Electronic Journal of Statistics
Vol. 11 (2017) 4347–4403
ISSN: 1935-7524
DOI: 10.1214/17-EJS1352

A provable smoothing approach for high dimensional generalized regression with applications in genomics

Fang Han

Department of Statistics, University of Washington, Seattle, WA 98195, USA
e-mail: fanghan@uw.edu

Hongkai Ji, Zhicheng Ji

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
e-mail: hji@jhu.edu; zji4@jhu.edu

and

Honglang Wang

Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA
e-mail: hlwang@iupui.edu

Abstract: In many applications, linear models fit the data poorly. This article studies an appealing alternative, the generalized regression model. This model only assumes that there exists an unknown monotonically increasing link function connecting the response Y to a single index $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}^*$ of explanatory variables $\boldsymbol{X} \in \mathbb{R}^d$. The generalized regression model is flexible and covers many widely used statistical models. It fits the data generating mechanisms well in many real problems, which makes it useful in a variety of applications where regression models are regularly employed. In low dimensions, rank-based M-estimators are recommended to deal with the generalized regression model, giving root-n consistent estimators of $\boldsymbol{\beta}^*$. Applications of these estimators to high dimensional data, however, are questionable. This article studies, both theoretically and practically, a simple yet powerful smoothing approach to handle the high dimensional generalized regression model. Theoretically, a family of smoothing functions is provided, and the amount of smoothing necessary for efficient inference is carefully calculated. Practically, our study is motivated by an important and challenging scientific problem: decoding gene regulation by predicting transcription factors that bind to cis-regulatory elements. Applying our proposed method to this problem shows substantial improvement over the state-of-the-art alternative in real data.

MSC 2010 subject classifications: Primary 47N30.
Keywords and phrases: Semiparametric regression, generalized regression model, rank-based M-estimator, smoothing approximation, transcription factor binding kwdgenomics.

Received December 2016.

Contents

1. Introduction

Regression models play a fundamental role in characterizing the relation among variables. Nonparametric and semiparametric regression models are commonly used alternatives to linear regression when the latter fails to fit the data well. Their advantages over simple linear regression models have been established in various fields [1, 2].

In this article, we study a semiparametric generalization of linear regression as such an alternative. We assume

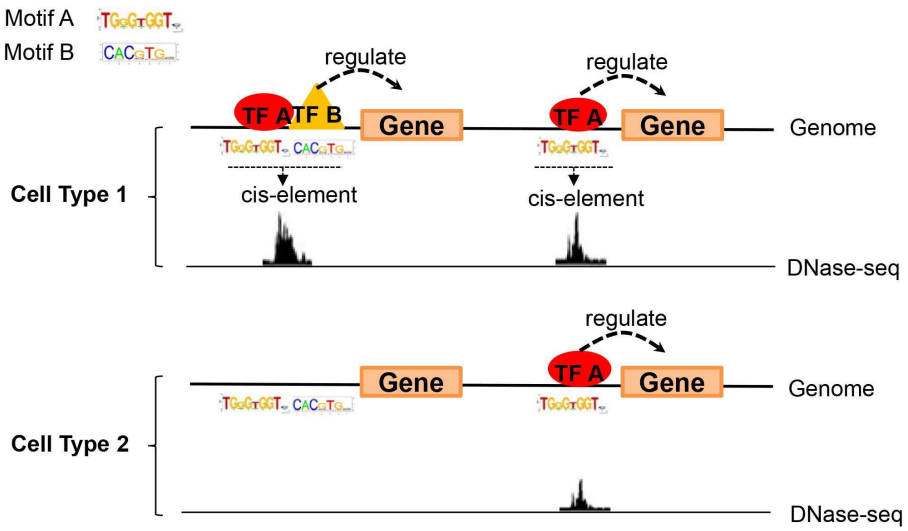
$$Y = D \circ \Lambda(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \epsilon), \quad \text{where} \quad Y, \epsilon \in \mathbb{R} \text{ and } \boldsymbol{X}, \boldsymbol{\beta}^* \in \mathbb{R}^d. \tag{1.1}$$

Here Y is the scalar response variable, $D(\cdot)$ is an unknown increasing function, $\Lambda(\cdot, \cdot)$ is an unknown strictly increasing function regarding each of its arguments, $\boldsymbol{X}$ represents the vector of explanatory variables, ϵ is an unspecified noise term independent of $\boldsymbol{X}$, and $\boldsymbol{\beta}^*$ is the regression coefficient characterizing the relation between $\boldsymbol{X}$ and Y. The coefficient $\boldsymbol{\beta}^*$ is assumed to be sparse. Model (1.1) is referred to as the generalized regression model. It was first proposed in econometrics [3], and is a very flexible semiparametric model, containing a parametric part, encoded in the linear term $\boldsymbol{X}^\top \boldsymbol{\beta}^*$, and a nonparametric part, encoded in link functions $D(\cdot), \Lambda(\cdot, \cdot)$, and the noise term ϵ. In practice, we assume that n independent realizations of $(Y, \boldsymbol{X})$, denoted as $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$, are observed. These observations will be used to fit the model.

1.1. A motivating genomic challenge

Model (1.1) naturally occurs in many applications. Below we elaborate a challenging scientific problem that motivates our study. To help put the model into context, some background introduction is necessary. One fundamental question in biology is how genes' transcriptional activities are controlled by transcription factors (TFs). TFs are an important class of proteins that can bind to DNA to induce or repress the transcription of genes nearby the binding sites (Figure 1(a)). Human has hundreds of different TFs. Each TF can bind to 10^2-10^5 different genomic loci to control the expression of 10-10^3 target genes. The genomic sites bound by TFs are also called cis-regulatory elements or cis-elements. Promoters and enhancers are typical examples of cis-elements. TF binding activities are context-dependent. The binding activity of a given TF at a given cis-element varies from cell type to cell type. It depends on the TF's expression level in each cell type as well as numerous other factors. One cis-element can be bound by multiple collaborating TFs that form a protein complex. Different TFs can bind to different cis-elements. In order to comprehensively understand the gene regulatory program, a crucial step is to identify all active cis-elements and their binding TFs in each cell type and biological condition.

The state-of-the-art technology for mapping genome-wide TF binding sites (TFBSs) is Chromatin Immunoprecipitation coupled with sequencing (ChIP-seq) [4]. Unfortunately, each ChIP-seq experiment can only analyze one TF. Using this technology to map binding sites of all human TFs would require one



(a) Illustration of TF, motif, cis-element and DNase-seq

	Y: DNase I hypersensitivity at a cis-element	X_1: TF 1 expression	X_2: TF 2 expression	...	X_d: TF d expression
Cell type 1	1.2	2.5	0.2	...	2.3
Cell type 2	2.3	0.9	3.5	...	2.4
...	...	...	...	...	...
Cell type n	0.1	0.7	0.1	...	0.9

(b) Overview of data structure

FIG 1. ***Background.*** *(a) An illustration of TF, motif, cis-element, and DNase I hypersensitivity measured by DNase-seq. For each cell type, two panels are displayed. In the top panel, the horizontal line represents the genome. TFs (ellipse and triangle) bind to cis-elements to activate or repress the transcription of nearby genes (rectangle). Each TF binds to a specific DNA motif located within the cis-element. In the bottom panel, DNase I hypersensitivity measured by DNase-seq at each cis-element correlates with the TF binding activity. Since TF binding activities are different in different cell types, the DNase-seq signals also vary across cell type. (b) Data structure used for predicting a cis-element's binding TFs. For each cis-element, Y is the DNase-seq signal measured in different cell types, and $\boldsymbol{X}$ is gene expression levels for d TFs in the same cell types.*

to conduct hundreds of such experiments. Moreover, ChIP-seq requires high-quality antibodies which are not available for all TFs. Therefore, mapping binding sites for all TFs using ChIP-seq is both costly and technically infeasible. An alternative approach to mapping TFBSs is based on genome-wide sequencing of DNase I hypersensitive sites (DNase-seq) [5]. TFBSs are often sensitive to the cleavage of DNase I enzyme. Thus, DNase I hypersensitivity (DH), which can be measured in a genome-wide fashion using DNase-seq, can be used to locate

cis-elements actively bound by TFs (Figure 1(a)). This approach is capable of mapping binding sites of all TFs in a biological sample through a single experimental assay. However, a major limitation of DNase-seq is that it does not reveal the identities of TFs that bind to each cis-element. If one could solve this problem by correctly predicting which TFs bind to each element, one would then be able to combine DNase-seq with computational predictions to identify all active cis-elements and their binding TFs in a biological sample using a single experimental assay. This would help scientists to remove a major roadblock in the study of gene regulation.

Many TFs recognize specific DNA sequence patterns called motifs (Figure 1(a)). Different TFs recognize different motifs. A conventional way to infer the identities of TFs that bind to a cis-element is to examine which TFs' DNA motifs are present in the cis-element. Unfortunately, motifs for 2/3 of all human TFs are unknown. Therefore, solely relying on DNA motifs is not sufficient to solve this problem. This motivates development of an alternative solution that leverages massive amounts of gene expression and DNase-seq data in public databases to circumvent the requirement for DNA motifs. The Encyclopedia of DNA Elements (ENCODE) project [6] has generated DNase-seq and gene expression data for a variety of different cell types. Using these data, one may examine how the protein-binding activity measured by DNase-seq at a cis-element varies across different cell types and how such variation is explained by variations in the expression of TFs (Figure 1(b)). Through this analysis, one may infer which TFs bind to each cis-element.

Formally, let Y be the activity of a cis-element in a particular cell type measured by DNase-seq, and let $\boldsymbol{X} = (X_1, \ldots, X_d)^\mathsf{T}$ be the expression level of d different TFs in the same cell type (Figure 1(b)). The relationship between cis-element activity and TF expression can be described using a generalized regression model, $Y = D \circ \Lambda(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}^*, \epsilon)$, with a high dimensional sparse regression coefficient vector $\boldsymbol{\beta}^*$. One expects the relationship between Y and $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}^*$ to be monotonic since a TF has to be expressed in order to be able to bind to cis-elements. Also, increased TF expression may lead to increased binding. The relationship may not be linear and the noise may not be normal since DNase-seq generates counts data. Although after normalization, the data may no longer be integers, they usually are still non-normal and may be zero-inflated. The model is high dimensional since there are hundreds of of TFs (i.e., $d = 10^2 - 10^3$), whereas the sample size (i.e., the number of ENCODE cell types with both DNase-seq and gene expression data) is small (n=50-100). Lastly, $\boldsymbol{\beta}^*$ has to be sparse since the number of TFs that can bind to a cis-element is expected to be small. This is because cis-elements are typically short. Each element cannot have physical contacts with too many different proteins. In this model, the non-zero components of $\boldsymbol{\beta}^*$ may be used to infer the identities of binding TFs.

1.2. Existing works on generalized regression

In order to properly position our results in the literature, below we briefly review existing methodological works that are most relevant to our study on deciphering

the generalized regression model. The generalized regression model contains many widely-used econometrical and statistical models, including important sub-classes such as the monotonic transformation model and monotonic single-index model, of the following forms:

$$Y = G(\boldsymbol{X}^\top \boldsymbol{\beta}^* + \epsilon) \quad \text{(monotonic transformation model)}, \tag{1.2}$$

$$Y = G(\boldsymbol{X}^\top \boldsymbol{\beta}^*) + \epsilon \quad \text{(monotonic single-index model)}, \tag{1.3}$$

where the univariate link function $G(\cdot)$ is assumed to be strictly increasing.

There has been research in estimating the generalized regression model or its variants in low dimensions. These works follow two tracks. In the first track, [3] and [7] proposed rank-based M-estimators for directly estimating $\boldsymbol{\beta}^*$, while treating link functions $D(\cdot)$ and $\Lambda(\cdot, \cdot)$ as nuisance parameters. The corresponding estimator $\widehat{\boldsymbol{\beta}}$ aims to maximize certain rank correlation measurement between Y and $\boldsymbol{X}^\top \boldsymbol{\beta}$, and hence often involves a discontinuous loss function. Based on an estimate of $\boldsymbol{\beta}^*$, [8], [9], and [10] further proposed methods to estimate the link function $D \circ \Lambda(\cdot, \cdot)$ under different parametric assumptions on link functions. This method is also extended in [11] and [12] to ultra-high dimensional settings via coupling it with a lasso-type penalty. In the second track, [13], [14], [15], among many others, focused on studying more specific models in (1.2) and (1.3), and suggested to approximate $D \circ \Lambda(\cdot, \cdot)$ for estimating $\boldsymbol{\beta}^*$ via exploiting the kernel regression and sieve approximation. These approaches therefore naturally require geometric assumptions and smoothness conditions on $D \circ \Lambda(\cdot, \cdot)$.

In high dimensions when d could be much larger than the sample size n, serious drawbacks are associated with methods in both tracks.

For the second track, first, simultaneous estimation of $D \circ \Lambda(\cdot, \cdot)$ and $\boldsymbol{\beta}^*$ requires extra prior assumptions on $D \circ \Lambda(\cdot, \cdot)$, which may not hold in practice. Secondly, nonparametric estimation is well known to be difficult in high dimensions. This could hurt the estimation of $\boldsymbol{\beta}^*$. Thirdly, these estimation procedures usually are very sensitive to outliers, which are common in real applications.

For the first track, rank-based M-estimators treat $D \circ \Lambda(\cdot, \cdot)$ as nuisance and hence could potentially gain efficiency and modeling flexibility in estimating $\boldsymbol{\beta}^*$. However, these rank-based methods also have serious computational and theoretical drawbacks. Computationally, the loss functions of rank-based M-estimators are commonly discontinuous. This violates the regularity conditions in most optimization algorithms [16, 17] and makes the optimization problem intractable. It could create serious computational issues, especially in high dimensions. Theoretically, the discontinuity of loss functions adds substantial difficulty for analysis, especially in high dimensions. This further makes the statistical performance of corresponding estimators intractable.

1.3. Overview of key results

This article studies a simple smooth alternative to the above rank-based methods. This results in estimators that are computationally efficient to calculate,

while keeping the modeling flexibility in estimating $\boldsymbol{\beta}^*$. The core idea is to replace the non-smooth rank-based loss function $\widehat{\mathcal{L}}(\cdot)$ by a smooth loss function $\widehat{\mathcal{L}}_n(\cdot)$, indexed by n. $\widehat{\mathcal{L}}_n(\cdot)$ is designed to become closer to $\widehat{\mathcal{L}}(\cdot)$ when n increases. A family of smoothing functions is accordingly studied.

