges weakly to a Gaussian process as well, which is the same as the limiting Gaussian process for Chen's rank-based estimator. From the smoothed rank correlation function of Λ_0, a close-form formula can be derived easily for the covariance function of the limiting Gaussian process. In addition, the covariance formula is consistent.

The rest of this chapter is organized as follows. In Section 4.2, the new methods are described and related large sample properties are developed. Section 4.3 contains discussions and some concluding remarks.

4.2 Main Results

In this section, the self-induced smoothing is applied to the rank correlation function of Λ_0. The section is further divided into two subsections, with the first introducing the estimation method and the covariance formula, the second establishing the large sample properties.

4.2.1 Methods

Note that model 4.1 continues to hold if Λ_0 and ϵ are replaced by $\Lambda_0 + \alpha$ and $\epsilon + \alpha$ for any constant α (a location shift), or Λ_0, ϵ and $\boldsymbol{\beta}$ by $\Lambda_0\alpha$, $\epsilon\alpha$ and $\alpha\boldsymbol{\beta}$ for any positive constant α (a scaling coefficient). To address this identifiability issue, we assume that $\Lambda_0(y_0) = 0$ for some finite y_0 and we reparameterize $\boldsymbol{\beta}$ as $(\boldsymbol{\theta}, 1)$. Define $d_{iy} = I[Y_i \leq y] = I[\mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i \leq \Lambda_0(y)]$ and $d_{iy_0} = I[Y_i \leq y_0] = I[\mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i \leq \Lambda_0(y_0)]$. The rank correlation function for Λ is defined as

$$Q_n^{\Lambda}(y, \Lambda, b) = \frac{1}{n(n-1)} \sum_{i \neq j} (d_{iy} - d_{jy_0}) I[\mathbf{X}_i'b - \mathbf{X}_j'b \geq \Lambda]. \tag{4.2}$$

Chen's rank-based estimate is defined as

$$\hat{\Lambda}_n(y) = \arg\max_{\Lambda \in M_\Lambda} Q_n^{\Lambda}(y, \Lambda, b_n), \tag{4.3}$$

for any given $y \in [y_2, y_1]$, where M_Λ is an appropriate compact set and b_n is the $\sqrt{n}$ consistent estimator for $\boldsymbol{\beta}$.

The objective function $Q_n^{\Lambda}(y, \Lambda, b)$ is a step function, to which we can apply the self-induced smoothing technique. Let Z be a random variable with mean 0 with standard normal distribution. Assume that Z is independent with data and let E_Z be the expectation with respect to Z given data. A self-induced smoothing for Q_n^{Λ} is $\tilde{Q}_n^{\Lambda}(y, \Lambda, b) = E_Z Q_n^{\Lambda}(y, \Lambda + Z/\sqrt{n}, b)$.

Following the same notations in Chapter 2, we calculate the smoothed rank correlation function for Λ as

$$\tilde{Q}_n^{\Lambda}(y, \Lambda, b) = \frac{1}{n(n-1)} \sum_{i \neq j} (d_{iy} - d_{jy_0}) \Phi\left(\sqrt{n}(\mathbf{X}_{ij}'b - \Lambda)\right). \tag{4.4}$$

We then use $\tilde{\Lambda}_n(y) = \arg\max_{\Lambda \in M_\Lambda} \tilde{Q}_n^{\Lambda}(y, \Lambda, b_n)$ as the smoothed rank estimator for $\Lambda_0(y)$.

Define

$$\hat{V}_n^{\Lambda}(y, y', \Lambda, b) = \frac{1}{n^3} \sum_{i=1}^{n} \left\{ \sum_{j} \left\{ n(d_{iy} - d_{jy_0})(d_{iy'} - d_{jy_0}) \right.\right. \tag{4.5}$$
$$\left.\left. \phi\left(\sqrt{n}(\mathbf{X}_{ij}' b - \Lambda(y))\right) \phi\left(\sqrt{n}(\mathbf{X}_{ij}' b - \Lambda(y'))\right) \right\} \right\}$$

and

$$\hat{A}_n^{\Lambda}(y, \Lambda, b) = \frac{1}{2n(n-1)} \sum_{i \neq j} \left\{ n(d_{iy} - d_{jy_0}) \dot{\phi}\left(\sqrt{n}(\mathbf{X}_{ij}' b - \Lambda(y))\right) \right\}, \tag{4.6}$$

where the notations follow from those in Chapter 2.

Let

$$\hat{D}_n^{\Lambda}(y, y', \Lambda, b) = \left[\hat{A}_n^{\Lambda}(y, \Lambda, b)\right]^{-1} \hat{V}_n^{\Lambda}(y, y', \Lambda, b) \left[\hat{A}_n^{\Lambda}(y', \Lambda(y'), b)\right]^{-1}. \tag{4.7}$$

Then in the next subsection we will see that $\hat{D}_n^{\Lambda}(y, y', \tilde{\Lambda}_n, b_n)$ converges in probability to the covariance function of the limiting Gaussian process for Chen's rank-based estimate.

4.2.2 Large-sample properties

In this section, we derive the large sample properties of the smoothed rank-based estimator as well as the covariance formula defined in the previous subsection. These properties are based on the following assumptions as well as Assumptions 2.1-2.3 in Chapter 2.

Assumption 4.1. Λ_0 is strictly increasing (or decreasing). $\Lambda_0(y_0) = 0$, $[\Lambda_0(y_2 - \epsilon^*), \Lambda_0(y_1 + \epsilon^*)] \subset M_\Lambda$ for a small positive number ϵ^* for some y_0, y_1, y_2 in the support of Y, where M_Λ is a compact interval.

Define $\tau^{\Lambda}(\omega, y, \Lambda, b) = E\left[h^{\Lambda}(\omega, W, y, \Lambda, b) + h^{\Lambda}(W, \omega, y, \Lambda, b)\right]$ where

$$h^{\Lambda}(\omega_1, \omega_2, y, \Lambda, b) = \left(I[y^1 \geq y] - I[y^2 \geq y_0]\right) I[x^1 b - x^2 b \geq \Lambda]$$

for $\omega_1 = (x^1, y^1)$, $\omega_2 = (x^2, y^2)$, and $W = (X, Y)$.

Assumption 4.2. $V^{\Lambda}(y) = E[\partial^2 \tau(W, y, \Lambda_0(y), \boldsymbol{\beta}_0)/\partial\Lambda^2]/2$ is negative for each $y \in [y_2, y_1]$, and uniformly bounded away from 0.

Assumption 4.3. The estimator b_n is $\sqrt{n}$-consistent, for example, Han's MRC estimator and the SMRCE defined in Chapter 2.

The above assumptions are the same regularity conditions required by Chen's rank-based estimator. Under the Assumptions 2.1-2.3 and 4.1-4.3, the following theorem shows the uniform consistency and the weak convergence of $\tilde{\Lambda}_n(y)$. In addition, the following theorem establishes the asymptotic equivalency between $\hat{\Lambda}_n(y)$, Chen's rank-based estimator, and $\tilde{\Lambda}_n(y)$, the smoothed rank estimator.

Theorem 4.1. *Under Assumptions 2.1-2.3 and 4.1-4.3: (i)* $\sup_{y_2 \le y \le y_1} |\tilde{\Lambda}_n(y) - \Lambda_0(y)| = o_p(1)$*; (ii) uniformly over* $y \in [y_2, y_1]$*,*

$$\sqrt{n}(\tilde{\Lambda}_n(y) - \Lambda_0(y)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J_{y_0,y}(\mathbf{X}_i, Y_i) + o_p(1), \tag{4.8}$$

and

$$\sqrt{n}(\tilde{\Lambda}_n(y) - \Lambda_0(y)) \xrightarrow{w} H_{\Lambda}(y_0, y) \tag{4.9}$$

where $H_{\Lambda}(y_0, y)$ *is a Gaussian process with mean 0 and a covariance function*

$$\Gamma^{\Lambda}(y, y'; y_0) = E[J_{y_0,y}(\mathbf{X}, Y) J_{y_0,y'}(\mathbf{X}, Y)] \tag{4.10}$$

with

$$J_{y_0,y}(\mathbf{X}, Y) = -V^{\Lambda}(y)^{-1} \frac{\partial}{\partial \Lambda} \tau^{\Lambda}(W, y, \Lambda_0(y), \boldsymbol{\beta}_0). \tag{4.11}$$

Moreover, the limiting Gaussian process for $\sqrt{n}(\tilde{\Lambda}_n(y) - \Lambda_0(y))$ *is the same as that for* $\sqrt{n}(\hat{\Lambda}_n(y) - \Lambda_0(y))$.

Theorem 4.2. *The covariance estimate* $\hat{D}_n^{\Lambda}(y, y', \tilde{\Lambda}_n, b_n)$ *based on* (4.7) *converges in probability to the covariance function* $\Gamma^{\Lambda}(y, y'; y_0)$ *in* (4.10) *uniformly over* $\{(y, y') : y \in [y_2, y_1], y' \in [y_2, y_1]\}$.

The proofs for the above two theorems are similar to those for Theorems 2.2.1-2 and therefore omitted.

4.3 Discussion

In this chapter, we apply the self-induced smoothing method developed in Chapter 2 to Chen's rank-based estimator for the strictly monotone function Λ_0. Through the smoothed rank correlation function of Λ_0, we derive a close form covariance formula for the limiting Gaussian process for Chen's rank-based estimate, and thus overcome the difficulty of making statistical inference on the rank-based estimate for Λ_0. There are remaining questions for estimating Λ_0 because both the original and the smoothed rank estimate are constructed point-wisely, and therefore fail to provide a unified functional estimate for Λ_0. The point-wise estimate $\tilde{\Lambda}_n$ or $\hat{\Lambda}_n$ can be used together with those nonparametric curve fitting methods such as monotonic splines methods to achieve a unified functional estimate. Although the fitting results from Figures 4.1-4 look good, the asymptotic properties remain unknown for this combined method. This is one of our future research directions.

Figure 4.1: Estimation of the exponential function

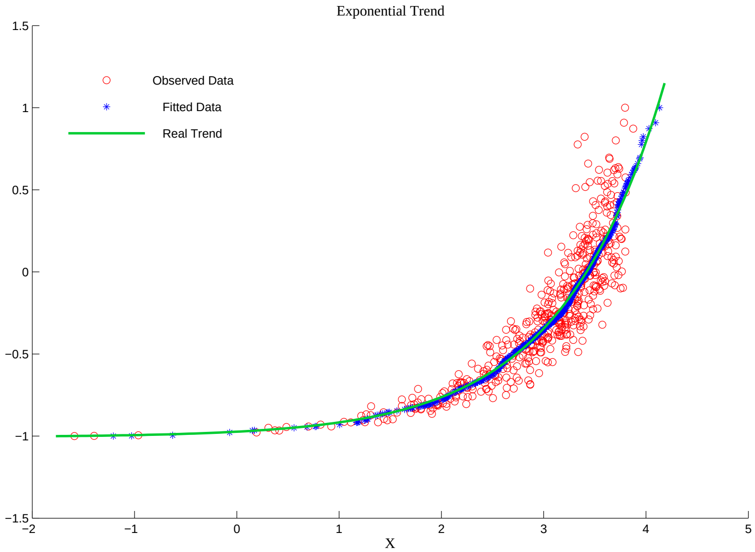


Figure 4.2: Estimation of the logarithm function

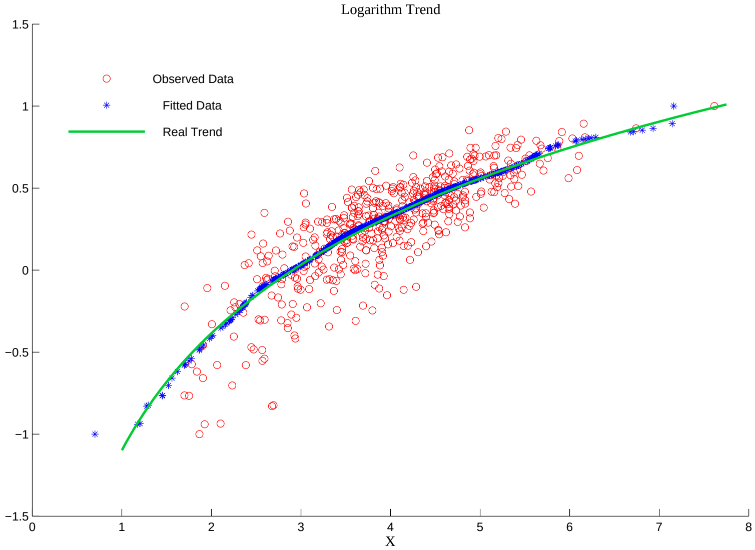


Figure 4.3: Estimation of the logistic function

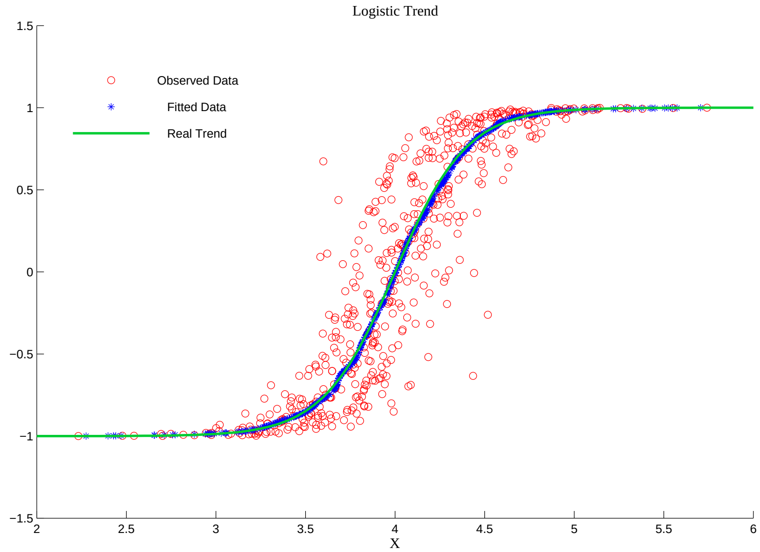
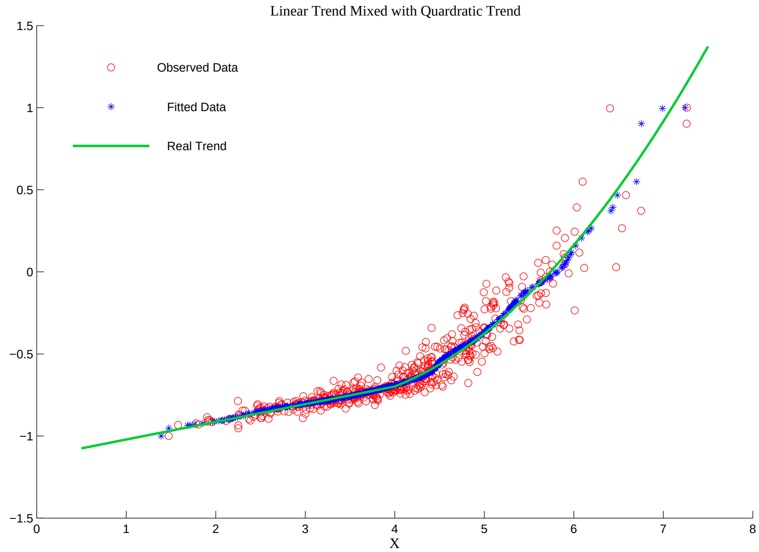


Figure 4.4: Estimation of the mixed function



Chapter 5

Multiple Hypotheses Testing

5.1 Introduction

In this chapter, we develop a new multiple hypotheses testing method for linear models, which can be generalized for the monotone transformation model through the estimation method developed in Chapter 4. This research is motivated by the applied research of REE studies with Dr. Wang from the New York City Obesity Research Center.

5.1.1 Background

The resting energy expenditure (REE) measures the amount of calories required for a 24-hour period by the whole body during a non-active period. It is one of the important quantities in nutrition study because it provides a reference level of daily energy consumption for human so as to prevent any under- or over-feeding. Therefore, the REE is so important that it is carefully monitored during all kinds of weight-loss programs.

There are direct and indirect methods to measure the resting energy expenditure. The direct method involves with putting a patient in a calorimeter and measuring the amount of heat produced by the body. The indirect method involves with analyzing the continuous gas exchange between O_2 and CO_2 of the patient under the resting condition, and then transform the exchange rates to the energy consumption. Both of these two methods require not only certain machine such as a calorimeter, sensor and metabolic cart, but also strict conditions such as a resting state of the patient, no consumption of food and calorie-containing beverages prior to the measurement and certain levels of environmental temperature and humidity. Therefore, a large-scale clinical study is prohibitive and the sample size is limited for certain groups of patients, which is one of the major difficulties to examine the REE.