Of note, the idea to approximate a non-smooth loss function using a smooth one has proven its successfulness in literature: see, for example, [18], [19], [20], [21], and [12] for smoothing Manski's maximum score estimator, an estimator targeting at the area under the receiver operating characteristic curve (AUC), quantile regression estimator, and Han's rank-based M-estimator in low and high dimensions. However, our results add new observations to this track of works both theoretically and in some new biological applications, which will be outlined below.

Theoretically, although a fairly studied approach in low dimensions, smoothing approximation to non-smooth and non-convex loss functions has received little attention in high dimensions. This is mainly due to the extreme irregularity of original loss functions, which form discontinuous U-processes [22]. Theoretically speaking, smoothing renders two types of errors: (i) the bias which characterizes the error introduced by employing $\widehat{\mathcal{L}}_n(\cdot)$ to approximate $\widehat{\mathcal{L}}(\cdot)$; (ii) the variance which characterizes the error introduced by the randomness of $\widehat{\mathcal{L}}_n(\cdot)$. Our study characterizes behaviors of both types of errors, based on which one can calculate the amount of smoothing necessary to balance the bias and variance. Our theory holds without any assumption on $D \circ \Lambda(\cdot, \cdot)$ other than monotonically increasing. Additionally, the noise term ϵ is allowed to be non-centered, arbitrarily heavy-tailed, including these Cauchy distributed ones with median possibly non-zero, and contain a substantial amount of outliers. Our theory hence confirms, mathematically rigorously, several advantages of smoothed Han's maximum rank estimator.

Practically, the aforementioned advantages of the studied procedure are important for problems where a large number of different regression models need to be fitted. Consider our motivating problem of predicting binding TFs of cis-elements. Since different cis-elements behave differently, the relationship between Y and $\boldsymbol{X}$ may have different functional forms for different cis-elements even though all these functions are monotonic. Additionally, different cis-elements may have different noise distributions. Despite this heterogeneity, different cis-elements can be conveniently handled in a unified fashion using our generalized regression procedure regardless of the form of $D \circ \Lambda(\cdot, \cdot)$ and ϵ. By contrast, manually constructing parametric models of different forms for differet cis-elements would be difficult due to the large number of models that need to be constructed. The advantages of our approach over existing high dimensional linear and robust regression couterparts (e.g., those in [23] and [24]) are hence clear.

Applying our method to the binding TF prediction problem, we find that our approach is substantially more accurate than the competing lasso method for predicting binding TFs. This demonstrates the practical utility of the smoothing approach and shows that it can provide a solution to a long-standing open problem in biology.

1.4. Other related works

Monotonic single-index and transformation models are important subclasses of the generalized regression model. In contrast to the monotonic transformation model, there exists an increasing amount of research studying the monotonic single-index model. These include [25], [15], and [26], to just name a few. However, they require much stronger modeling assumptions, and are sensitive to different types of data contamination.

There are two more related semiparametric approaches to generalize the linear regression model. The first is the general single-index model, with no explicit geometric constraint on $G(\cdot)$ in (1.3). In high dimensions, [27] and [28] studied such a relaxed model. For this, besides being sensitive to data contamination, they need to first sphericalize the data, which is extremely difficult in high dimensions. Recently, [29] proposed an alternative approach that does not require data sphericity. However, boundedness assumption on $\boldsymbol{X}$, subgaussianity of ϵ, and certain smoothness conditions on $G(\cdot)$ are still required.

The second is sufficient dimension reduction. Related literature in this direction includes [30], [31], [32], [33], and [34]. Sufficient dimension reduction approaches only assume Y is independent of $\boldsymbol{X}$ conditional on some linear projections of $\boldsymbol{X}$. However, a data sphericalization step is also crucial in all these approaches, and in each step of derivation, we need $d/n \to 0$ to proceed. These make the sufficient dimension reduction approaches vulnerable to high dimensionality, which will be further illustrated in simulations and real data experiments.

1.5. Paper organization

The rest of the paper is organized as follows. The next section presents our smoothing approach to estimating the generalized regression model. In Section 3, we use this method to solve our motivating scientific problem of decoding transcription factor binding. We demonstrate that this semiparametric regression approach is capable of substantially improving the accuracy over the state-of-the-art alternatives in real data. Section 4 gives theoretical results for understanding the proposed approach and calculates the appropriate smoothing amount. Section 5 provides discussions. Finite-sample simulation results and proofs are relegated to an appendix.

1.6. Notation

Let $\boldsymbol{v} = (v_1, \ldots, v_d)^{\mathsf{T}}$ and $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{d\times d}$ be a d dimensional real vector and a d by d real matrix. For sets $I, J \subset \{1, \ldots, d\}$, let $\boldsymbol{v}_I$ be the subvector of $\boldsymbol{v}$ with entries indexed by I, and $\mathbf{M}_{I,J}$ be the submatrix of $\mathbf{M}$ with rows and columns indexed by I and J. Let $\mathrm{card}(I)$ represent the cardinality of the set I. For $0 < q < \infty$, we define the vector ℓ_0, ℓ_q, and ℓ_∞ (pseudo-)norms of $\boldsymbol{v}$ to be $\|\boldsymbol{v}\|_0 := \mathrm{card}(\{j : v_j \neq 0\})$, $\|\boldsymbol{v}\|_q := (\sum_{i=1}^d |v_i|^q)^{1/q}$, and $\|\boldsymbol{v}\|_\infty := \max_{1\le i\le d} |v_i|$.

For the symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ represent the largest and smallest eigenvalues of $\mathbf{M}$. We write $\mathbf{M} \succeq 0$ if $\mathbf{M}$ is positive semi-definite. For any $x \in \mathbb{R}$, we define the sign function $\mathrm{sign}(x) := x/|x|$, where by convention we write $0/0 = 0$. For any two random vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$, we write $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed. We let c, C be two generic absolute positive constants, whose actual values may vary at different locations. We write $\mathbb{1}(\cdot)$ to be the indicator function. We let $\mathbb{S}^{d-1}$ represent the ball $\{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$. For any two real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$, or equivalently $b_n \gtrsim a_n$, if there exists a constant C such that $|a_n| \leq C|b_n|$ for any large enough n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The symbols $\overset{\mathbb{P}}{\lesssim}, \overset{\mathbb{P}}{\gtrsim}$, and $\overset{\mathbb{P}}{\asymp}$ are analogous to $\lesssim, \gtrsim$, and $\asymp$, but these relations hold stochastically.

2. Problem setup and methods

In this section, we provide some background on the generalized regression model and associated rank-based M-estimators. We further describe the class of non-convex smoothness approximations that will be exploited and covered by our theory.

Throughout the paper, we assume Model (1.1) holds, and we have n independent and identically distributed (i.i.d.) observations $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ generated from (1.1). We do not pose any parametric assumption on either the link functions $D(\cdot), \Lambda(\cdot, \cdot)$ or the noise term ϵ, except for assuming ,

$$D(\cdot) \text{ is non-degenrate}, \quad D(a) \geq D(b), \quad \Lambda(a, \cdot) > \Lambda(b, \cdot), \quad \text{and} \quad \Lambda(\cdot, a) > \Lambda(\cdot, b), \tag{2.1}$$

as long as $a > b$. For model identifiability, in the sequel, we assume we know at least one specific entry of $\boldsymbol{\beta}^*$ that is nonzero, and fix it to be one. Without loss of generality, we assume $\beta_1^* = 1$.

The generalized regression model of form (1.1) represents a large class of models. These include the monotonic transformation and single-index models of the forms (1.2) and (1.3). More specifically, the generalized regression model covers the linear regression model, with $D \circ \Lambda(u, v) = u + v$; the Box-Cox transformation model [35], with $D \circ \Lambda(u, v) = (u + v)^\lambda$ for some $\lambda > 0$; the log-linear model and accelerated failure time model [36], with $D \circ \Lambda(u, v) = \exp(u + v)$; the binary choice model [37], with $D \circ \Lambda(u, v) = \mathbb{1}(u + v \geq 0)$; the censored regression model [38], with $D \circ \Lambda(u, v) = (u + v)\mathbb{1}(u + v \geq 0)$.

We focus on the following rank-based M-estimator, called the maximum rank correlation (MRC), which is first proposed in [3]:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_1 = 1}{\mathrm{argmax}} \left\{ \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sign}(Y_i - Y_{i'}) \mathrm{sign}(\boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta} - \boldsymbol{X}_{i'}^\mathsf{T} \boldsymbol{\beta}) \right\}. \tag{2.2}$$

The formulation of MRC is a U-process [39, 21].

Intrinsically, MRC aims to maximize the Kendall's tau correlation coefficient [40] between Y and $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$, while treating the link functions as nuisance. This

is attainable only via assuming $D \circ \Lambda(\cdot, \cdot)$ is monotonic, and has its roots in transformation and copula models [41]. For such models, rank-based approaches have been well-known to be of certain efficiency properties [42, 43].

For fully appreciating the rationality of MRC, we provide a proposition that characterizes MRC's relation to the linear regression model, and further addresses the identifiability issue. Although the result in the first part is very straightforward, we do not find an explicit one in the literature that shows this relation,

Proposition 2.1. Suppose $\boldsymbol{X}$ is continuous and has a positive definite covariance matrix. We have, (i) when the link function $D \circ \Lambda(u, v) = u + v$ corresponds to the linear regression model (without requiring $\boldsymbol{X}$ and ϵ to be centered), $\boldsymbol{\beta}^*$ is the unique optimum to maximize the Pearson correlation between Y and $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$ up to a scaling:

$$\operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^d} \Big\{ \frac{\mathbb{E}[(Y_1 - Y_2)(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})]}{\sqrt{\mathrm{Cov}(Y_1 - Y_2)}\sqrt{\mathrm{Cov}(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})}} \Big\};$$

(ii) As long as the link functions $D(\cdot)$ and $\Lambda(\cdot, \cdot)$ satisfy (2.1), $\boldsymbol{\beta}^*$ is the unique optimum to maximize the Kendall's tau correlation coefficient between Y and $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$ up to a scaling:

$$\operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^d} \Big\{ \mathbb{E}\Big(\mathrm{sign}(Y_1 - Y_2) \mathrm{sign}(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta}) \Big) \Big\}.$$

Throughout, we are interested in the settings where the dimension d could be much larger than the sample size n. In such settings, due to restrictive information obtainable, we have to further regularize the parameter space. In particular, sparsity on the parametric space is commonly assumed. A seemingly natural regularized MRC estimator is as follows:

$$\operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_1 = 1} \Big\{ \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \mathrm{sign}(Y_i - Y_{i'}) \mathrm{sign}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_{i'}^\mathsf{T}\boldsymbol{\beta}) - \lambda_n \sum_{j=2}^{d} |\boldsymbol{\beta}_j| \Big\}, \tag{2.3}$$

or its equivalent formulation:

$$\operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_1 = 1} \Big\{ \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \Big(\mathbb{1}(Y_i > Y_{i'}) \mathbb{1}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} > \boldsymbol{X}_{i'}^\mathsf{T}\boldsymbol{\beta}) + \mathbb{1}(Y_i < Y_{i'}) \mathbb{1}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} < \boldsymbol{X}_{i'}^\mathsf{T}\boldsymbol{\beta}) \Big) - \lambda_n \sum_{j=2}^{d} |\boldsymbol{\beta}_j| \Big\}.$$

However, (2.3) involves a discontinuous loss function that has abrupt changes. As stated in the introduction, this incurs serious computational and statistical issues. For tackling such issues, we propose the following smoothing approximation to (2.3), using a set of cumulative distribution functions (CDFs)

$\{F_{ii'}(\cdot), 1 \le i < i' \le n\}$ to approximate the indicator function $\mathbb{1}(\cdot)$:

$$\widehat{\boldsymbol{\beta}}_{\alpha_n} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_1=1}{\text{local-argmax}} \Big\{ \frac{2}{n(n-1)} \sum_{1\le i<i'\le n} \Big(S_{ii'} F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) + (1-S_{ii'})(1-F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))) \Big) - \lambda_n \sum_{j=2}^{d} |\boldsymbol{\beta}_j| \Big\}.$$

Here "local-argmax$\{\cdot\}$" and "local-argmin$\{\cdot\}$" represent the sets of local maxima and minima for a given function respectively. In addition, we write

$$Z_{ii'}(\boldsymbol{\beta}) := (\boldsymbol{X}_i - \boldsymbol{X}_{i'})^{\mathsf{T}}\boldsymbol{\beta} \quad \text{and} \quad S_{ii'} := \mathbb{1}(Y_i > Y_{i'}).$$

Of note, α_n is an explicitly stated smoothness parameter controlling the approximation speed, presumably increasing to infinity with n. For any pair (i, i'), $F_{ii'}(\cdot)$ is a pre-determined fixed smooth continuous CDF, satisfying $F_{ii'}(-u) = 1 - F_{ii'}(u)$ for arbitrary $u \ge 0$.

Note we can equivalently write

$$\widehat{\boldsymbol{\beta}}_{\alpha_n} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_1=1}{\text{local-argmin}} \Big\{ \underbrace{-\frac{2}{n(n-1)} \sum_{1\le i<i'\le n} F_{ii'}(\widetilde{S}_{ii'}\alpha_n Z_{ii'}(\boldsymbol{\beta}))}_{\widehat{\mathcal{L}}_n(\boldsymbol{\beta})} + \lambda_n \sum_{j=2}^{d} |\boldsymbol{\beta}_j| \Big\}, \tag{2.4}$$

where $\widetilde{S}_{ii'} := \text{sign}(Y_i - Y_{i'})$ is the signed pairwise difference.

On one hand, the smoothed loss function $\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ is close to the MRC loss function in (2.3) when α_n is large enough. This guarantees "the bias term" is small enough. On the other hand, $\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ is smooth, giving computational and statistical guarantees for convergence in optimizing the loss function (2.4).

There are several notable remarks for the proposed smoothing approach.

Remark 2.2. In practice, we can take the approximation function $F_{ii'}(u)$ of the following forms:

- sigmoid function: $F_s(u) := 1/(1+\exp(-u))$;
- standard Gaussian CDF: $F_g(u) := \Phi(u)$, where $\Phi(\cdot)$ represents the standard Gaussian CDF;
- standard double exponential CDF: $F_e(u) := 1/2 + \text{sign}(u)(1-\exp(-|u|))/2$.