Besides the importance of measuring the REE, it is also crucial to understand the composition of the resting energy expenditure for both nutrition study and clinical practice. To better understand the composition of REE, a lot of prediction models were proposed, for instance, Harris Benedict equation (Harris and Benedict, 1918 and 1919; Roza and Shizgal, 1984), organ-tissue level model and cellular level model. In this chapter, we examine the organ-tissue level model which decomposes the whole-body resting energy expenditure as a summation of resting energy consumptions in all organs and tissues of human body. According to the literature [see Elia (1992) and Gallagher et al. (1998)], there are six major organs and tissues considered in the mechanism of resting energy consumption, i.e., liver, brain, heart, kidneys, skeletal muscle and adipose tissue. Therefore, the organ-tissue level REE model is

$$REE = \sum_{i=1}^{7}(R_i \times T_i), \tag{5.1}$$

where T_i $(i = 1, ..., 6)$ is the individual organ/tissue mass in kilogram, i is the index for one organ/tissue, and R_i is the resting metabolic rate of corresponding organ/tissue in kilocalorie per kilogram and per day. In addition, T_7 is the residual mass of other organs and tissues, i.e., excluding the masses of 6 major organs and tissues from the whole body mass. Therefore the 7^{th} (last) component of the summation in (5.1) accounts for the total energy consumption of the rest of the human body. The choice of major organs and tissues in (5.1) is reasonable because these 4 organs, i.e., liver, brain, heart and kidneys, have particularly high basal specific metabolic rate and the 2 tissues, i.e., skeletal muscle and adipose tissue, have relatively large masses. The masses of major organs can be measured by magnetic resonance imaging (MRI) and the masses of tissues can be measured by a dual-energy X-ray absorptiometry scanner. Therefore, the main problem of exploring the relationship between REE and body composition becomes how to evaluate the resting organ/tissue metabolic rate for a certain population. Some studies, e.g. by Wang et al. (2005, 2007, 2010), Bosy et al. (2003), Gallagher et al. (2000) and Wakabayashi et al. (2002) find that the changes in the whole body and liver showed an expansion of extracellular compartments and a relative loss of cellularity among older adults, which results in the decrease in both metabolic rates and REE for older adults. Therefore, it is also important to evaluate the R_i values for different population so as to understand the basic energy requirement of a certain group.

Elia (1992) suggested the following set of coefficients for healthy adults as the resting metabolic rates in (5.1): 200 for liver, 240 for brain, 440 for heart and kidneys, 13 for skeletal muscle, 4.5 for adipose tissue and 12 for the residual organs and tissues. However, this set of coefficients has never been closely examined let alone whether Elia's R_i values are applicable to a certain group of patients.

5.1.2 The issue of collinearity

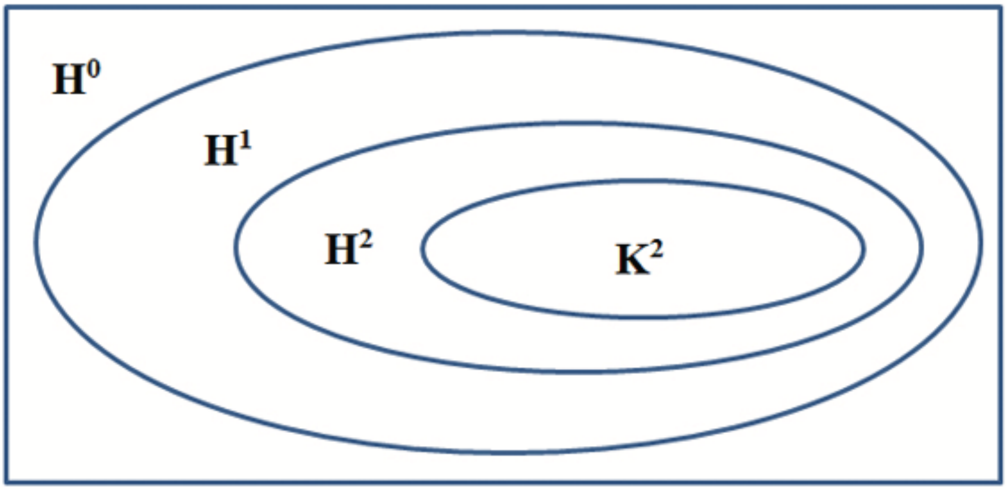
The mechanistic REE model (5.1) appears to be a natural candidate for the use of multiple linear regression. However, the significant dependency among the organ and tissue masses and the limited sample size make the least squares fit to (5.1) unstable. Specifically, the standard errors for the estimates of R_i's are exceptionally large and thus the individual 95% confidence interval of each R_i provides little information of the true value. In addition, the high collinearity among T_i values raises questions of using standard multiple testing procedures such as Bonferroni's correction and step-up/step-down procedures. For instance, the Holm's (1979) step-down procedure and the Benjamini-Hochberg (1995) false discovery rate (FDR) controlling procedure, both of which are conservative because of ignoring the dependent structure of the estimates. To deal with the dependent case, Cohen et al. (2009) proposed the maximum residual down (MRD) method for testing the means of correlated normal random variables. They showed that the MRD method is intuitive and have a desirable convexity property required for admissibility. However, there are still two problems of directly applying the MRD method to our study. At first, it is difficult to find the optimal choice of the critical values. Secondly, it is still unclear whether the MRD method is applicable to linear regression models, especially for large p and small n problems.

Another issue associated with testing the REE model (5.1) is how to formulate the testing hypothesis. One candidate of testing Elia's R_i values is to examine the following multiple hypotheses,

$$\boldsymbol{H}_i : R_i = R_{i0} \quad vs \quad \boldsymbol{K}_i : R_i \neq R_{i0}, \quad i = 1, ..., 7, \tag{5.2}$$

where R_{i0} is the metabolic rates of i^{th} organ/tissue suggested by Elia. However, since the organ and tissue masses are highly correlated, the standard multiple testing

Figure 5.1: Nested testing hypotheses



procedures may accept all Elia's coefficients for older adults, which is actually the opposite to the empirical findings (Wang et al., 2005, 2007 and 2010; Bosy et al., 2003; Gallagher et al., 2000; Wakabayashi et al., 2002), with a large probability.

Therefore, in this chapter, we consider the following sequence of hypotheses

$$\begin{array}{rcl}
\boldsymbol{H}^0 : \text{for all i's, } R_i = R_{i0} & vs & \boldsymbol{K}^0 : \text{at least 1 i, } R_i \neq R_{i0} \\
\boldsymbol{H}^1 : \text{exactly 1 i, } R_i \neq R_{i0} & vs & \boldsymbol{K}^1 : \text{at least 2 i's, } R_i \neq R_{i0} \,\big|\boldsymbol{K}^0 \\
 & \vdots & \\
 & \vdots & \\
\boldsymbol{H}^{p-1} : \text{exactly p-1 i's, } R_i \neq R_{i0} & vs & \boldsymbol{K}^{p-1} : \text{for all i's, } R_i \neq R_{i0} \quad \big|\boldsymbol{K}^{p-2}
\end{array} \tag{5.3}$$

where $\boldsymbol{H}^i$ is nested in $\boldsymbol{K}^{i-1}$ [see Figure 5.1 for illustration]. Obviously, a sequential decision-making is desirable for this situation.

To test (5.3), we developed a step-wise multiple testing procedure which is very intuitive and powerful. The new method is based on performing block-wise marginal regressions repeatedly, which makes it essentially different from all existing methods, although our procedure bears the similarity with MRD method and SURE independence screening method (Fan and Lv, 2008). The new method, which is called the minimax of marginal regression distances (MMRD) step-down procedure, have the following advantages. At first, the MMRD procedure controls the family-wise error rate in the strong sense meanwhile it is more powerful to detect the possible deviation from null hypotheses. Secondly, the MMRD procedure breaks down the dependent structure among the explanatory variables through block-wise marginal regressions. It reduces the effect of the troublesome collinearity among the covariates. Specifically, the earlier the individual hypothesis is selected, the less the collinearity is and thus the more powerful the test of that hypothesis becomes. Thirdly, the MMRD testing statistics are minimax solutions to some optimization problems and thus achieve certain optimality.

In the next section, we define the marginal statistics as well as the block-wise marginal regression, which are the basis of the MMRD step-down procedure. In Section 5.3, we describe the MMRD procedure and use the new method to evaluate Elia's resting organ/tissue metabolic rates for different groups of people. In Section 5.4, some distributional properties of MMRD statistics are developed and the validity of the MMRD method is established. In Section 5.5, other properties of the MMRD procedure are discussed.

5.2 Basics

5.2.1 Marginal statistics

Let us now focus on the following linear regression model:

$$Y = X\beta + \epsilon. \tag{5.4}$$

Here $Y = (Y_1, ..., Y_n)^T$ is the vector of n sample responses where n is the sample size; $X = (X^{(1)}, ..., X^{(p)})$ is the matrix of p predictors, i.e., $X^{(i)}$'s; ϵ is a normal random vector with mean 0 and a diagonal covariance matrix with all diagonal elements being σ^2.

The standard estimation procedure for (5.4) is to perform a least squares (LS) fit to (5.4) for the whole data. The resulting least squares estimate is

$$\widetilde{\beta} = [X^T X]^{-1} X^T Y, \tag{5.5}$$

which is the best linear unbiased estimator (BLUE). In other word, the variance of the $\widetilde{\beta}$ achieves the lower bound for those of unbiased linear estimators which are all in the form of $T(X)Y$. Even though the least squares estimator achieves the lower bound, the component-wise variances are exceptionally large when $X^T X$ is nearly singular. Therefore the power of the test statistics based on the LS fit is not powerful enough to detect potential bias of Elia's coefficients.

Recalling the first step of forward stage-wise regression method (Goldberger, 1961; Goldberger and Jochemes, 1961; Freund et al., 1961a, b) and the SURE independence screening method (Fan and Lv, 2008), define the following marginal statistics,

$$\hat{\beta}^{(i)} = \left[\left[X^{(i)}\right]^T X^{(i)} \right]^{-1} \left[X^{(i)}\right]^T \left[Y - X^{(-i)} \beta_0^{(-i)} \right], \tag{5.6}$$

where $X^{(i)}$ is the sample vector of i^{th} predictor, $X^{(-i)}$ is the sub-matrix by deleting the i^{th} column vector in X and $\beta_0^{(-i)}$ is the sub-vector by deleting the i^{th} element in β_0, the hypothesized values. Actually, $\hat{\beta}^{(i)}$ is the least squares estimate for the following regression problem

$$Y - X^{(-i)} \beta_0^{(-i)} = X^{(i)} \beta^{(i)} + \epsilon. \tag{5.7}$$

Since

$$\begin{aligned}\hat{\beta}^{(i)} &= \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1} \left[X^{(i)}\right]^T \left[Y - X^{(-i)}\beta_0^{(-i)}\right] \\ &= \beta^{(i)} + \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1} \left[X^{(i)}\right]^T X^{(-i)} \left[\beta^{(-i)} - \beta_0^{(-i)}\right] \\ &+ \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1} \left[X^{(i)}\right]^T \epsilon,\end{aligned} \tag{5.8}$$

it is obvious that the covariance matrix of $\hat{\beta} = (\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(p)})^T$ is

$$\widehat{\Sigma} = \sigma^2 G \left[X^T X\right] G \tag{5.9}$$

where G is a diagonal matrix,

$$G = diag\left\{\left[\left[X^{(1)}\right]^T X^{(1)}\right]^{-1}, \cdots, \left[\left[X^{(p)}\right]^T X^{(p)}\right]^{-1}\right\}.$$

Its proof is given in Section 5.5. Then the variance of $\hat{\beta}^{(i)}$ is

$$\sigma^2 \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1}.$$

For $\tilde{\beta}^{(i)}$, the least squares estimator, it can be shown that its variance is

$$\sigma^2 \left[\left[X^{(i)}\right]^T X^{(i)} - c\right]^{-1}$$

where

$$c = \left[X^{(i)}\right]^T X^{(-i)} \left[\left[X^{(-i)}\right]^T X^{(-i)}\right]^{-1} \left[X^{(-i)}\right]^T X^{(i)} \geq 0$$

Therefore, the variance of the marginal statistics $\hat{\beta}^{(i)}$ is smaller than that of the LS estimate $\tilde{\beta}^{(i)}$.

Remark 5.1. By definition,

$$\widetilde{\beta} = \beta + [X^T X]^{-1} X^T \epsilon, \tag{5.10}$$

and

$$\begin{aligned}\widehat{\beta}^{(i)} &= \beta^{(i)} + \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1} \left[X^{(i)}\right]^T \epsilon \\ &\quad + \left[\left[X^{(i)}\right]^T X^{(i)}\right]^{-1} \left[X^{(i)}\right]^T X^{(-i)} \left[\beta^{(-i)} - \beta_0^{(-i)}\right].\end{aligned} \tag{5.11}$$

Therefore, the least squares estimate $\tilde{\beta}$ is unbiased but the marginal fit $\hat{\beta}$ is biased. Specifically, if the true parameter β is not equal to the hypothesized value β_0, then the third part, which contains $\beta^{(-i)} - \beta_0^{(-i)}$, in the above decomposition of $\hat{\beta}^{(i)}$, introduces bias to the estimation of β. In other words, $\widehat{\beta}$ is no longer a pivotal quantity and thus the interpretation of the resulting confidence interval is different. This is a major difficulty in doing interval estimation for multiple testing problems. The resulting confidence intervals or the P-values should be interpreted as conditional one [see Cohen et al. (2009)]. However, $\hat{\beta}$ is conditional unbiased in the sense that

$$E\left[\hat{\beta}^{(i)} \middle| \beta^{(-i)} = \beta_0^{(-1)}\right] = \beta^{(i)}. \tag{5.12}$$

5.2.2 Chi-squared statistics

From $\tilde{\beta}$ and $\hat{\beta}$, Chi-squared statistics can be constructed in a sandwich form as follows,

$$\widehat{\eta} = (\widehat{\beta} - \beta_0)^T \Sigma_{\widehat{\beta}}^{-1} (\widehat{\beta} - \beta_0), \tag{5.13}$$

and

$$\widetilde{\eta} = (\widetilde{\beta} - \beta_0)^T \Sigma_{\widetilde{\beta}}^{-1} (\widetilde{\beta} - \beta_0). \tag{5.14}$$

It is claimed that the two Chi-squared tests have the same power for any specific alternative hypotheses $K : \beta = \beta^*$. The proof requires some calculation and thus is provided in Section 5.5 with details.

Remark 5.2. This property reveals that although the marginal statistics are less variable, it is still not aggressive to reject the null hypotheses (5.2) simultaneously in the sense that the resulting ellipsoid statistics has the same power of detecting the possible deviation from null as that of the ellipsoid statistics based on LS estimate.

5.3 Method and Application

5.3.1 Generalized marginal statistics

In the last section, we described the marginal statistics and discussed its properties. In this section, the definition of marginal statistics is generalized.

Instead of fixing $(p-1)$ coefficients at their hypothesized values each time such as in the model (5.7), let us fix $0 < m < p-1$ coefficients at hypothesized values and estimate others by a least squares fit. Specifically, for a given set $B = \{b_1, ..., b_{m+1}\}$ whose cardinality $|B| = m+1$, we consider the following models

$$Y - X^{(B\backslash\{b_j\})}\beta_0^{(B\backslash\{b_j\})} = X^{(\{b_j\}\cup B^c)}\beta^{(\{b_j\}\cup B^c)} + \epsilon, \quad j = 1, ..., m+1. \tag{5.15}$$

Here for a given set $A = \{a_1, ..., a_l\}$, $X^{(A)} = (X^{(a_1)}, X^{(a_2)}, ..., X^{(a_l)})$, which is a n-by-l matrix, and $\beta^{(A)} = (\beta^{(a_1)}, \beta^{(a_2)}, ..., \beta^{(a_l)})^T$, which is a vector of l elements.