As will be shown later, the above approximation functions all guarantee efficient inference. For the approximation parameter α_n, theoretically, we recommend choosing it using a specific rate. Such a rate depends on n, d, and a sparseness parameter, and will be stated more explicitly in Section 4. In practice, cross-validation is recommended [44].

Remark 2.3. We note the formulation (2.4) is related to those smoothing approaches introduced in [45], [21], and [12]. Actually, their approaches could

be regarded as special cases of (2.4), by taking $F_{ii'}(\cdot)$ to be the Gaussian CDF or sigmoid function. However, as will be seen in Sections 3 and 4, we will add new contributions to literature in both theory and applications.

Remark 2.4. It is also worth comparing the formulation in (2.4) to the other robust regression formulations introduced in the high dimensional statistics literature. The original lasso estimator is well-known to be vulnerable to non-linear link functions [46], and heavy-tailed noise term ϵ [47]. [48], [24], and [23] proposed different robust (non)convex approaches to address the possible heavy-tailedness issue of ϵ. In particular, [23] provided a framework for investigating a group of (non)convex loss functions (e.g., Huber's loss and Tukey's biweight), and studied the corresponding estimators. However, these procedures all stick to the linear link function, and hence will lead to inconsistent estimation, invalid statistical inference, and erroneous predictions when the link function is non-linear. As is discussed in the introduction, non-linearity is common in complex biology systems.

3. Real data example

In this section, we apply our approach to the motivating scientific problem – predicting TFs that bind to individual cis-elements. As introduced before, solving this problem is crucial for studying gene regulation. DNase-seq experiments can be used to map active cis-elements in a genome-wide fashion. If one can correctly predict which TFs bind to each cis-element, one would be able to couple DNase-seq with computational predictions to efficiently predict genome-wide binding sites of a large number of TFs simultaneously in a new biological sample. This cannot be achieved using any other existing experimental technology.

The conventional approach that predicts binding TFs based on DNA motifs is contingent on the availability of known TF motifs. However, 2/3 of human TFs do not have known motifs. This motivates us to investigate an alternative solution that does not require DNA motif information. This new approach leverages large amounts of publicly available DNase-seq and gene expression data generated by the ENCODE project. Using data from multiple ENCODE cell types, this approach models the relationship between a cis-element's protein-binding activity Y measured by DNase-seq and the gene expression levels of d TFs, $\mathbf{X}$, measured by exon arrays. It predicts binding TFs by identifying important variables in $\mathbf{X}$ in the regression model. Below we present real data results from a small-scale pilot study as a way to illustrate our semiparametric regression approach and compare it with other methods. A comprehensive whole-genome analysis and investigation of biology are beyond the scope of this article and will be addressed elsewhere.

In our pilot study, human DNase-seq and matching gene expression exon array data from 57 cell types were obtained from the ENCODE project. After data processing and normalization (see the appendix Section C.1 for details), 169 TFs whose gene expression varies across the 57 cell types were obtained. In parallel, 1,000 cis-elements were randomly sampled from a total of over 10^6

cis-elements in the human genome for method evaluation. For each cis-element, the objective is to identify which of the 169 TFs may bind to it. Let Y be a cis-element's DNase I hypersensitivity (a surrogate for protein-binding activity), measured by its normalized and log-transformed DNase-seq read count, in a cell type. Let $\boldsymbol{X}$ be the normalized gene expression values of the 169 TFs in the same cell type. We want to use Y and $\boldsymbol{X}$ observed from 57 different cell types to learn their relationship and subsequently predict binding TFs.

Six competing methods are compared, listed below. For all methods except random prediction, the non-zero components of the estimate were used to predict which TFs can bind to a cis-element. The code that implements our method has been released online (https://github.com/zji90/RMRCE).

- RMRCE: the generalized regression model $Y = D \circ \Lambda(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}^*, \epsilon)$ was fitted using Regularized Maximum Rank Correlation Estimator (RMRCE) in (2.4). The tuning parameter α_n was selected using cross validation and the Gaussian smoothing approximation was used.
- Hinge: the indicator function in Han's proposal is replaced by a hinge loss approximation. Specifically, the loss function is changed to:
$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d,\beta_1=1}{\text{local-argmax}}\Big\{\frac{1}{n(n-1)}\sum_{i\neq i'}\mathbb{1}(Y_i > Y_{i'})[\max\{0,(\boldsymbol{X}_i-\boldsymbol{X}_{i'})^{\mathsf{T}}\boldsymbol{\beta}+1\}]-\lambda_n\sum_{j=2}^{d}|\boldsymbol{\beta}_j|\Big\}.$$
- Lasso: the lasso [49] was used.
- SIM: the method as proposed by [29].
- SDR: the sufficient dimension reduction method as proposed by [34].
- Random: TFs randomly sampled from the 169 TFs were treated as the predicted binding TFs.

Among these methods, the lasso represents the state-of-the-art linear model for characterizing the relationship between a response and sparse predictors. SIM and SDR represent competing semiparametric regression models. Hinge is a simple convex relaxation of Han's proposal. We include this comparison to find out whether the smoothed rank correlation is better than convex relaxation. The random method serves as a negative control. For all methods except random prediction, we tried different tuning parameters (α_n in RMRCE was selected using cross validation) and calculated the overall accuracy under each parameter setting. Detailed implementation strategy for the methods is presented in Section A.1 in the appendix.

Prediction accuracy of different methods is compared using DNA motif information. The rationale is that DNA motifs were not used to make predictions, therefore they provide an independent source of information for validation. For a method with better prediction accuracy, one should expect that its predicted binding TFs for a cis-element are more likely to be supported by the presence of corresponding motifs in the cis-element. By contrast, it is less likely to find motifs in a cis-element for incorrectly predicted binding TFs. Based on this reasoning, we downloaded all known vertebrate motif matrices from the JASPAR database [50] and mapped them to human genome using the CisGenome software [51] under its default setting. For a given TF, if its motif(s) occurred one or

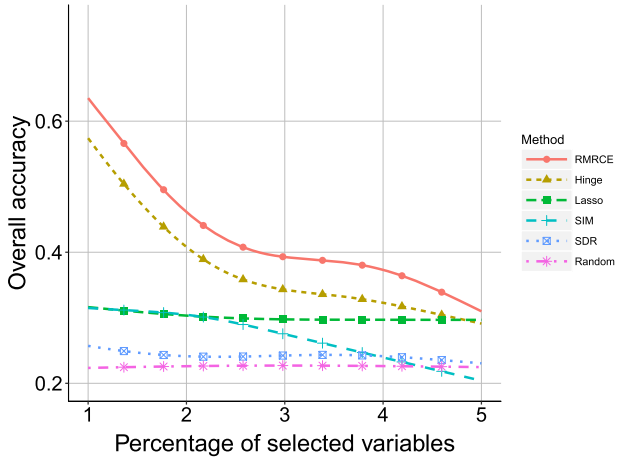


FIG 2. *Overall accuracy of RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, SDR, and the random prediction method. X-axis shows the averaged percentage of selected TFs out of all 169 TFs. Y-axis shows the overall accuracy.*

multiple times within 500 base pairs (bp's) of the center of a cis-element, then the TF's motif was called to be found in the cis-element. Let $i \in \{1, \ldots, 1,000\}$ be the index of cis-elements. Let M_i denote the set of TFs whose motifs were found in cis-element i. In order to characterize a method's prediction accuracy, we applied the method to predict binding TFs for each cis-element. If a predicted TF does not have any known DNA motif, we lack information for evaluating the correctness of the prediction. Therefore, for each cis-element, we only retained the predicted TFs that had known DNA motifs in the JASPAR database for estimating the prediction accuracy. Among all the 169 TFs, 63 TFs had known DNA motifs and were included in the evaluation. Let A_i be the set of retained TFs for cis-element i, and let $|A_i|$ be the number of TFs in A_i. Let $B_i = A_i \bigcap M_i$ be the subset of TFs in A_i whose motifs were found in cis-element i (and hence validated), and let $|B_i|$ be the number of TFs in B_i. The prediction accuracy of a method was characterized by the following ratio

$$\sum_{i=1}^{1,000} |B_i| \Big/ \sum_{i=1}^{1,000} |A_i|.$$

This ratio is the percentage of all testable predictions that were validated by the presence of DNA motifs. The higher the ratio, the more accurate a method is.

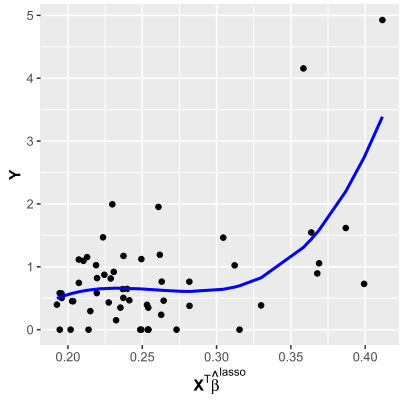
Figure 2 compares the accuracy of different methods. For each method, we gradually increased the number of reported TFs by relaxing the tuning parameter (e.g., setting a smaller tuning parameter λ_n will result in more TFs being reported by RMRCE), and the accuracy was plotted as a function of increasing number of predicted binding TFs. This figure shows that RMRCE is significantly

more accurate than all the other methods, and the random prediction method has the worst performance. Of note, as the number of selected TFs increases, the accuracy of all methods drops (except for the random, which remains stable). This is as expected since the overall signal strength decreases as more TFs are reported. RMRCE performance with different choices of α_n is presented in the appendix (Section C.2). A model diagnostic heuristic is developed to check the monotonicity assumption of the proposed model. The detailed descriptions and model diagnostic results in real data application can be found in the appendix (Section C.3).

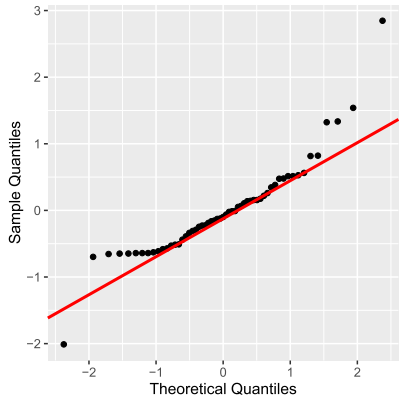
To shed light on why RMRCE substantially outperformed the lasso, Figure 3 shows data from two cis-elements as examples. For each cis-element, we used the lasso to identify binding TFs from the 169 TFs. The observed response $\boldsymbol{Y}$ was then plotted against its fitted value $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ in Figure 3(a) and Figure 3(c). The blue curve in each plot represents a smooth curve fitted using the generalized additive model with cubic splines and default parameters as implemented in the R package **mgcv**. It treats $\boldsymbol{Y}$ as response and $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ as independent variable. Clearly, $\boldsymbol{Y}$ and $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ do not have a linear relationship. Moreover, the figures show that the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ for different cis-elements have different functional forms. This makes the use of parametric models complicated as one would need to build models with different functional forms for different cis-elements, which would be tedious if one wants to analyze millions of cis-elements in the whole genome. Figure 3(b) and Figure 3(d) show the normal qqplots for the residuals that were obtained from the fitted smooth curves. These figures show that, even when a non-linear smoothed function was fitted to the data, the residuals are still non-normal and may have a complicated distribution.

Figure 4 shows a similar analysis using RMRCE. In Figure 4(a) and Figure 4(c), $\boldsymbol{Y}$ was plotted against the RMRCE fitted $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{RMRCE}}$. Clearly, the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{RMRCE}}$ is non-linear, but such a non-linear, yet monotonically increasing, relationship can be handled by our method in a unified fashion regardless of the specific functionnon-linearal forms. Figure 4(b) and Figure 4(d) are the residual qqplots where the residuals were obtained from the fitted smooth curves in Figure 4(a) and Figure 4(c). These qqplots show that the distributions of the residuals are non-normal. The non-normal residuals, however, can be naturally handled by our generalized regression model.

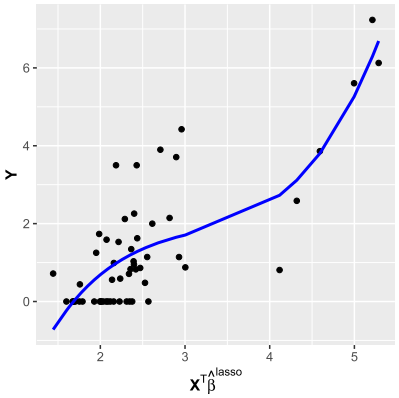
The above analyses demonstrate the value of our approach for handling noisy, monotonic, non-normal, and non-linear data. Whereas the simple linear regression and marginal screening cannot fully capture the delicateness of such a complex system, RMRCE handles these challenges very well. Thus, RMRCE is a more appealing method to tackle the studied problem than simple linear models based methods such as the lasso. Similar conclusion applies to the comparison with the other four competing methods. In particular, as will be shown in the Section of synthetic data analysis (Section A), Hinge is usually a convex approximation too crude to the studied method, and when the generalized regression model is reasonable in modeling the data (which is hinted by the experimental



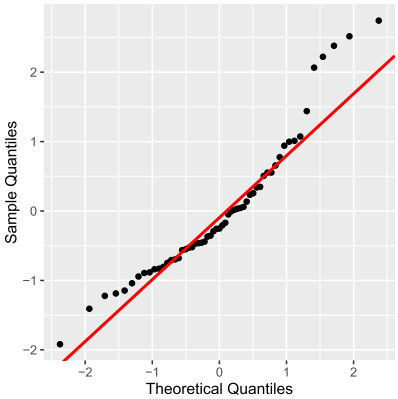
(a) Fitted curve for cis-element 1 using the lasso



(b) Residual qqplot for cis-element 1 using the lasso



(c) Fitted curve for cis-element 2 using the lasso



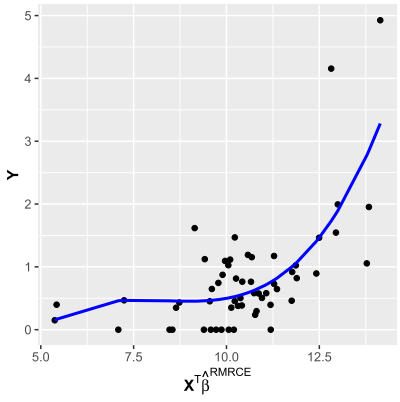
(d) Residual qqplot for cis-element 2 using the lasso

FIG 3. ***Fitted curves and residual qqplots using the lasso for two cis-elements.*** *For each cis-element, the lasso was fitted to obtain the regression coefficients* $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$. *Blue lines in (a) and (c) are fitted smoothed curves treating* $\boldsymbol{Y}$ *as response and* $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ *as independent variable. (b) and (d) are residuals qqplots for the smoothed curves. The quantiles of the residuals (sample quantiles) are compared to the quantiles of a normal distribution (theoretical quantiles). The residuals are calculated as* $\boldsymbol{Y} - g(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\text{lasso}})$ *where* g *represents the smoothed functions fitted in (a) and (c).*

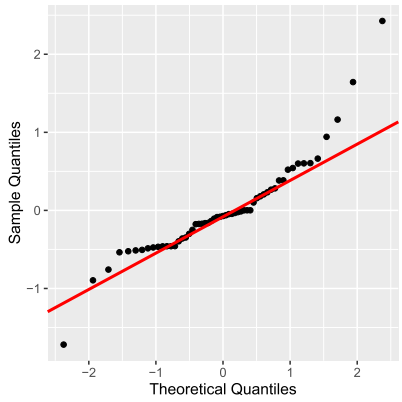
results in this section), SIM and SDR are much less efficient in handling the data than RMRCE.