For each $k \in \{1, ..., m+1\}$, we use the least squares method to fit one model in (5.15) when $j = k$ and get an estimate for $\beta^{(\{b_k\}\cup B^c)}$ as

$$\begin{aligned}\bar{\beta}^{(\{b_k\}\cup B^c)} = &\left[\left[X^{(\{b_k\}\cup B^c)}\right]^T X^{(\{b_k\}\cup B^c)}\right]^{-1} \\ &\times \left[X^{(\{b_k\}\cup B^c)}\right]^T \left[Y - X^{(B\backslash\{b_k\})}\beta_0^{(B\backslash\{b_k\})}\right].\end{aligned} \tag{5.16}$$

As in (5.11), $\bar{\beta}^{(\{b_k\}\cup B^c)}$ can be decomposed as follows,

$$\begin{aligned}\bar{\beta}^{(\{b_k\}\cup B^c)} = \beta^{(\{b_k\}\cup B^c)} + \left[\left[X^{(\{b_k\}\cup B^c)}\right]^T X^{(\{b_k\}\cup B^c)}\right]^{-1} \left[X^{(\{b_k\}\cup B^c)}\right]^T \epsilon \\ + \left[\left[X^{(\{b_k\}\cup B^c)}\right]^T X^{(\{b_k\}\cup B^c)}\right]^{-1} \left[X^{(\{b_k\}\cup B^c)}\right]^T X^{(B\backslash\{b_k\})} \left[\beta^{(B\backslash\{b_k\})} - \beta_0^{(B\backslash\{b_k\})}\right].\end{aligned} \tag{5.17}$$

Therefore the variance of $\bar{\beta}^{(\{b_k\}\cup B^c)}$ is

$$\sigma^2 \left[\left[X^{(\{b_k\}\cup B^c)} \right]^T X^{(\{b_k\}\cup B^c)} \right]^{-1},$$

and $\bar{\beta}^{(\{b_k\}\cup B^c)}$ is still conditional unbiased in the sense that

$$E\left[\bar{\beta}^{(\{b_k\}\cup B^c)} \middle| \beta^{(B\backslash\{b_k\})} = \beta_0^{(B\backslash\{b_k\})} \right] = \beta^{(\{b_k\}\cup B^c)}. \tag{5.18}$$

Hence for a given set B, let

$$\hat{\beta}^{(B)} = (\bar{\beta}^{(b_1)}, \bar{\beta}^{(b_2)}, ..., \bar{\beta}^{(b_{m+1})})^T \tag{5.19}$$

be the generalized estimator of $\beta^{(B)}$.

5.3.2 MMRD step-down procedure

In this section, we combine the generalized marginal statistics with a maximizing step-down procedure to establish a sequential decision-making for testing the multiple hypotheses (5.3). To describe the MMRD step-down testing procedure, we firstl define the testing statistics, the minimax of marginal regression distances, as well as the critical values and testing functions for each individual hypothesis in (5.3).

Definition 5.1. For simplicity, the following definitions are used in the MMRD step-down procedure.

1. For any matrix A, $diag(A)$ is a diagonal matrix with diagonal elements the same as A's.

2. For a diagonal matrix $D = diag\{d_1, ..., d_p\}$, $D^{1/2} \triangleq diag\{\sqrt{d_1}, ..., \sqrt{d_p}\}$ and $D^{-1/2} \triangleq diag\{1/\sqrt{d_1}, ..., 1/\sqrt{d_p}\}$.

3. $U_j^B \triangleq \left[diag\left(COV(\hat{\beta}^{(B)})\right)\right]^{-1/2} \times \left[\hat{\beta}^{(B)} - \beta_0^{(B)}\right]$, for any index set $B \in \{B' : B' \subset \{1, 2, ..., p\}, |B'| = p + 1 - j\}$. Here $\hat{\beta}^{(B)}$ is defined as in (5.19) and $COV(\hat{\beta}^{(B)})$ is the covariance matrix of $\hat{\beta}^{(B)}$.

4. The test statistics U_j's are defined as follows.

   For $j = 1$,
   $$U_1 \triangleq \left|U_1^B\right|_\infty;$$
   For $j \geq 2$,
   $$U_j \triangleq \min_{|B|=p+1-j} \left|U_j^B\right|_\infty$$
   where $|v|_\infty$ is the l_∞ norm of vector v.

5. The critical values C_j's are defined as follows.

   For $j = 1$, C_1 is the upper α quantile of U_1 such that
   $$P\left(U_1 \geq C_1 \middle| H^0\right) = \alpha;$$
   For $j \geq 2$, $C_j \triangleq \max_{|B|=p+1-j} C_j^B$ where C_j^B is the upper α quantile of $\left|U_j^B\right|_\infty$ such that
   $$P\left(\left|U_j^B\right|_\infty \geq C_j^B \middle| \beta^{(B)} = \beta_0^{(B)}\right) = \alpha.$$

6. For H^{j-1} in (5.3), define testing functions $\phi_j(X, Y; \alpha) = \mathrm{I}(U_j > C_j)$.

Remark 5.3. U_j^B is a random vector of $(p+1-j)$ correlated standard normal random variables given $\beta^{(B)} = \beta_0^{(B)}$ and without any constrain on $\beta^{(B^c)}$. Specifically, U_1^B is a normal random vector with mean 0 and covariance matrix $G^{1/2}\left[X^T X\right]G^{1/2}$ given $\beta = \beta_0$. Here G is defined in (5.9).

Then based on the testing functions ϕ_j's, the testing procedure is defined as follows.

Algorithm 5.1. *(MMRD)*

1. *If $\phi_1 = 0$, then accept H^0 and terminate the procedure. Otherwise reject H^0 and continue the following steps.*

2. *If $\phi_j = 0$ $(j \geq 2)$, then accept H^{j-1} and terminate the procedure. Otherwise reject H^{j-1} and continue to test H^j.*

3. *Repeat step 2 until $j = p$.*

5.3.3 Application to the REE studies

We applied the MMRD step-down testing procedure to evaluating Elia's metabolic rates of major organs and tissues for model (5.1). The distribution of U_1 is simulated for REE model based on data from Wang and et al. (2011).

The test statistics calculated from MMRD procedure are summarized in Table 5.1. From Table 5.1 we know that there are exactly two organ/tissue metabolic rates different from Elia's coefficients for the elderly. Those two major organs/tissues are probably kidneys and liver. This finding matches the results in Wang et al. (2005, 2007, 2010), Bosy et al. (2003), Gallagher et al. (2000) and Wakabayashi et al. (2002)

5.4 Theory

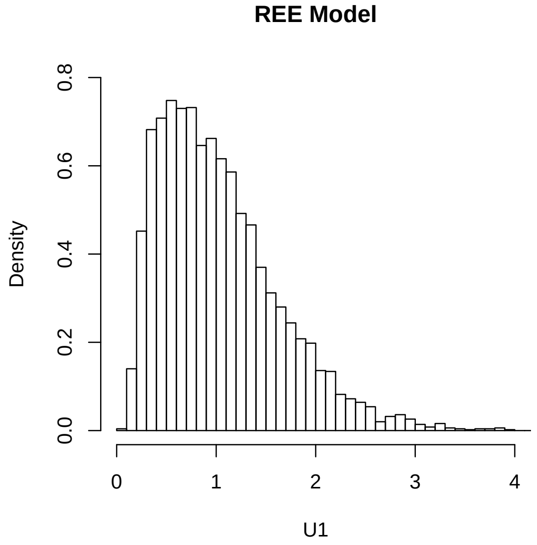
5.4.1 Controlling the error rate

In this section, we show that the MMRD procedure controls the family-wise error rate (FWER) in a natural way. Recall the definition of family-wise error rate in the strong sense as follows.

Table 5.1: The testing results of MMRD

	All	Young	Mid-age	Elderly
Stage	(n=131)	(n=43)	(n=51)	(n=37)
1	2.433	[1.026]	[0.933]	5.285
$(a(p))$	(Kidneys)	(Liver)	(Kidneys)	(Kidneys)
2	3.620	0.378	2.924	2.807
$(a(p))$	(Liver)	(Brain)	(Liver)	(Liver)
3	[1.054]	0.260	0.694	[1.338]
$(a(p))$	(Brain)	(SM)	(Brain)	(Heart)
4	0.654	0.238	0.630	1.502
$(a(p))$	(Res)	(Res)	(Heart)	(Brain)
5	0.610	0.203	0.761	0.462
$(a(p))$	(AT)	(Heart)	(SM)	(AT)
6	0.604	0.053	0.506	0.210
$(a(p))$	(Heart)	(AT)	(AT)	(Res)
7	0.657	0.054	0.030	0.056
$(a(p))$	(SM)	(Kidneys)	(Res)	(SM)

Figure 5.2: The distribution of U_1 based on the REE data



Definition 5.2. For the null hypotheses H^0 to H^{p-1} and corresponding testing function ϕ_1 to ϕ_p, if the following condition is always satisfied for any subset I of $\{1, 2, ..., p\}$,

$$Pr\left(\phi_j(X, Y; \alpha) = 1 \middle| \text{ all } H^{i-1}, i \in I\right) \leq \alpha, \forall j \in I, \tag{5.20}$$

then it is said that $\phi_j(X, Y; \alpha)$'s control the family-wise error rate at the level of α for testing the multiple hypotheses (5.3).

Theorem 5.1. *The MMRD step-down procedure as defined in Algorithm 5.1 controls the family-wise error rate at the level of α.*

Proof. Since the null hypotheses H^j (j=0,...,p-1) are disjoint, the condition of controlling the family-wise error rate reduces to the following one,

$$\forall j \in I,\ Pr\left(\phi_j(X, Y; \alpha) = 1 \middle| H^{j-1}\right) \leq \alpha. \tag{5.21}$$

Let us show the above inequality for $j = 1$ and $j \geq 2$ separately.

For $j = 1$, recalling the definition of U_1 and C_1, we have

$$Pr\left(\phi_1(X, Y; \alpha) = 1 \middle| H^0\right) = Pr\left(\left|U_j^{\{1,2,\dots,p\}}\right|_\infty > C_1 \middle| \ \beta = \beta_0\right) = \alpha.$$

For $j \geq 2$, recalling the definition of U_j and C_j, we have

$$\begin{aligned} &Pr\left(\phi_j(X, Y; \alpha) = 1 \middle| H^{j-1}\right) = Pr\left(\min_{|B|=p+1-j} \left|U_j^B\right|_\infty > C_j \middle| H^{j-1}\right) \\ =&\sum_{|B|=p+1-j} Pr\left(\min_{|B|=p+1-j} \left|U_j^B\right|_\infty > C_j \middle| \ \beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c\right) \\ &\qquad \times Pr\left(\beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c \middle| H^{j-1}\right) \end{aligned}$$

where the second equality is due to the law of total probability and the truth that $\{\beta : \beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c\} \subset H^{j-1}$ for any set B satisfying $B \subset \{1, 2, ..., p\}$ and $|B| = p+1-j$. Since the following is true for any $B' \in \{B : B \subset \{1, 2, ..., p\}, |B| = p + 1 - j\}$

$$\left(\min_{|B|=p+1-j} \left|U_j^B\right|_\infty > C_j\right) \subset \left(\left|U_j^{B'}\right|_\infty > C_j\right) \subset \left(\left|U_j^{B'}\right|_\infty > C_j^{B'}\right),$$

we have

$$\begin{aligned} &Pr\left(\phi_j(X, Y; \alpha) = 1 \middle| H^{j-1}\right) \\ \leq &\sum_{|B|=p+1-j} Pr\left(\left|U_j^B\right|_\infty > C_j^B \middle| \ \beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c\right) \\ &\qquad \times Pr\left(\beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c \middle| H^{j-1}\right) \\ =&\alpha \times \sum_{|B|=p+1-j} Pr\left(\beta^{(B)} = \beta_0^{(B)}, \beta^{(l)} \neq \beta_0^{(l)}, \forall l \in B^c \middle| H^{j-1}\right) = \alpha \end{aligned}$$

where the second last equality is due to the definition of C_j^B

□

5.4.2 Distributional properties of $\left|U_j^B\right|_\infty$

To make the sequential decisions of whether accepting and terminating the MMRD procedure or rejecting the current marginal null hypothesis and carrying on the MMRD procedure, the distributional theory of the maximal marginal regression statistics, $\left|U_j^B\right|_\infty$, is developed in this subsection. Note that in the following derivation, let us only focus on two-sided test. We are going to examine the distribution of U_1. The distributional theories for $\left|U_j^B\right|_\infty$'s, $j \geq 2$, are the same.

By definition, $U_1 = \left|U_1^{\{1,2,\ldots,p\}}\right|_\infty$ is the maximal absolute value of p correlated standard normal distributions. Therefore, let us consider the following setup.

Suppose $(W_1, ..., W_p)$ follows multivariate normal distribution $N(0, \mathbf{D})$ where $\mathbf{D}$ is the variance-covariance matrix with unit diagonal elements. Let $R_1, ..., R_p$ be the rank statistics of $|W_i|_{i=1}^p$ s.t. $|W_{R_1}| \leq |W_{R_2}| \leq ... \leq |W_{R_p}|$. Define the anti-rank function as $a(\cdot)$ s.t. $a(R_i) = i, \forall i$.

Then at first, let us focus on $W_{a(p)}$. Actually

$$W_{a(p)} = \begin{cases} W_{\min}, & \text{if } |W_{\min}| > |W_{\max}| \\ W_{\max}, & \text{if } |W_{\max}| > |W_{\min}| \end{cases},$$

where $W_{\max}$ is the biggest order statistics of W and vice versa. Then by law of total probability,

$$\begin{aligned} P(W_{a(p)} \leq t) = & P(W_{\min} < t \big| |W_{\min}| > |W_{\max}|) \times p_1 \\ & + P(W_{\max} < t \big| |W_{\max}| > |W_{\min}|) \times (1 - p_1). \end{aligned}$$

where $p_1 = P(|W_{\min}| > |W_{\max}|)$. Since $(W_1, ..., W_p)$ is symmetric about 0, $W_{\min} \overset{d}{=} -W_{\max}$. So $p_1 = 1/2$. Hence, the density function of $W_{a(p)}$ is a mixture of two normal density as follows,

$$f_{W_{a(p)}}(t) = \frac{1}{2} \times f_{W_{\min}\Big||W_{\min}|>|W_{\max}|}(t) + \frac{1}{2} \times f_{W_{\max}\Big||W_{\max}|>|W_{\min}|}(t).$$

Therefore, $f_{W_{a(p)}}(t)$ is bimodal.

Example 5.1. (D is an identity matrix) If the variance-covariance matrix is diagonal (independent case), then from lemmas in the appendices, we can compute $F_{W_{a(p)}}(t) \equiv P(W_{a(p)} \leq t)$ as,

$$F_{W_{a(p)}}(t) = \frac{1}{2} - \frac{1}{2}(1 - 2\Phi(t \wedge 0))^p + \frac{1}{2}(2\Phi(t \vee 0) - 1)^p \tag{5.22}$$

Therefore, the critical value (two-sided test) C_1^* for U_1 as a significant level of α is

$$C_1^* = \Phi^{-1}(\frac{1}{2} + \frac{1}{2}(1 - \alpha)^{1/p}), \tag{5.23}$$

where $\Phi^{-1}(\cdot)$ is the inverse-CDF of standard normal distribution.

Next, let us focus on $U_1 = |W_{a(p)}|$.

Example 5.2. ($p = 2$) If the variance-covariance matrix is non-diagonal (dependent case), it is not easy to compute or even approximate the cumulative distribution function of U_1. Instead, there are some interesting results as follows.

Lemma 5.1. *Suppose $p = 2$. Then the probability density function of U_1, denoted as $f_{U_1}(t; \rho)$, equals*

$$2\phi(t) \times \left[\Phi\left(\sqrt{\frac{1+\rho}{1-\rho}}t\right) - \Phi\left(-\sqrt{\frac{1-\rho}{1+\rho}}t\right)\right] \times I_{\{t>0\}}, \tag{5.24}$$

where $\phi(t)$ and $\Phi(t)$ are respectively probability density function and cumulative distribution function of standard normal distribution; ρ is the correlation or off-diagonal entry of $D_{2\times2}$.

Remark 5.4. Then from this lemma, we know that there is no-close form expression to compute the cumulative distribution function of U_1; instead, we need numerical integrations or other approximation methods such as MCMC simulations.

Corollary 5.1. *Suppose $p = 2$, then $f_{U_1}(t;\rho) = f_{U_1}(t;-\rho)$.*

Lemma 5.2. *Suppose $p = 2$ and $t > \sqrt{2}$, then $f_{U_1}(t;\rho) \leq f_{U_1}(t;\rho = 0)$. Furthermore, for $\rho \in (0, +\infty)$, $f_{U_1}(t;\rho)$ is decreasing in ρ; for $\rho \in (-\infty, 0)$, $f_{U_1}(t;\rho)$ is increasing in ρ.*

Lemma 5.3. *Suppose $p = 2$, then $q_{.05}(\rho) = q_{.05}(-\rho)$ and $q_{.05}(|\rho|)$ is decreasing in $|\rho|$.*

Remark 5.5. The influence of the collinearity can be explained as follows. From Lemma 5.3 and Eqn. (5.9), we know how the correlation between predictors influence the decision making of MMRD procedure in a special case of $p = 2$. The more dependency in predictors (the bigger $|\rho|$ is), the more difficult it is to reject the null. This makes sense because the evidence of a deviation from null hypothesis becomes weaker when the association between predictors turns higher. The 'significant' evidence may be not that significant because of the strong linear dependency between predictors.