4. Theory

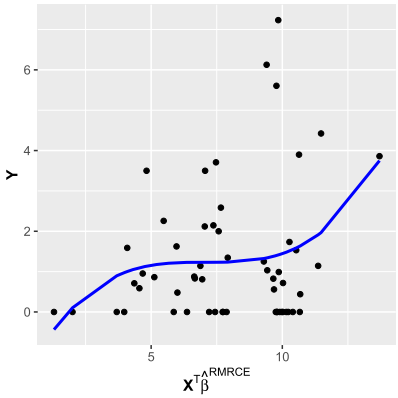
This section is devoted to investigating the statistical performance of the smoothing estimator $\widehat{\boldsymbol{\beta}}_{\alpha_n}$ in (2.4). For characterizing the estimation error of $\widehat{\boldsymbol{\beta}}_{\alpha_n}$ to $\boldsymbol{\beta}^*$,



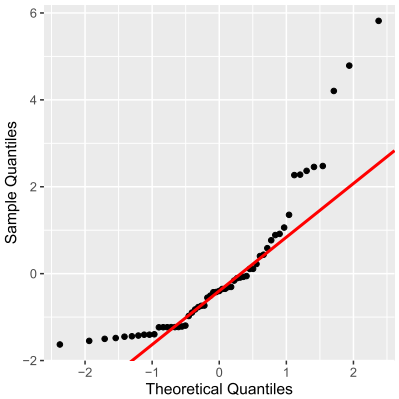
(a) Fitted curve for cis-element 1 using RMRCE



(b) Residual qqplot for cis-element 1 using RMRCE



(c) Fitted curve for cis-element 2 using RMRCE



(d) Residual qqplot for cis-element 2 using RMRCE

FIG 4. ***Fitted curves and residual qqplots using RMRCE for two cis-elements.*** *For each cis-element, RMRCE was fitted to obtain the regression coefficients* $\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$. *Blue lines in (a) and (c) are fitted smoothed curves treating* $\boldsymbol{Y}$ *as response and* $\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ *as independent variable. (b) and (d) are residuals qqplots for the smoothed curves. The quantiles of the residuals (sample quantiles) are compared to the quantiles of a normal distribution (theoretical quantiles). The residuals are calculated as* $\boldsymbol{Y} - g(\boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}})$ *where* g *represents the smoothed functions fitted in (a) and (c).*

we separately study the approximation error (bias part) and the stochastic error (variance part).

Before introducing main theorems, let's first introduce some additional notation. Let's first define the population approximation minimizer $\boldsymbol{\beta}^*_{\alpha_n}$ as

$$\boldsymbol{\beta}^*_{\alpha_n} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_1=1}{\text{local-argmin}} \mathcal{L}_n(\boldsymbol{\beta}),$$

where $\mathcal{L}_n(\boldsymbol{\beta}) := \mathbb{E}\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ can be easily derived as

$$\mathcal{L}_n(\boldsymbol{\beta}) = -\frac{2}{n(n-1)} \sum_{1\le i<i'\le n} \mathbb{E}\Big\{ S_{ii'} F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) + (1-S_{ii'})(1-F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})))\Big\}. \tag{4.1}$$

Note, analogously, Proposition 2.1 shows

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_1=1}{\operatorname{argmin}} \underbrace{-\mathbb{E}\big(S_{ii'}\mathbb{1}(Z_{ii'}(\boldsymbol{\beta})>0) + (1-S_{ii'})\mathbb{1}(Z_{ii'}(\boldsymbol{\beta})<0)\big)}_{\mathcal{L}(\boldsymbol{\beta})}. \tag{4.2}$$

We decompose the statistical error $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2$ into two parts:

$$\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2 \le \underbrace{\|\boldsymbol{\beta}^*_{\alpha_n} - \boldsymbol{\beta}^*\|_2}_{\text{approximation error}} + \underbrace{\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*_{\alpha_n}\|_2}_{\text{stochastic error}}.$$

We will first characterize the approximation error in the next section. Studies of the stochastic error are in Section 4.2.

Because the loss functions $\mathcal{L}(\boldsymbol{\beta})$, $\mathcal{L}_n(\boldsymbol{\beta})$, and $\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ are all non-convex, adopting a similar argument as in [23], we first focus on the following specified minima:

$$\widehat{\boldsymbol{\beta}}_{r,\alpha_n} := \underset{\beta_1=1, \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le r}{\operatorname{argmin}} \Big\{\widehat{\mathcal{L}}_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=2}^{d} |\beta_j|\Big\} \quad \text{and} \quad \boldsymbol{\beta}^*_{r,\alpha_n} := \underset{\beta_1=1, \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le r}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta}).$$

The estimators $\widehat{\boldsymbol{\beta}}_{r,\alpha_n}$ and $\boldsymbol{\beta}^*_{r,\alpha_n}$ are constructed merely for theoretical purposes. In addition, the parameter r there, controlling the convexity region around the truth $\boldsymbol{\beta}^*$, is no need to be specified in practice.

As will be shown in the next two subsections, $\widehat{\boldsymbol{\beta}}_{r,\alpha_n}$ and $\boldsymbol{\beta}^*_{r,\alpha_n}$ are local optima to $\widehat{\mathcal{L}}_n(\cdot)$ and $\mathcal{L}_n(\cdot)$ under some explicitly characterized regularity conditions. Thus, we will end this section with a general theorem for quantifying the behaviors of some local optimum $\widehat{\boldsymbol{\beta}}_{\alpha_n}$.

4.1. Approximation error

This section first shows $\mathcal{L}_n(\cdot)$ in (4.1) could well approximate $\mathcal{L}(\cdot)$ in (4.2). We will then study the behaviors of minima of $\mathcal{L}_n(\cdot)$ and $\mathcal{L}(\cdot)$.

4.1.1. Generalized regression model

For the generalized regression model of the form (1.1), to guarantee the fast approximation of $\mathcal{L}_n(\cdot)$ to $\mathcal{L}(\cdot)$, we require the following two assumptions on data generating schemes and approximation functions $\{F_{ii'}(\cdot), 1\le i<i'\le n\}$.

> **(A1).** Model (1.1) and the constraint (2.1) hold with $\boldsymbol{\beta}^*$ non-degenerate, $\{\boldsymbol{X}_i,\ i=1,\ldots,n\} \overset{i.i.d.}{\sim} N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and independent of i.i.d. absolutely continuous noises $\{\epsilon_i,\ i=1,\ldots,n\}$. Of note, the noise term ϵ_i is not necessarily of mean or median zero.

(A2). Assume $F_{ii'}(u)$ satisfies $F_{ii'}(-u) = 1 - F_{ii'}(u)$ for arbitrary $u \geq 0$. Further assume there exist some absolute constants $C_1, C_2 > 0$ such that $1 - F_{ii'}(u) \leq C_1 \exp(-C_2 u)$ for any $u > 0$.

On one hand, Assumption **(A1)** is a commonly adopted assumption in the high dimensional regression literature [52, 53]. Of note, Assumption **(A1)** is by no means necessary. By checking the proof of Lemma 4.1, we could readily relax it to a multivariate subgaussian assumption [54]. However, for presentation clearness, this paper is focused on the multivariate Gaussian case. On the other hand, Assumption **(A2)** is mild. It is easy to check sigmoid function, Gaussian, and double exponential CDFs discussed in Remark 2.2 all satisfy **(A2)**.

Under Assumptions **(A1)** and **(A2)**, the next lemma investigates the approximation error of $\mathcal{L}_n(\cdot)$ to $\mathcal{L}(\cdot)$.

Lemma 4.1. Assume Assumptions **(A1)** and **(A2)** hold. We then have

$$\sup_{\boldsymbol{\beta}:\beta_1=1} |\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \leq \frac{2C_1}{C_2\alpha_n} \sup_{\boldsymbol{\beta}:\beta_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}},$$

where absolute constants C_1 and C_2 are given in Assumption **(A2)**.

Lemma 4.1 shows $\mathcal{L}_n(\cdot)$ uniformly approximates $\mathcal{L}(\cdot)$ linearly with regard to $1/\alpha_n$. However, it is not enough to show convergence for $\boldsymbol{\beta}^*_{\alpha_n}$ to $\boldsymbol{\beta}^*$. For guaranteeing this, we require a "local strong convexity" of $\mathcal{L}(\cdot)$ around the truth $\boldsymbol{\beta}^*$. To this end, we write

$$\boldsymbol{\Gamma} := \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*) \in \mathbb{R}^{d\times d}$$

to be the second derivative of the population loss function $\mathcal{L}(\cdot)$ in (4.2). The next assumption regularizes the behavior of $\boldsymbol{\Gamma}$'s spectrum.

(A3). Assume $\lambda_{\min}(\boldsymbol{\Gamma})$ and $\lambda_{\max}(\boldsymbol{\Gamma})$ are respectively lower and upper bounded by two absolute positive constants.

Assumption **(A3)** is also posed inexplicitly in [22]. With Assumption **(A3)**, the next proposition, essentially coming from [22], shows the local strong convexity of $\mathcal{L}(\cdot)$.

Proposition 4.2 ([22])**.** Assume Assumptions **(A1)** and **(A3)** hold. We can then pick positive absolute constant set $\gamma := \{\gamma_1, \gamma_2\}$ with $\gamma_2/\gamma_1 - 1$ close to 0, such that for some small enough $r(\gamma) > 0$ only depending on γ, as long as $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r(\gamma)$, we have

$$\gamma_1 \lambda_{\min}(\boldsymbol{\Gamma})\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq \gamma_2 \lambda_{\max}(\boldsymbol{\Gamma})\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

Combining Lemma 4.1 and Proposition 4.2, the next lemma characterizes the approximation error of $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$ to the truth $\boldsymbol{\beta}^*$.

Lemma 4.3. Assume Assumptions **(A1)**, **(A2)**, and **(A3)** hold. We then have

$$\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2^2 \lesssim \alpha_n^{-1} \cdot \sup_{\boldsymbol{\beta}:\beta_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}}.$$

Lemma 4.3 verifies that, when we choose α_n to increase to infinity with n, the approximation error decays to zero at a rate $\alpha_n^{-1/2}$ while assuming the boundedness of the eigenvalues of $\mathbf{\Gamma}$. We are focused on such $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$ and the corresponding estimator $\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n}$.

4.1.2. Monotonic transformation model

This section aims to study how sharp the results in the last subsection are. We focus on the monotonic transformation model (1.2). We first show Lemma 4.1 cannot be improved too much, even under a much more restrictive monotonic transformation model.

Lemma 4.4. Under Model (1.2), assume Assumption **(A1)** and $F_{ii'}(u)$ satisfies $F_{ii'}(-u) = 1 - F_{ii'}(u)$ for arbitrary $u \geq 0$. For arbitrary vectors $\|\boldsymbol{\beta}\|_2 = \|\boldsymbol{\beta}^*\|_2 = M < \infty$, we then have the following three statements true.

(1). Supposing $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{\beta}^* = 0$, we have $\mathcal{L}_n(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta})$ for any $\alpha_n > 0$.
(2). With ϵ Gaussian distributed with bounded parameters, fixing α_n and supposing $\mathbf{\Sigma} = \mathbf{I}_d$, $|\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})|$ is an increasing function of $|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}^*|$ for all three examples of approximations given in Remark 2.2.
(3). If we further assume sigmoid approximation and ϵ Gaussian distributed with bounded parameters, we have

$$|\mathcal{L}_n(\boldsymbol{\beta}^*) - \mathcal{L}(\boldsymbol{\beta}^*)| \asymp \alpha_n^{-2}.$$

Furthermore, Combined with the second fact yields, for any $\|\boldsymbol{\beta}\|_2 = M$,

$$|\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \lesssim \alpha_n^{-2}.$$

Lemma 4.4 shows, even under a very ideal parametric model, the approximation error in Lemma 4.1 can only be improved slightly from linearly to quadratically decaying.

Secondly, we note Proposition 4.2 relies on an inexplicit assumption **(A3)**. To fully appreciate this proposition, we provide an alternative way to clearly reveal its connection to the data generating parameter $\mathbf{\Sigma}$. For this, we assume the monotonic transformation model (1.2) and one more assumption.

(A3'). Assume Model (1.2) holds with explanatory variables $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$ and noises $\{\epsilon_i, i = 1, \ldots, n\}$ satisfying Assumption **(A1)**. We further assume the noises are absolutely continuous, satisfying

$$\int_{-\infty}^{0} f_\epsilon(x) \exp(-x^2/(2b_n^2))dx = Cb_n(1 + o(1)) \quad \text{as} \quad \mathrm{b_n} \to 0.$$

Here $f_\epsilon(\cdot)$ represents the probability density function (PDF) of $\epsilon_2 - \epsilon_1$.

We note the noise assumption in **(A3')** is mild. It does not require any moment condition on the noises $\{\epsilon_i, i = 1, \ldots, n\}$. In particular, the next proposition shows both Gaussian and Cauchy distributions satisfy Assumption **(A3')**.

Proposition 4.5. Suppose noises $\{\epsilon_i,\ i = 1, \ldots, n\}$ are arbitrarily Gaussian or Cauchy distributed with bounded parameters. Then they satisfy Assumption **(A3')**.

With Assumption **(A3')**, we are now ready to prove the local strong convexity of $\mathcal{L}(\cdot)$.

Lemma 4.6. Assume Assumptions **(A1)** and **(A3')** hold. We can then pick positive absolute constant set $\gamma := \{\gamma_1, \gamma_2\}$ with $\gamma_2/\gamma_1 - 1$ close to 0, such that for some small enough $r(\gamma) > 0$ only depending on γ, as long as $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq r(\gamma)$, we have

$$\gamma_1 \cdot \Big(1 - \frac{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sqrt{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}}\Big) \leq \mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq \gamma_2 \cdot \Big(1 - \frac{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sqrt{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}}\Big).$$

As a simple consequence, for $\boldsymbol{\Sigma} = \mathbf{I}_d$ and $\|\boldsymbol{\beta}^*\|_2 = \|\boldsymbol{\beta}\|_2 = M$, we have

$$\frac{\gamma_1}{2M^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq \frac{\gamma_2}{2M^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2.$$

Remark 4.7. Compared to the result in Proposition 4.2, Lemma 4.6 gives explicit inequalities based on $\boldsymbol{\Sigma}$, and the proof techniques exploited are utterly different from these in [22] that are based on Taylor's expansion.

4.2. Stochastic error

This section investigates the stochastic error term $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*_{\alpha_n}\|_2$. This falls in the application regime of the high dimensional M-estimators theory [53] with some slight modifications due to the additional constraint on $\boldsymbol{\beta}$: $\beta_1 = 1$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r$.

4.2.1. A general framework for constrained M-estimators

In this section we consider studying general M-estimators. In detail, let's be focused on the following constrained M-estimator:

$$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\mathcal{A}\subset\mathbb{R}^d}{\operatorname{argmin}} \Big\{L_n(\boldsymbol{\theta}) + \lambda_n P(\boldsymbol{\theta})\Big\},$$

which aims to estimate the truth

$$\boldsymbol{\theta}^* := \underset{\boldsymbol{\theta}\in\mathcal{A}\subset\mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}L_n(\boldsymbol{\theta}).$$

Here $\mathcal{A} \subset \mathbb{R}^d$ is a subset of the d-dimensional real space, $L_n(\cdot)$ is the loss function, and $P(\cdot)$ is the penalty term. We pose the following five assumptions on $\mathcal{A}$, $L_n(\cdot)$, and $P(\cdot)$.

- **(B1).** Assume $\mathcal{A} - \boldsymbol{\theta}^* := \{\boldsymbol{v} \in \mathbb{R}^d : \boldsymbol{v} = \boldsymbol{\theta} - \boldsymbol{\theta}^* \text{ for some } \boldsymbol{\theta} \in \mathcal{A}\}$ is star-shaped.

- **(B2).** Assume $L_n(\cdot)$ is convex differentiable in $\mathcal{A}$, and $P(\cdot)$ is a semi-norm.
- **(B3).** (Decomposability). There exist subspaces $\mathcal{M} \subset \overline{\mathcal{M}} \subset \mathbb{R}^d$ such that

$$P(\boldsymbol{\theta}+\boldsymbol{\gamma}) = P(\boldsymbol{\theta}) + P(\boldsymbol{\gamma}) \ \text{ for all } \boldsymbol{\theta} \in \mathcal{M} \text{ and } \boldsymbol{\gamma} \in \overline{\mathcal{M}}^{\perp}.$$

- **(B4).** (Restricted strong convexity). Define the set

$$\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \boldsymbol{\theta}^*) := \{\boldsymbol{\Delta} \in \{\mathcal{A} - \boldsymbol{\theta}^*\} : P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}}) \le 3P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}}) + 4P(\boldsymbol{\theta}^*_{\mathcal{M}^{\perp}})\},$$

  where $\boldsymbol{\Delta}_{\mathcal{N}}$ represents the projection of $\boldsymbol{\Delta}$ to $\mathcal{N}$ for arbitrary subspace $\mathcal{N}$ of $\mathbb{R}^d$. We assume, for all $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \boldsymbol{\theta}^*)$, we have

$$L_n(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - L_n(\boldsymbol{\theta}^*) - \langle \nabla L_n(\boldsymbol{\theta}^*), \boldsymbol{\Delta} \rangle \ge \kappa_L \|\boldsymbol{\Delta}\|_2^2 - \delta_L \|\boldsymbol{\Delta}\|_2 - \tau_L^2(\boldsymbol{\theta}^*),$$

  where κ_L, δ_L, and $\tau_L^2(\boldsymbol{\theta}^*)$ are three constants.
- **(B5).** We assume

$$\Psi(\overline{\mathcal{M}}) := \sup_{\boldsymbol{v} \in \overline{\mathcal{M}} \setminus \mathbf{0}} \frac{P(\boldsymbol{v})}{\|\boldsymbol{v}\|_2} < \infty \ \text{ and } \ \lambda_n \ge 2P^*(\nabla L_n(\boldsymbol{\theta}^*)).$$

  Here $P^*(\cdot)$ is the dual norm of $P(\cdot)$.

Assumptions **(B1)-(B5)** are posed for the purpose of involving estimators like $\widehat{\beta}_{\alpha_n}$, and hence also (slightly) generalize the corresponding ones in [53]. Under the above assumptions, we have the following theorem hold, which is a straightforward extension to Theorem 1 in [53].

Theorem 4.8. Assume Assumptions **(B1)-(B5)** hold. We then have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \le (2\lambda_n \Psi(\overline{\mathcal{M}}) + \delta_L)^2/\kappa_L^2 + 2(\tau_L^2(\boldsymbol{\theta}^*) + 2\lambda_n P(\boldsymbol{\theta}^*_{\mathcal{M}^{\perp}}))/\kappa_L.$$

4.2.2. Stochastic error analysis

For analyzing the high dimensional stochastic error, sparsity is commonly encouraged for model identifiability and efficient inference. We adopt this idea. In particular, we assume the following regularity condition:

> **(A0).** The data $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ are generated in a triangular array setting as in [55]. We assume, for some pre-specified α_n, $\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_0 \le s_n$, while the parameter s_n changes with n. Note, in this setting, $\boldsymbol{\beta}^*$ is not necessarily sparse.

This formulation of sparseness can be regarded as a working assumption, and intrinsically comes from the sieve idea [56, 57]. Note a similar assumption is inexplicitly stated in [24].

For guaranteeing efficient inference, we still require three more assumptions, listed below.

(A4). Assume $F_{ii'}(\cdot)$ is twice-differentiable, and there exists an absolute constant $C_3 > 0$ such that $\sup_{u\in\mathbb{R}} |\frac{d}{du}F_{ii'}(u)| \leq C_3$ for arbitrary pair (i, i') with $1 \leq i \neq i' \leq n$.
(A5). We assume $\nabla^2\widehat{\mathcal{L}}_n(\boldsymbol{\beta}) \succeq 0$ for arbitrary $\boldsymbol{\beta} \in \mathbb{R}^d$ with $\beta_1 = 1$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_2 \leq C_4 \cdot r(\gamma)$ for some $C_4 > 1$.
(A6). There exists an absolute constant $C_5 > 0$ such that $C_5^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C_5$.

Here Assumption **(A4)** is easy to check. Actually, it is straightforward to verify sigmoid function, Gaussian, and double exponential CDFs as approximation functions introduced in Remark 2.2 all satisfy Assumption **(A4)**. On the other hand, for Assumption **(A6)**, we require a little bit more stringent assumption on $\lambda_{\min}(\boldsymbol{\Sigma})$ than what is required for the lasso. This is because of the additional effort on controlling the smoothness approximation error. Finally, noticing

$$\nabla^2\widehat{\mathcal{L}}_n(\boldsymbol{\beta}) = -\frac{2}{n(n-1)}\sum_{i<i'}\alpha_n^2(\boldsymbol{X}_i - \boldsymbol{X}_{i'})(\boldsymbol{X}_i - \boldsymbol{X}_{i'})^\top F''_{ii'}(\widetilde{S}_{ii'}\alpha_n Z_{ii'}(\boldsymbol{\beta}))$$

is very easy to calculate, we note Assumption **(A5)** could be verified empirically. As a matter of fact, in the appendix, we will show a lot of statistical models satisfy Assumption **(A5)** via exhaustive simulation studies.

Remark 4.9. Of note, [22] has shown that, under certain regularity conditions, $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ is positive definite. Using very similar arguments, we can show $\nabla^2\mathcal{L}_n(\boldsymbol{\beta}^*) = \mathbb{E}\nabla^2\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*)$ is positive definite. Accordingly, by continuity, Assumption (**A5**) holds with high probability when d/n is relatively small. However, empirical results in the appendix Section B.1 show that, even when the population design matrix's condition number is very small and for d/n very large (e.g., $n = 50$ and $d = 800$), the convexity property still holds with high probability.

Before presenting the main result in this section, let's define an additional parameter. We write the cone

$$\mathcal{H} := \{\boldsymbol{v} \in \mathbb{R}^d : v_1 = 1, \|\boldsymbol{v} - \boldsymbol{\beta}^*\|_2 \leq r(\gamma), \|\boldsymbol{v}_{S^c}\|_1 \leq 3\|\boldsymbol{v}_S\|_1\}$$

with S representing the index set of nonzero-entries in $\boldsymbol{\beta}^*_{r,\alpha_n}$. We define a parameter κ_n controlling the uniform convergence of $\widehat{\mathcal{L}}_n(\cdot)$ to $\mathcal{L}_n(\cdot)$ within the cone $\mathcal{H}$:

$$\kappa_n := \mathbb{E}\sup_{\mathcal{H}}|\widehat{\mathcal{L}}_n(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta})|,$$

where the expectation is in the outer probability sense. Of note, because both $\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ and $\mathcal{L}_n(\boldsymbol{\beta})$ are bounded, we have $\kappa_n \lesssim 1$, but κ_n could be much smaller. In particular, we have $\kappa_n = O(n^{-1/2})$ by standard U-statistics empirical process theory [58, 59, 60] when we fix d and choose $r(\gamma) = o(1)$. More refined theoretical evaluation on the order of κ_n could be found in [61].

With Assumptions **(A0)-(A6)**, the following lemma then characterizes the stochastic error of the specified minimum $\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n}$ to its population counterpart $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$.

Lemma 4.10. If Assumptions **(A0)-(A6)** all hold, we then have, when $\lambda_n \asymp \alpha_n\sqrt{\log d/n}$, for large enough n, $\|\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_2^2 \overset{\mathbb{P}}{\lesssim} \alpha_n^2 s_n \log d/n + \alpha_n^{-1} + \kappa_n$.

4.3. Main results

Combining Lemmas 4.3 and 4.10, we are now ready to provide the main theorem. It characterizes the consistency property of the proposed smoothing estimators under a specified scaling condition.

Theorem 4.11. Assume Assumptions **(A0)-(A6)** hold for some pre-chosen

$$\alpha_n \asymp \{n/(s_n \log d)\}^{1/3}.$$

Assume $\widehat{\boldsymbol{\beta}}_{\alpha_n}$ is a stationary point of optimization problem (2.4), satisfying $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2 \leq r$ for some constant r depending on $r(\gamma)$. Then $\widehat{\boldsymbol{\beta}}_{\alpha_n}$ exists and satisfies that, as long as $s_n \log d/n \to 0$ and $\kappa_n \to 0$, we have $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2 \xrightarrow{\mathbb{P}} 0$. In particular, when d is fixed, we have $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2 \xrightarrow{\mathbb{P}} 0$.

In practice, it is very difficult to theoretically calculate exactly how large d is allowed to be using Theorem 4.11. However, a rule of thumb in our case, mimicking the corresponding ones in robust statistics (cf. [62]), is $10 \log d \leq n^{1/3}$.

5. Discussions

In the manuscript, we are focused on studying the lasso-type penalty of formulation (2.3). It should be highlighted that SCAD [63] and MCP [64] type penalties could be implemented and studied in the same manner, and similar theoretical and empirical performances can be expected. Since the extension from lasso-type penalty to non-convex ones is beyond the interest of this paper, we decide to leave them for future studies. Another important issue which is only mildly touched is model diagnostic check. To our knowledge, there has not been much study on goodness-of-fit test of Han's model (1.1) in high dimensions. In this manuscript we provided several heuristics to check monotonicity of Y and a single index of $\boldsymbol{X}$ (cf. Section 3 and Section C.3 in the appendix). In the future, it will be interesting to develop a theoretically solid test for it.

This manuscript has shown the advantage of smoothed maximum rank correlation method both theoretically and empirically. We close this section with a discussion on some limitations. (i) Computationally, as shown in the appendix Section A.2.3, though better than several semiparametric competitors, RMRCE is significantly worse than the lasso in terms of demanding much more time in implementation. (ii) Theoretically, if the true generating model is linear and the noise is Gaussian, RMRCE loses efficiency compared to the lasso. Accordingly, if the practitioner has a strong belief that, for instance, a simple linear model of Gaussian noise applies to the studied data, then the lasso is recommended compared to RMRCE.

Appendix A: Synthetic data analysis

This section empirically examines the finite sample performance of the proposed Regularized Maximum Rank Correlation Estimator (RMRCE) $\widehat{\boldsymbol{\beta}}_{\alpha_n}$ in (2.4) using the synthetic data. We demonstrate various empirical results to compare the proposed methods to the competing methods. Our simulation highlights the distinctive attributes of the proposed method, that is, it has comparable performance to the lasso under the high dimensional linear model, but beats the lasso under non-linear models. The proposed methods outperform Hinge, SIM, and SDR under both linear and non-linear models.

A.1. Algorithm description

We exploit the coordinate descent algorithm [65] without penalization for the first term to solve (2.4). This problem falls in the application regime of the coordinate descent algorithm theory [17]. For the comparison fairness, in the sequel we also do not penalize the first term in implementing the lasso.