Example 5.3. $(p > 2)$

Let us first build a strict lower bound for $q_{.05}$, the upper-0.05 quantile of U_1.

Lemma 5.4. *(A strict lower bound for $q_{.05}$ of U_1) Let $q_{.05}$ be the upper-0.05 quantile of U_1. Then for $p \geq 2$, we have $q_{.05} \geq 1.96$, or*

$$P(U_1 \geq 1.96) \geq 0.05. \tag{5.25}$$

Actually, for dimension $p \geq 2$, $P(|W_{a(p)}| \geq 1.96) > 0.05$ and $q_{.05} > 1.96$. The lower bound 1.96 cannot be improved; in other word, it is a strict lower bound for

Figure 5.3: The impact of correlation on the distribution of U_1

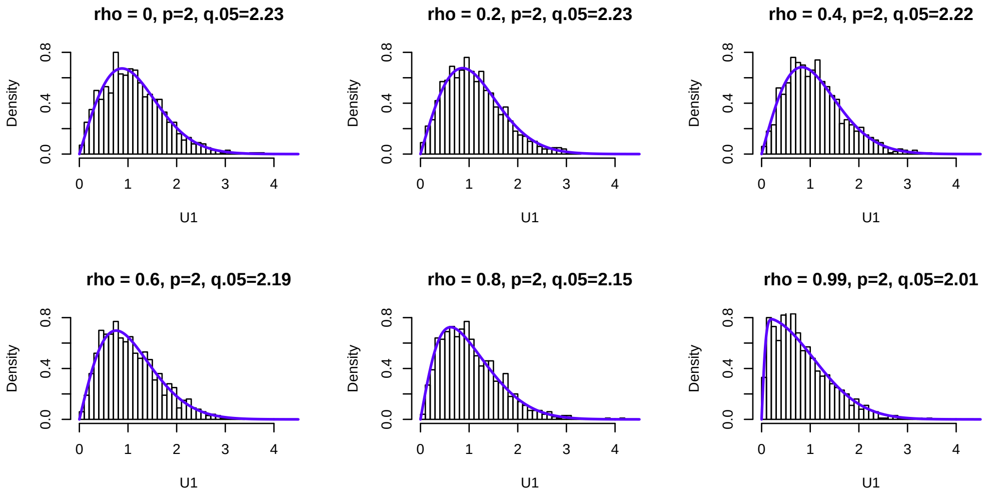


Table 5.2: The simulated quantiles

correlation ρ^*	0.99	0.999	0.9999	0.99999
simulated $q_{.05}$	2.085	1.997	1.970	1.960

$q_{.05}$. This is illustrated by an example and simulations, which also sheds light on the situation under which $q_{.05}$ is approaching 1.96 for a fix dimension parameter p.

We give an example of approaching the 1.96 lower bound here. Without loss of generality, let us assume the dimension p is 7, the same as that of REE model. For the covariance matrix D, let us assume that all of its off-diagonal element is ρ^*. For big enough ρ^*, $D(\rho^*)$ is positive-definite. Let ρ^* be 0.99, 0.999, 0.9999, and 0.99999. The corresponding $q_{.05}$'s are estimated by simulations (sample size $N = 5,000$; number of replications $M = 5,000$) and summarized in Table 5.2.

We then constructs a strict upper bound for $q_{.05}$ of U_1 based on Khatri-Sidak inequality (Khatri, 1967; Sidak, 1967 and 1968) and show how the upper bound is obtained.

Let us first introduce Khatri-Sidak inequality in the following proposition, which was proved by Khatri (1967) and Sidak (1967, 1968) .

Proposition 5.1. *(Khatri-Sidak Inequality) Suppose* $(V_1, ..., V_p)$ *is a centered, Gaussian random vector, then*

$$P(\max_{1\leq i\leq p} |V_i| \leq t) \geq P(|V_1| \leq t) \times P(\max_{2\leq i\leq p} |V_i| \leq t). \tag{5.26}$$

A short proof can be found in Li and Shao (2001).

By mathematical induction and (5.26), we get

$$P\left(\max_{1\leq i\leq p} |W_i| \leq t\right) \geq \left[P(|W_1| \leq t)\right]^p. \tag{5.27}$$

Consequently, we have the following lemma.

Lemma 5.5. *Let* $q_{.05}$ *be the upper*-0.05 *quantile of* U_1*. Then* $q_{.05} \leq C_1^*$*, which is defined in* (5.23)*.*

In general, the exact value of C_1 in Algorithm 5.1 (MMRD) can be determined for a given correlation matrix D by simulations.

5.5 Other Properties

5.5.1 Variance of marginal fit

In this section, we show the computation for the covariance matrix of $\hat{\beta}^{(B)}$ for $|B| = p + 1 - j$ and $j = 1, ..., p$ where $\hat{\beta}^{(B)}$ is defined as in (5.11) and (5.19).

For $j = 1$, according to equation (5.11),

$$\begin{aligned}\widehat{\beta}^{(i)} - \beta_0^{(i)} = &\left[\left[X^{(i)}\right]^T X^{(i)})\right]^{-1} \left[X^{(i)}\right]^T X\left[\beta - \beta_0\right] \\ &+ \left[\left[X^{(i)}\right]^T X^{(i)})\right]^{-1} \left[X^{(i)}\right]^T \epsilon.\end{aligned}$$

Therefore, by block-wise matrix multiplication,

$$\widehat{\beta} - \beta_0 = G\left[X^T X\right][\beta - \beta_0] + G X^T \epsilon \tag{5.28}$$

where G is a diagonal matrix,

$$G = diag\left\{\left[\left[X^{(1)}\right]^T X^{(1)}\right]^{-1}, \cdots, \left[\left[X^{(p)}\right]^T X^{(p)}\right]^{-1}\right\}.$$

Therefore, the variance of marginal fit $\widehat{\beta}$ is $\sigma^2 G\left[X^T X\right] G$ as in (5.9).

5.5.2 Chi-squared statistics

It has been claimed in Subsection 5.2.2 that the two sandwich statistics $\widehat{\eta}$ and $\widetilde{\eta}$ in (5.13) and (5.14) have the same power under any alternative hypothesis $K : \beta = \beta^*$.

Actually, from (5.28) we know that both $\widehat{\eta}$ and $\widetilde{\eta}$ are distributed as non-central chi-squared r.v. Hence we only need to show that $\widehat{\eta}$ has the same non-central parameter as $\widetilde{\eta}$.

From (5.10), the non-central parameter of $\widetilde{\eta}$ is

$$[\beta^* - \beta_0]^T \left[X^T X\right][\beta^* - \beta_0]/\sigma^2$$

under alternative K.

From (5.9) and (5.28), the non-central parameter of $\widehat{\eta}$ is

$$\begin{aligned} & [\beta^* - \beta_0]^T \left[X^T X\right] G \times \Sigma_{\widehat{\beta}}^{-1} \times G\left[X^T X\right][\beta^* - \beta_0] \\ = & [\beta^* - \beta_0]^T \left[X^T X\right][\beta^* - \beta_0]/\sigma^2 \end{aligned}$$

Therefore, $\widehat{\eta}$ has the same non-central parameter as $\widetilde{\eta}$.

5.5.3 Distributional theories

Proof of Lemma 5.2 . According to Lemma 5.1,

$$\frac{\partial}{\partial\rho} f_{U_1}(t;\rho) = \frac{2t\phi^2(t)}{\sqrt{1-\rho^2}} \times g(t,\rho) \times I_{\{t>0\}} \tag{5.29}$$

where

$$g(t,\rho) = \frac{e^{-\frac{t^2}{1-\rho}}}{1-\rho} - \frac{e^{-\frac{t^2}{1+\rho}}}{1+\rho}. \tag{5.30}$$

Then we just need to prove that $g(t,\rho) < 0$ for $\rho > 0$ and $g(t,\rho) > 0$ for $\rho < 0$ when $t > \sqrt{2}$. From Eqn. (5.30), we know $g(t,\rho) = -g(t,-\rho)$ and $g(t,0) = 0$. Therefore, it is sufficient to prove $\partial g(t,\rho)/\partial\rho < 0$ for $\rho > 0$ and $t > \sqrt{2}$. After some calculation, we get

$$\frac{\partial}{\partial\rho} g(t,\rho) = \frac{e^{-\frac{t^2}{1-\rho}}}{(1-\rho)^2}\left[1 - \frac{t^2}{1-\rho}\right] - \frac{e^{-\frac{t^2}{1+\rho}}}{(1+\rho)^2}\left[\frac{t^2}{1+\rho} - 1\right].$$

Since $t^2 > 2$ and $0 < (1 \pm \rho) < 2$, we obtain that $\partial g(t,\rho)/\partial\rho < 0$ for $\rho > 0$ and $t > \sqrt{2}$. This finishes the proof.