One issue in the implementation is on choosing the tuning parameter λ_n and the smoothing parameter α_n. For tuning λ_n and α_n, we propose to use five-fold cross-validation. Using five randomly split subsets of equal size, we define the following loss function (A.1):

$$CV(\lambda_n, \alpha_n) := \frac{1}{5}\sum_{k=1}^{5}\binom{n_k}{2}^{-1}\sum_{1\le l<l'\le n_k}\operatorname{sign}(Y_{i_l}-Y_{i_{l'}})\operatorname{sign}(\boldsymbol{X}_{i_l}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\alpha_n}^{(-k)}(\lambda_n)-\boldsymbol{X}_{i_{l'}}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\alpha_n}^{(-k)}(\lambda_n)), \tag{A.1}$$

where n_k is the number of data points in the k-th part, and $\widehat{\boldsymbol{\beta}}_{\alpha_n}^{(-k)}(\lambda_n)$ is obtained from the other 4 parts of the training data with the tuning parameter λ_n and smoothing parameter α_n. We then select the λ_n and α_n that maximize $CV(\lambda_n, \alpha_n)$ over a grid of possible values (λ_n, α_n). In addition, when α_n is pre-chosen, λ_n is chosen to optimize (A.1) with the chosen α_n.

Another issue in the implementation is on choosing the starting point for the coordinate descent algorithm. For this, we consider the following strategy proposed in [66]. Specifically, let $\{\mathbf{e}_j \in \mathbb{R}^d, 1 \le j \le d\}$ be the standard basis with the j-th entry equal to 1, and 0 at all the other entries. The algorithm starts with $\widehat{\boldsymbol{\beta}}^{(0)} = \operatorname{sign}(L_{j^*})\mathbf{e}_{j^*}$ by selecting the j^*-th coordinate with the index j^* that maximizes the absolute value of L_j with respect to $j \in \{2, \ldots, d\}$, i.e.,

$$j^* = \underset{j\in\{2,\ldots,d\}}{\operatorname{argmax}}\left\{|L_j| : L_j := \binom{n}{2}^{-1}\sum_{1\le i<i'\le n}\operatorname{sign}(Y_i-Y_{i'})\operatorname{sign}(X_{ij}-X_{i'j})\right\}.$$

In practice, we could combine other starting point candidates like the lasso and sieve solutions, and select the one that maximizes the objective function in (2.2). However, our numerical results indicate that the above simple strategy has worked very well in a variety of applications.

A.2. Synthetic data analysis

This section investigates the empirical performance of the proposed methods via synthetic data analysis. We compare the proposed methods to Hinge, the lasso, SIM, and SDR. The descriptions of these methods are in Section 3. First, we show, under the high dimensional linear model, the proposed methods perform reasonably well compared to the lasso. Secondly, we show the proposed methods beat the lasso under a monotonic transformation model (1.2). Thirdly, the proposed methods beat Hinge, SIM, and SDR under both linear and non-linear models.

We also explore the performance of the proposed methods with the tuning parameter α_n chosen by cross validation or set to be fixed values $1, 3, 5, 7, 9$. The results show that, as long as α_n is comparatively large ($\alpha_n \geq 3$), the estimates and selection results are not sensitive to α_n. In addition, choosing α_n by cross validation only leads to marginal improvement of performance compared to fixed α_n.

We tried $d = 50$ and 200 with a series of sample sizes for variable selection and estimation. Due to the space limit, below we mainly show the results about variable selection and estimation performance with $d = 200$. Similar patterns hold for $d = 50$, some of whose results are put in later sections.

A.2.1. High dimensional linear model

This section is focused on the high dimensional linear model[1],

$$Y_i = \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\beta}^0 + \epsilon_i, i = 1, 2, \cdots, n, \tag{A.2}$$

with $\boldsymbol{\beta}^0 = (5, 4, 3, 2, 1, -1, -3, -5, 0, \cdots, 0)^{\mathsf{T}}$ and $\boldsymbol{X}_i$ a d dimensional random vector generated from a multivariate normal distribution $\mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma} = ((\sigma_{jk})))$ with $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \leq j, k \leq d$. Here the noise was generated from the standard normal and independent of the covariates. Note the difference between (A.2) and (1.2) in terms of $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0/\beta_1^0$. We also consider all three smoothing approximations.

We first compare the variable selection performance with $d = 200$. To this end, Figures 5 and 6 plot the receiver operating characteristic (ROC) curves for RMRCE with different choices of tuning parameter α_n, and all the competing methods. These results are calculated based on 200 replications. It shows that the proposed methods have overall good variable selection performance compared to the competing methods. Detailed comments are as follows. (i) With sample size increasing, the ROC curves shift towards the upper left corner, as what we expect. (ii) The difference between different smoothing approximations and different choices of tuning parameter α_n is very slight via observing Figure 5. The performance of RMRCE with α_n selected using cross validation

[1] We have tried a lot of different combinations of $D(\cdot)$ and $\Lambda(\cdot, \cdot)$, confirming that the results are very similar to the linear case. Due to the space limit, we do not list them all in this paper.

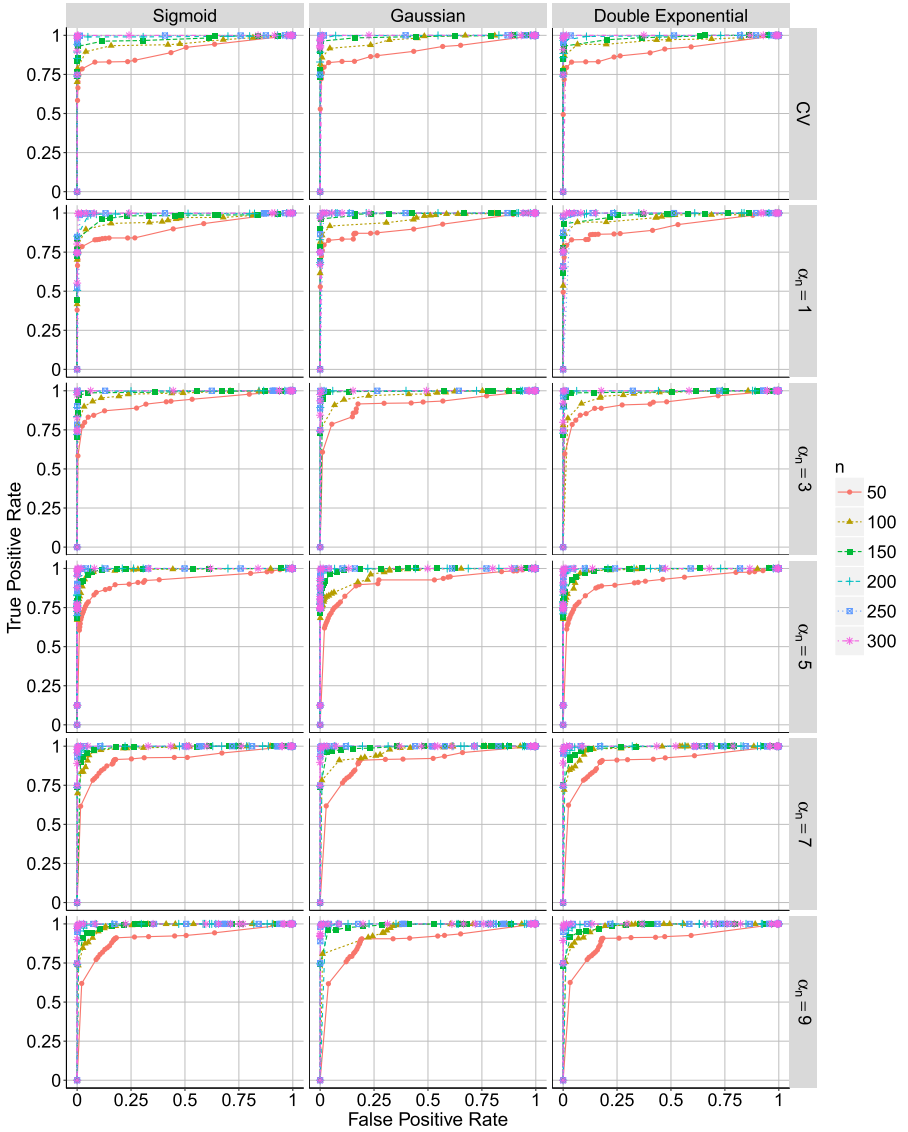


FIG 5. ***ROC curves for RMRCE under the high dimensional linear model.*** *α_n is selected by cross validation (CV) or set to be $1, 3, 5, 7, 9$. The results for Sigmoid, Gaussian and double exponential smoothing approximations are presented. The results are based on 200 replications with $d = 200$.*

is slightly better than those with fixed α_n. (iii) The lasso has slightly better variable selection performance as shown in Figure 6, especially regarding the small sample size. This is reasonable since the simulation setup in (A.2) is the right model for the lasso. (iv) RMRCE has better performance than Hinge. This

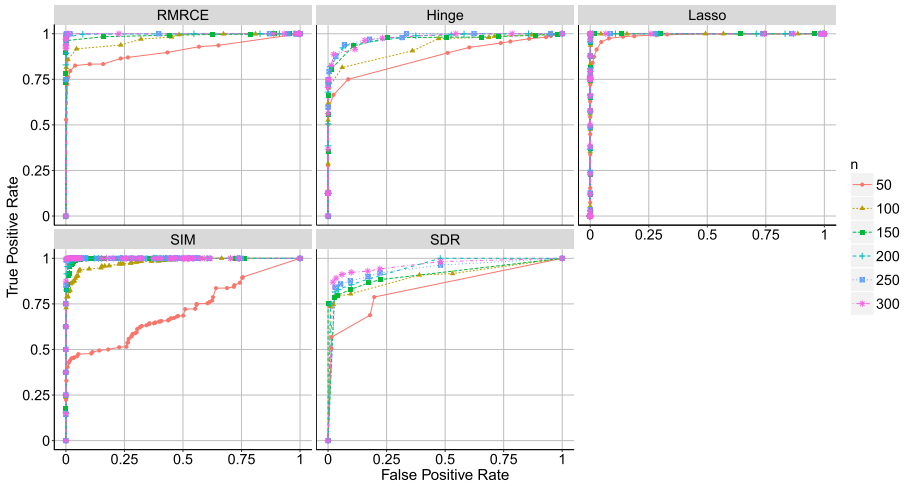


FIG 6. ***ROC curves for all competing methods under the high dimensional linear model.*** *Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $d = 200$.*

shows that the smoothed rank correlation in RMRCE is better than the convex relaxation in Hinge. (v) RMRCE has better performance than the other two semiparametric regression methods: SIM and SDR. Table 1 gives the averaged false positive rates and the averaged true positive rates for all the competing methods. Similar observations to Figure 6 were observed.

Secondly, we focus on comparing the estimation errors with $d = 200$. Tables 2 and 3 present the estimation errors to $\boldsymbol{\beta}^*$ compared to the tuning parameter λ_n for RMRCE with different choices of tuning parameter α_n, and all the competing methods across 200 replications. Note, for lasso, the estimation error is calculated as $\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda_n)/\widehat{\beta}_1^{\text{lasso}}(\lambda_n) - \boldsymbol{\beta}^*\|_2$. The estimation errors are calculated in the same way for SIM and SDR. The results show that for RMRCE, Hinge, and the lasso, the estimation error decreases first and then increases as tuning parameter increases. For SIM and SDR, the estimation error decreases first and then almost stays constant as the number of selected variables increases. Besides, with increased sample size, the estimation error curves shift downward to 0. Table 2 also shows that the difference between the proposed methods with different approximation functions and different choices of α_n is very mild as long as α_n is large enough ($\alpha_n \geq 3$). The performance of RMRCE with $\alpha_n = 1$ is comparatively worse than RMRCE with $\alpha_n \geq 3$. Table 2 and 3 further show, under the high dimensional linear model (A.2), the proposed methods perform reasonably well, although slightly worse than the lasso[2]. In

[2]Note that for fair comparison, the results for the lasso are based on the standardized estimation error $\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}/\widehat{\beta}_1^{\text{lasso}} - \boldsymbol{\beta}^*\|_2$ with the lasso estimator $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ of $\boldsymbol{\beta}^0$. Same for SIM and SDR.

TABLE 1

Variable selection performance for all competing methods under the high dimensional linear model. *Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $n = 100$ and $d = 200$. "FPR" stands for averaged false positive rates with standard deviations in parentheses. "TPR" stands for averaged true positive rates with standard deviations in parentheses. For SIM and SDR, N stands for the number of selected variables.*

(a) RMRCE

λ_n	FPR	TPR
0.005	0.321 (0.022)	0.999 (0.012)
0.010	0.234 (0.025)	0.999 (0.012)
0.030	0.009 (0.012)	0.858 (0.111)
0.050	0.001 (0.002)	0.792 (0.062)
0.100	0.001 (0.002)	0.751 (0.110)

(b) Hinge

λ_n	FPR	TPR
0.040	0.870 (0.166)	0.994 (0.027)
0.060	0.687 (0.188)	0.986 (0.039)
0.080	0.472 (0.141)	0.980 (0.049)
0.100	0.360 (0.307)	0.906 (0.098)
0.150	0.060 (0.081)	0.815 (0.117)

(c) Lasso

λ_n	FPR	TPR
0.100	0.157 (0.022)	1.000 (0.000)
0.300	0.014 (0.009)	0.999 (0.009)
0.500	0.001 (0.003)	0.937 (0.080)
0.700	0.000 (0.001)	0.833 (0.081)
1.000	0.000 (0.000)	0.775 (0.050)

(d) SIM

N	FPR	TPR
5.000	0.000 (0.001)	0.623 (0.015)
10.000	0.018 (0.003)	0.818 (0.076)
20.000	0.166 (0.262)	0.948 (0.083)
30.000	0.309 (0.354)	0.976 (0.054)
50.000	0.719 (0.352)	0.996 (0.025)

(e) SDR

N	FPR	TPR
2.000	0.009 (0.004)	0.789 (0.059)
5.000	0.022 (0.011)	0.736 (0.091)
10.000	0.031 (0.020)	0.748 (0.086)
25.000	0.097 (0.004)	0.805 (0.089)
50.000	0.390 (0.033)	0.908 (0.070)

addition, the proposed methods have better performance than Hinge, SIM, and SDR.

A.2.2. High dimensional generalized regression model

This section considers the monotonic transformation model (1.2) with $G(x) = x^3$:

$$Y_i = (\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}^0 + \epsilon_i)^3, i = 1, 2, \cdots, n, \tag{A.3}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{\beta}^0$ are the same as in Model (A.2), and ϵ_i follows the standard normal distribution and is independent of $\boldsymbol{X}_i$. This is the setting where the lasso could still possibly provide a consistent estimator [67]. But the estimation/model selection efficiency is expected to be low, which will be demonstrated empirically here. Note that under Model (A.3), the proposed methods and Hinge have exactly the same results as under Model (A.2). For the convenience of comparison, we still include the results of Hinge and RMRCE (α_n chosen by cross validation, Gaussian approximation).