Proof of Lemma 5.3 . Since

$$\begin{aligned} &\{W : |W_1| > t, |W_2| > |W_1|\} \\ &\qquad \in \{W : |W_2| > t, |W_2| > |W_1|\} \end{aligned}$$

and

$$\begin{aligned} P(U_1 > t) &= P(|W_1| > t, |W_1| > |W_2|) \\ &\quad + P(|W_2| > t, |W_2| > |W_1|), \end{aligned}$$

we get $P(U_1 > t) > P(|W_1| > t)$. Therefore, the upper 0.05-quantile for U_1 is bigger than 1.96, the upper 0.05-quantile for standard normal distribution.

Then by Lemma 5.2, it is not difficult to prove that $q_{.05}(|\rho|)$ is decreasing in $|\rho|$.

Proof of Lemma 5.4 . By definition of anti-rank function $a(\cdot)$, we observe

$$\big\{\boldsymbol{\omega} : |W_1| \geq 1.96\big\} \subseteq \big\{\boldsymbol{\omega} : |W_{a(p)}| \geq 1.96\big\}.$$

Hence $P(|W_{a(p)}| \geq 1.96) \geq 0.05$. Therefore $q_{.05} \geq 1.96$.

Bibliography

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**, 289-300.

[2] Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Med.*, **2**, 273-277.

[3] Bosy-Westphal, A., Eichhorn, C., Kutzner, D., Illner, K., Heller, M. and M´uller, M. J. (2003). The age-related decline in resting energy expenditure in humnas is due to the loss of fat-free mass and to alterations in its metabolically active components. *J. Nutr.*, **133** 2356-2362.

[4] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc. B*, **26**, 211-252.

[5] Brown, B. M. and Wang, Y. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, **92**, 149-158.

[6] Brown, B. M. and Wang, Y. (2007). Induced smoothing for rank regression with censored survival times. *Statist. Med.*, **26**, 828-836.

[7] Chen, S. (2002). Rank estimation of transformation models. *Econometrica*, **70**, 1683-1697.

[8] Cohen, A., Sackrowitz, H. B. and Xu, M. (2009). A new multiple testing method in the dependent case. *Ann. Statist.*, **37**, 1518-1544.

[9] Cox, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc. B*, **34**, 187-220.

[10] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

[11] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, **7**, 1-26.

[12] Elia, M. (1992). Organ and tissue contribution to metabolic rate. *Energy metabolism: tissue determinants and cellular corollaries*, 61-80, New York, NY: Raven Press.

[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

[14] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional Hazards Model and Frailty Model. *Ann. Statist.*, **30**, 74-99.

[15] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, **70**, 849-911.

[16] Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109-148.

[17] Freund, R. J., Vail, R. W. and Clunies-Ross, C. W. (1961). Residual analysis. *J. Amer. Statist. Assoc.*, **56**, 98-104.

[18] Freund, R. J., Vail, R. W. and Clunies-Ross, C. W. (1961). Corrigendum to the article on residual analysis. *J. Amer. Statist. Assoc.*, **56**, 1005.

[19] Friedman, J.H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1-22.

[20] Gallagher, D., Belmonte D., Deurenberg, P., Wang, Z., Krasnow N., Pi-Sunyer, F. X. and Heymsfield S. B. (1998). Organ-tissue mass measurement allows modeling of REE and metabolically active tissue mass. *Am. J. Physiol.*, **275**, 249-258.

[21] Gallagher, D., Allen, A., Wang, Z., Heymsfield S. B., Krasnow N. (2000). Smaller organ tissue mass in the elderly fails to explain lower resting metabolic rate. *Ann. N. Y. Acad. Sci.*, **904**, 449-455.

[22] Goldberger, A. S. (1961). Stepwise least squares: residual analysis and specification error. *J. Amer. Statist. Assoc.*, **56**, 998-1000.

[23] Goldberger, A. S. and Jochemes, D. B. (1961). Note on stepwise least squares. *J. Amer. Statist. Assoc.*, **56**, 105-110.

[24] Gørgens, T. and Horowitz, J. L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *J. Econometrics*, **90**, 155-191.

[25] Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. Econometrics*, **35**, 303-316.

[26] Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251-1271.

[27] Harris, J. A. and Benedict, F. G. (1918). A biometric study of basal metabolism in man. *Proceedings of the National Academy of Sciences*, **4**, 370-373.

[28] Harris, J. A. and Benedict, F. G. (1919). *A biometric study of basal metabolism in man*, 190, Washington, DC: Carnegie Institution.

[29] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65-70.

[30] Jin, Z., Ying, Z. and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, **88**, 381-390.

[31] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data (2nd Ed.).* New York: Wiley.

[32] Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. *J. Econometrics*, **136**, 251-280.

[33] Khatri, C. G. (1967). On certain inequalities for normal distributions and their applications to simultaneous confidence bounds. *Ann. Math. Stat.*, **38**, 1853-1867.

[34] Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications, in the *Handbook of Statistics. Stochastic Processes: Theory and Methods*, **19**, 533-597, Amsterdam: Elsevier.

[35] Maddala, G. S. (1983). Limited dependent and qualitative variables in econometrics. *Econometric Society Monograph No.3.* Cambridge: Cambridge University Press.

[36] McFadden, D. L. (1984). Econometric analysis of qualitative response models. *Handbook of Econometrics, Vol.2*, 1395-1457, Amsterdam: Elsevier.

[37] Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.*, **15**, 780-799.

[38] Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *J. Econometrics*, **25**, 303-325.

[39] Roza, A. M. and Shizgal, H. M. (1984). The Harris Benedict equation reevaluated: resting energy requirements and the body cell mass. *Am. J. Clin. Nutr.*, **40**, 168-182.

[40] Sherman, R. P. (1993). The limit distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123-137.

[41] Sherman, R. P. (1994a). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory*, **10**, 372-395.

[42] Sherman, R. P. (1994b). Maximum inequalities for degenerate U-processes with applications to maximization estimators. *Ann. Statist.*, **22**, 439-459.

[43] Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.*, **62**, 626-633.

[44] Sidak, Z. (1968). On multivariate normal probabilities of rectangules: their dependence on correlations. *Ann. Math. Stat.*, **39**, 1425-1434.

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

[46] Tibshirani, R.J. (1997). The LASSO method for variable selection in the Cox model. *Statist. Med.*, **16**, 385-395.

[47] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-26.

[48] van der Vaart, A.W. (1998). *Asymptotic Statistics.*, New York: Cambridge University Press.

[49] Wakabayashi, H., Nishiyama, Y., Ushivama, T., Maeba, T. and Maeta, H.(2002). Evaluation of the effect of age on functioning hypatocyte mass and liver blood flow using liver scintigraphy in preoperative estimations for surgical patients: comparison with CT volumetry. *J. Surg. Res.*, **106**, 246-253.

[50] Wang, Z., Heshka, S., Zhang, K., Boozer, C.N. and Heymsfield, S.B. (2001). Resting energy expenditure: systematic organization and critique of prediction methods. *Obes. Res.*, **9**, 331-336.

[51] Wang, Z., Heshka, S., Heymsfield, S.B., Shen, W., Gallagher, D. (2005). A cellular level approach to predicting resting energy expenditure across the adult years. *Am. J. Clin. Nutr.* , **81**, 799-806.

[52] Wang, Z., Heshka, S., Wang, J., Gallagher, D., Deurenberg, P., Zhao, C. and Heymsfield, S.B. (2007). Metabolically active portion of fat-free mass: a cellular body composition level modeling analysis. *Am. J. Physiol. Endocrinol. Metab.*, **292**, 49-53.

[53] Wang, Z., Heymsfield, S.B., Ying, Z., Pierson, R.N.Jr., Gallagher, D. and Gidwani, S. (2010). A cellular level approach to predicting resting energy expenditure: Evaluation of applicability in adolescents. *Am. J. Hum. Biol.*, **22**, 476-483.

[54] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, **67**, 301-320.