TABLE 2

Averaged estimation errors for RMRCE under the high dimensional linear model. *Standard deviations are shown in parentheses.* α_n *is selected by cross validation or set to be* $1, 3, 5, 7, 9$. *The results for Sigmoid, Gaussian and double exponential (DE) smoothing approximations are presented. The results are based on 200 replications with* $d = 50$ *or* $d = 200$.

(a) α_n selected by cross validation

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.068 (0.038)	0.139 (0.089)	0.099 (0.047)	0.137 (0.104)	0.081 (0.039)	0.136 (0.101)
100	0.021 (0.013)	0.033 (0.018)	0.026 (0.014)	0.035 (0.026)	0.028 (0.017)	0.046 (0.022)
150	0.013 (0.006)	0.010 (0.008)	0.012 (0.017)	0.006 (0.010)	0.014 (0.008)	0.016 (0.014)
200	0.008 (0.004)	0.008 (0.009)	0.003 (0.004)	0.003 (0.008)	0.010 (0.005)	0.006 (0.008)
250	0.006 (0.003)	0.007 (0.005)	0.002 (0.003)	0.004 (0.002)	0.006 (0.005)	0.005 (0.004)
300	0.004 (0.003)	0.002 (0.004)	0.001 (0.007)	0.004 (0.002)	0.004 (0.003)	0.003 (0.002)

(b) $\alpha_n = 1$

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.239 (0.136)	0.242 (0.197)	0.113 (0.047)	0.137 (0.104)	0.109 (0.049)	0.136 (0.101)
100	0.150 (0.061)	0.156 (0.061)	0.088 (0.022)	0.077 (0.021)	0.081 (0.022)	0.088 (0.022)
150	0.133 (0.044)	0.123 (0.038)	0.084 (0.019)	0.071 (0.011)	0.082 (0.018)	0.066 (0.015)
200	0.130 (0.037)	0.114 (0.022)	0.083 (0.013)	0.071 (0.008)	0.078 (0.015)	0.066 (0.011)
250	0.125 (0.031)	0.110 (0.017)	0.078 (0.013)	0.070 (0.008)	0.078 (0.014)	0.066 (0.009)
300	0.120 (0.025)	0.112 (0.015)	0.074 (0.011)	0.080 (0.008)	0.076 (0.012)	0.070 (0.007)

(c) $\alpha_n = 3$

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.068 (0.038)	0.139 (0.089)	0.099 (0.047)	0.169 (0.098)	0.081 (0.039)	0.170 (0.091)
100	0.021 (0.013)	0.033 (0.018)	0.026 (0.014)	0.035 (0.026)	0.028 (0.017)	0.046 (0.022)
150	0.015 (0.009)	0.010 (0.008)	0.013 (0.006)	0.006 (0.010)	0.014 (0.008)	0.016 (0.014)
200	0.011 (0.007)	0.008 (0.004)	0.003 (0.004)	0.003 (0.008)	0.010 (0.005)	0.006 (0.008)
250	0.009 (0.005)	0.007 (0.003)	0.002 (0.003)	0.004 (0.002)	0.006 (0.005)	0.005 (0.004)
300	0.007 (0.004)	0.007 (0.002)	0.001 (0.002)	0.004 (0.002)	0.004 (0.003)	0.003 (0.002)

(d) $\alpha_n = 5$

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.095 (0.045)	0.175 (0.099)	0.119 (0.054)	0.208 (0.098)	0.111 (0.051)	0.169 (0.090)
100	0.025 (0.014)	0.051 (0.020)	0.036 (0.073)	0.072 (0.027)	0.033 (0.015)	0.061 (0.021)
150	0.013 (0.006)	0.026 (0.008)	0.012 (0.017)	0.042 (0.016)	0.017 (0.007)	0.036 (0.011)
200	0.008 (0.004)	0.017 (0.006)	0.009 (0.011)	0.013 (0.013)	0.011 (0.005)	0.023 (0.010)
250	0.006 (0.003)	0.009 (0.006)	0.008 (0.008)	0.010 (0.007)	0.008 (0.004)	0.013 (0.008)
300	0.004 (0.002)	0.002 (0.004)	0.008 (0.005)	0.008 (0.005)	0.006 (0.003)	0.005 (0.005)

(e) $\alpha_n = 7$

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.104 (0.050)	0.171 (0.091)	0.140 (0.068)	0.304 (0.130)	0.122 (0.056)	0.232 (0.105)
100	0.030 (0.016)	0.046 (0.027)	0.044 (0.022)	0.064 (0.037)	0.040 (0.020)	0.063 (0.033)
150	0.017 (0.007)	0.017 (0.012)	0.018 (0.012)	0.031 (0.018)	0.021 (0.009)	0.028 (0.015)
200	0.010 (0.007)	0.008 (0.009)	0.012 (0.009)	0.019 (0.009)	0.012 (0.008)	0.015 (0.011)
250	0.007 (0.005)	0.007 (0.005)	0.008 (0.006)	0.016 (0.006)	0.009 (0.006)	0.012 (0.006)
300	0.004 (0.003)	0.006 (0.003)	0.001 (0.007)	0.017 (0.005)	0.006 (0.004)	0.011 (0.004)

(f) $\alpha_n = 9$

n	Sigmoid (d=50)	Sigmoid (d=200)	Gaussian (d=50)	Gaussian (d=200)	DE (d=50)	DE (d=200)
50	0.120 (0.055)	0.211 (0.099)	0.157 (0.073)	0.395 (0.389)	0.140 (0.065)	0.313 (0.133)
100	0.037 (0.018)	0.075 (0.027)	0.049 (0.023)	0.076 (0.041)	0.044 (0.022)	0.100 (0.035)
150	0.020 (0.009)	0.021 (0.019)	0.026 (0.012)	0.021 (0.067)	0.024 (0.011)	0.029 (0.023)
200	0.014 (0.007)	0.013 (0.011)	0.013 (0.010)	0.023 (0.010)	0.017 (0.008)	0.019 (0.013)
250	0.008 (0.005)	0.011 (0.006)	0.009 (0.007)	0.019 (0.007)	0.010 (0.007)	0.015 (0.007)
300	0.006 (0.004)	0.010 (0.004)	0.001 (0.007)	0.019 (0.005)	0.008 (0.004)	0.014 (0.005)

TABLE 3

Averaged estimation errors for all competing methods under the high dimensional linear model. *Standard deviations are shown in parentheses. Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $d = 50$ or $d = 200$. The tuning parameter λ_n (RMRCE, Hinge, and the lasso) or number of selected variables (SIM and SDR) is determined through cross-validation.*

(a) RMRCE

	d	
n	50	200
50	0.099 (0.047)	0.137 (0.104)
100	0.026 (0.014)	0.035 (0.026)
150	0.012 (0.017)	0.006 (0.010)
200	0.003 (0.004)	0.003 (0.008)
250	0.002 (0.003)	0.004 (0.002)
300	0.001 (0.007)	0.004 (0.002)

(b) Hinge

	d	
n	50	200
50	1.963 (0.248)	2.143 (0.353)
100	2.580 (0.230)	2.199 (0.271)
150	2.430 (0.203)	2.220 (0.232)
200	2.260 (0.212)	2.259 (0.155)
250	2.323 (0.191)	2.591 (0.121)
300	2.362 (0.217)	2.257 (0.109)

(c) Lasso

	d	
n	50	200
50	0.034 (0.021)	0.064 (0.035)
100	0.012 (0.007)	0.020 (0.009)
150	0.008 (0.004)	0.011 (0.005)
200	0.006 (0.003)	0.008 (0.004)
250	0.004 (0.002)	0.006 (0.003)
300	0.003 (0.002)	0.005 (0.002)

(d) SIM

	d	
n	50	200
50	1.109 (0.424)	2.080 (0.937)
100	0.822 (0.535)	0.902 (0.000)
150	0.815 (0.004)	0.831 (0.014)
200	0.812 (0.003)	0.824 (0.008)
250	0.809 (0.002)	0.816 (0.000)
300	0.810 (0.002)	0.814 (0.000)

(e) SDR

	d	
n	50	200
50	1.952 (3.210)	2.931 (2.246)
100	1.038 (3.691)	1.046 (3.630)
150	0.950 (3.734)	0.921 (0.015)
200	0.888 (3.746)	1.025 (3.452)
250	0.861 (3.728)	0.876 (3.676)
300	0.854 (3.764)	0.854 (3.729)

First, we reveal the variable selection performance for all competing methods under Model (A.3) by plotting the ROC curves in Figure 7 with $d = 200$. It confirms that the lasso as well as other competing methods perform much worse than the proposed methods. Table 4 further illustrates the averaged TPRs and FPRs along with their standard deviations over 200 replications. As shown in Table 4, Hinge, the lasso, SIM, and SDR have much worse performance than the proposed methods. Secondly, Table 5 shows the normalized ℓ_2 estimation error. For the lasso, the estimation error is calculated as $\|\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda_n)/\widehat{\beta}_1^{\text{lasso}}(\lambda_n) - \boldsymbol{\beta}^*\|_2$. The estimation errors are calculated in the same way for SIM and SDR. With $d = 200$, the comparison between the methods confirms the advantage of the proposed method.

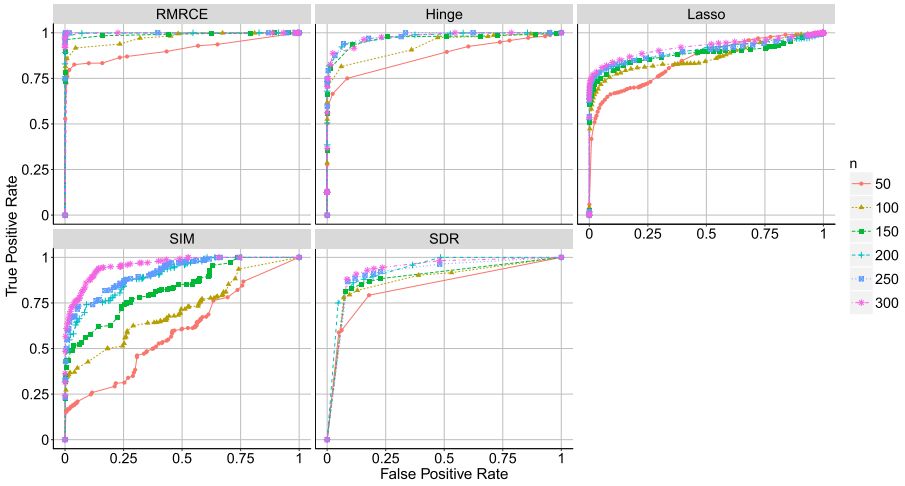


FIG 7. ***ROC curves for all competing methods with*** $G(x) = x^3$***.*** *Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $d = 200$.*

TABLE 4

Variable selection performance for all competing methods with $G(x) = x^3$***.*** *Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $n = 100$ and $d = 200$. "FPR" stands for averaged false positive rates with standard deviations in parentheses. "TPR" stands for averaged true positive rates with standard deviations in parentheses. For SIM and SDR, N stands for the number of selected variables.*

(a) RMRCE

λ_n	FPR	TPR
0.005	0.321 (0.022)	0.999 (0.012)
0.010	0.234 (0.025)	0.999 (0.012)
0.030	0.009 (0.012)	0.858 (0.111)
0.050	0.001 (0.002)	0.792 (0.062)
0.100	0.001 (0.002)	0.751 (0.110)

(b) Hinge

λ_n	FPR	TPR
0.040	0.870 (0.166)	0.994 (0.027)
0.060	0.687 (0.188)	0.986 (0.039)
0.080	0.472 (0.141)	0.980 (0.049)
0.100	0.360 (0.307)	0.906 (0.098)
0.150	0.060 (0.081)	0.815 (0.117)

(c) Lasso

λ_n	FPR	TPR
1.000	0.881 (0.076)	0.985 (0.044)
5.000	0.600 (0.042)	0.889 (0.094)
10.000	0.542 (0.032)	0.859 (0.095)
50.000	0.432 (0.032)	0.832 (0.098)
100.000	0.365 (0.044)	0.828 (0.093)
500.000	0.116 (0.061)	0.774 (0.089)

(d) SIM

N	FPR	TPR
5	0.012 (0.005)	0.327 (0.132)
10	0.052 (0.102)	0.393 (0.191)
20	0.423 (0.379)	0.652 (0.324)
30	0.578 (0.387)	0.756 (0.304)
40	0.646 (0.380)	0.796 (0.280)
50	0.727 (0.201)	0.883 (0.204)

(e) SDR

N	FPR	TPR
5	0.071 (0.035)	0.779 (0.098)
10	0.096 (0.050)	0.795 (0.098)
20	0.126 (0.068)	0.809 (0.084)
30	0.354 (0.217)	0.851 (0.096)
40	0.531 (0.397)	0.916 (0.099)
50	0.390 (0.033)	0.902 (0.066)

TABLE 5

Averaged estimation errors for all competing methods with $G(x) = x^3$***.*** *Standard deviations are shown in parentheses. Methods included are: RMRCE (α_n selected by cross validation, Gaussian smoothing approximation), Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with $d = 50$ or $d = 200$. The tuning parameter λ_n (RMRCE, Hinge, and the lasso) or number of selected variables (SIM and SDR) is determined through cross-validation.*

(a) RMRCE

	d	
n	50	200
50	0.099 (0.047)	0.137 (0.104)
100	0.026 (0.014)	0.035 (0.026)
150	0.012 (0.017)	0.006 (0.010)
200	0.003 (0.004)	0.003 (0.008)
250	0.002 (0.003)	0.004 (0.002)
300	0.001 (0.007)	0.004 (0.002)

(b) Hinge

	d	
n	50	200
50	1.963 (0.248)	2.143 (0.353)
100	2.580 (0.230)	2.199 (0.271)
150	2.430 (0.203)	2.220 (0.232)
200	2.260 (0.212)	2.259 (0.155)
250	2.323 (0.191)	2.591 (0.121)
300	2.362 (0.217)	2.257 (0.109)

(c) Lasso

	d	
n	50	200
50	1.514 (0.347)	2.163 (0.293)
100	0.629 (0.129)	0.998 (0.213)
150	0.478 (0.092)	0.614 (0.133)
200	0.407 (0.076)	0.415 (0.084)
250	0.306 (0.058)	0.382 (0.077)
300	0.260 (0.045)	0.306 (0.054)

(d) SIM

	d	
n	50	200
50	2.589 (0.277)	2.806 (0.926)
100	0.898 (1.301)	2.061 (0.812)
150	0.860 (1.491)	0.910 (0.243)
200	0.831 (0.023)	0.897 (0.928)
250	0.836 (2.932)	0.859 (3.495)
300	0.827 (1.480)	0.850 (0.007)

(e) SDR

	d	
n	50	200
50	3.947 (1.824)	2.504 (3.062)
100	1.096 (3.586)	1.054 (3.636)
150	1.009 (3.636)	0.970 (3.712)
200	0.889 (3.759)	1.599 (3.606)
250	0.862 (3.746)	0.883 (3.732)
300	0.857 (3.763)	0.858 (3.702)

A.2.3. Comparison of computation time

Table 6 shows the median of the computation time of 200 replicated runs for RMRCE, Hinge, the lasso, SIM, and SDR with different n and $d = 50$. For demonstration purpose we only present the computation time with fixed tuning parameters. For RMRCE, the Gaussian approximation is used and the tuning parameters are set as $\alpha_n = 5$ and $\lambda_n = 1$. For Hinge and the lasso we set the tuning parameter as $\lambda_n = 1$. SIM and SDR are run with the default parameters. The results are similar with other choices of the tuning parameters, as long as they are in a reasonable scale.

The lasso is the most efficient among all methods, and its computation time is close to zero. RMRCE takes significantly more time to finish compared to the lasso, but is still much faster than SIM and SDR. In addition, RMRCE is slightly slower than Hinge since in Hinge the non-convex rank correlation is

TABLE 6

Median computation time (seconds) for all competing methods. *Methods included are: RMRCE, Hinge, the lasso, SIM, and SDR. The results are based on 200 replications with different n, $d = 50$, and fixed tuning parameters α_n and λ_n.*

	n					
Method	50	100	150	200	250	300
RMRCE	1.350	5.177	14.872	24.83	30.638	33.771
Hinge	0.263	0.822	1.627	3.242	4.910	6.618
Lasso	0.002	0.011	0.157	0.002	0.002	0.002
SIM	72.202	144.404	168.132	190.23	218.331	316.347
SDR	59.849	122.514	132.736	157.172	198.831	237.852

replaced with a convex function, making the optimization faster. In summary RMRCE is much faster compared to SIM and SDR, but not as efficient as Hinge and the lasso.

Appendix B: Empirical verification of the theory

This section examines Theorem 4.11 and Assumption **(A5)** in Section 4.2 using synthetic data.

B.1. Convexity verification

For investigating the stochastic error $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*_{\alpha_n}\|_2$, Assumption **(A5)** is required to hold. In this section, we verify Assumption **(A5)** via various empirical studies on the positive definiteness of the following Hessian matrix:

$$\nabla^2 \widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*) = -\frac{2}{n(n-1)} \sum_{i<i'} \alpha_n^2 (\boldsymbol{X}_i - \boldsymbol{X}_{i'})(\boldsymbol{X}_i - \boldsymbol{X}_{i'})^{\mathsf{T}} F''_{ii'}(\widetilde{S}_{ii'} \alpha_n Z_{ii'}(\boldsymbol{\beta}^*)).$$

For presentation clearness, we focus on the high dimensional linear model (A.2) with $\boldsymbol{\beta}^0 = (5, 4, 3, 2, 1, -1, -3, -5, 0, \cdots, 0)^{\mathsf{T}}$ and $\boldsymbol{X}_i$ a d dimensional random vector generated from a multivariate normal distribution $\mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma} = ((\sigma_{jk})))$ with $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \le j, k \le d$. We further generate ϵ_i from a mixture of the standard normal and $\delta = 0\%$ or 20% of the outliers following Cauchy$(0, 0.01)$. The noise is independent of $\boldsymbol{X}_i$.

We demonstrate the results under different situations with different noise distributions as well as different smoothing approximations. Table 7 gives the results, considering all three examples of smoothing approximations in Remark 2.2, with pure Gaussian noise and the mixture of Gaussian noise with 20% Cauchy outliers. Here, via exhaustive simulation studies, α_n is recommended to be 5 for sigmoid, Gaussian, and double exponential CDF approximations. The results are calculated based on 200 replications.

There are several noteworthy discoveries. First, the performances of all cases are similar. Specifically, as sample size n increases, the proportion of positive definite Hessian matrices increases to 1. In addition, small dimension d leads

TABLE 7

Proportion of positive definite Hessian matrices. *Results are based on 200 replications. The first row is for $\delta = 0$ and the second row is for $\delta = 0.2$. "DE" stands for double exponential CDF approximation.*

(a) Sigmoid, $\delta = 0$

	d		
n	50	200	800
50	0.830	0.815	0.850
100	1.000	0.925	0.910
150	1.000	0.985	0.935
200	1.000	1.000	0.955
250	1.000	1.000	0.940
300	1.000	1.000	0.970

(b) Gaussian, $\delta = 0$

	d		
n	50	200	800
50	0.865	0.840	0.875
100	1.000	0.955	0.930
150	1.000	0.990	0.975
200	1.000	1.000	0.965
250	1.000	1.000	0.970
300	1.000	1.000	0.980

(c) DE, $\delta = 0$

	d		
n	50	200	800
50	0.815	0.845	0.870
100	1.000	0.920	0.925
150	1.000	0.975	0.950
200	1.000	0.995	0.945
250	1.000	1.000	0.980
300	1.000	1.000	0.975

(d) Sigmoid, $\delta = 0.2$

	d		
n	50	200	800
50	0.890	0.845	0.880
100	1.000	0.925	0.925
150	1.000	0.945	0.955
200	1.000	0.980	0.950
250	1.000	0.990	0.960
300	1.000	0.990	0.965

(e) Gaussian, $\delta = 0.2$

	d		
n	50	200	800
50	0.900	0.870	0.910
100	1.000	0.930	0.930
150	1.000	0.975	0.945
200	1.000	0.990	0.955
250	1.000	0.990	0.950
300	1.000	0.995	0.960

(f) DE, $\delta = 0.2$

	d		
n	50	200	800
50	0.845	0.865	0.865
100	1.000	0.945	0.925
150	1.000	0.965	0.935
200	1.000	0.970	0.930
250	1.000	0.985	0.955
300	1.000	0.995	0.950

to higher proportion of convexity. Secondly, for comparison between different noise terms, we find pure Gaussian noise enjoys higher proportion of convexity compared to the mixture noise with large n. Thirdly, for comparison between different smoothing approximations, they perform similarly, though the Gaussian CDF approximation enjoys slightly higher proportions of positive definite Hessian matrices.

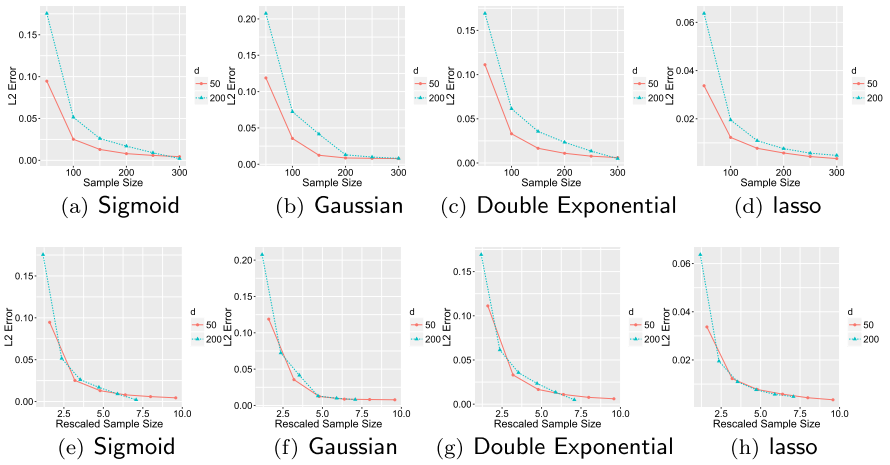


FIG 8. ***Estimation error curves for different methods.*** *The first row is for estimation error curves compared to the sample size n, and the second row is for estimation error curves compared to the rescaled sample size $n/(s \log d)$.*

B.2. Verification of the theorem

This section verifies the property given in Theorem 4.11. Simulations with different generating models have shown that the proposed methods are robust to the monotonic functions $D(\cdot)$ and $\Lambda(\cdot, \cdot)$. Hence to demonstrate the scaling property of the estimation errors $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2$, we only focus on the high dimensional linear model (A.2) with ϵ_i generated from the standard normal and independent of $\boldsymbol{X}_i$. For simplicity, we only show the results for RMRCE with $\alpha = 5$. The results are similar for other choices of α_n.

To this end, Figure 8 illustrates the scaling property of the averaged estimation errors $\|\widehat{\boldsymbol{\beta}}_{\alpha_n} - \boldsymbol{\beta}^*\|_2$ compared to n and $n/(s \log d)$ across 200 replications for different approximation functions. The performance is also compared to the lasso.

Figure 8 clearly shows a "stacking curve" phenomena. Specifically, regarding the rescaled sample size $n/(s \log d)$, the error curves corresponding to different d's are all aligned together. In addition, the error decays to zero as sample size n increases, and increases as d increases. Hence Figure 8 confirms the theoretical discovery in Theorem 4.11. For comparison, we also include the lasso's stacking curves in Figure 8(d) and 8(h). They show similar "stacking curve" phenomena.

Appendix C: Additional materials for real data example

C.1. Data processing and normalization for TF prediction

We downloaded DNase-seq and gene expression exon array data from 57 different cell types generated by ENCODE at the University of Washington [68]. Exon

array samples were consistently normalized using the GeneBASE software [69]. The output of GeneBASE was gene-level expression values for all genes. From these values, we extracted expression values for all human TFs documented in the Animal Transcription Factor Database (AminalTFDB) [70]. We then filtered out TFs whose expression values were nearly constant across all samples (defined as coefficient of variation < 0.2) because one would expect them to behave like an intercept term in a regression model. These nearly-constant TFs were not expected to provide much information to explain variation of the response variable across different cell types. After filtering, 169 TFs were retained and they were used as predictors $\boldsymbol{X}$ in our regression models. We obtained the observed values of $\boldsymbol{X}$ for all samples. Replicate samples from the same cell type were averaged. This resulted in a gene expression matrix $\mathbb{X}$ with 57 rows and 169 columns. Here rows correspond to 57 cell types ($n = 57$). Columns correspond to 169 TFs ($d = 169$). Each row is a realization of $\boldsymbol{X}$.

For the responses, we processed the DNase-seq data from the same 57 cell types as follows. First, the human genome was divided into 200 base pair (bp) non-overlapping windows, yielding approximately 1.65×10^7 windows. For each window and each DNase-seq sample, we calculated the DNase-seq signal by counting the number of DNase-seq reads overlapping with the window. The read count was then normalized by the library size. To do so, the window read count was divided by the total read count of the sample and then multiplied with a constant 17,002,867, which is the minimum sample read count of all samples. The normalized counts were $\log_2$ transformed after adding a pseudocount of 1. Since windows without any DNase-seq signal are unlikely to be cis-elements, we only retained windows that had non-zero read count in at least 10 cell types and had coefficient of variation no less than 1. From these retained windows, we randomly sampled 1,000 windows to serve as our testing cis-elements. For each cis-element, we extracted the normalized and log-transformed read count from all DNase-seq samples to serve as the measurements of their DNase I hypersensitivity. Replicate samples from the same cell type were averaged. This produced a matrix $\mathbb{Y}$ with 57 rows and 1,000 columns. The rows represent 57 cell types, and the columns represent 1,000 cis-elements. Each column corresponds to a response Y, and it contains the observed values of Y in all 57 cell types.

Lastly, we used $\mathbb{X}$ and $\mathbb{Y}$ to perform the analyses. Our regression model was fitted for each cis-element (i.e., each column of $\mathbb{Y}$) separately. Each regression has sample size $n = 57$ and dimension $d = 169$. An analysis of the whole dataset involves fitting 1,000 regression models.

C.2. RMRCE performance with difference choices of α_n

With regard to the real data experiment in Section 3, Figure 9 further compares the accuracy of RMRCE with α_n chosen by cross validation or set to be fixed values $1, 3, 5, 7, 9$. The performances are similar for RMRCE with difference choices of α_n. α_n chosen by cross validation only leads to marginal performance gain compared to fixed α_n. This agrees with the conclusion from the synthetic data analysis.

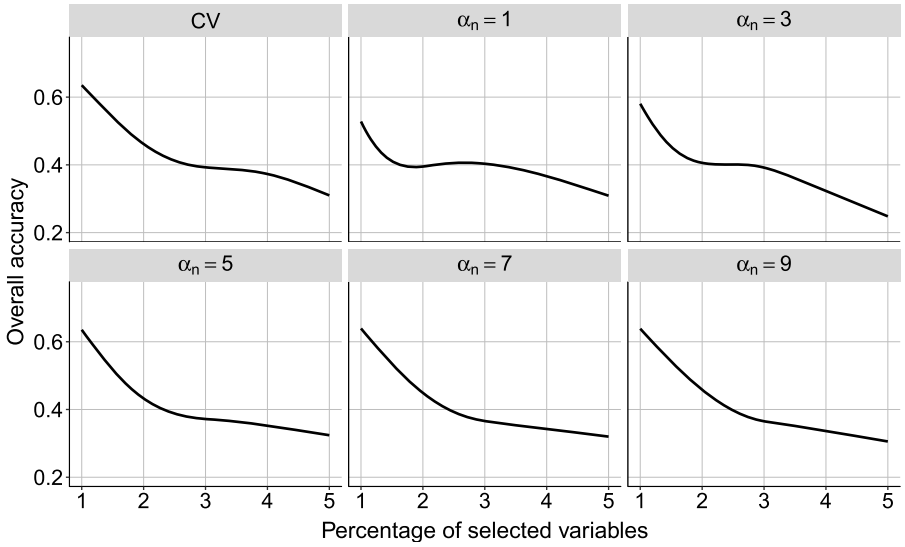


FIG 9. *Overall accuracy of RMRCE with α_n chosen by cross validation (CV) or set to be fixed values $1, 3, 5, 7, 9$. X-axis shows the averaged percentage of selected TFs out of all 169 TFs. Y-axis shows the overall accuracy.*

In real data applications, since different choices of α_n (as long as α_n is large enough) lead to fairly robust results, we recommend using a fixed $\alpha_n = 5$, although choosing the optimal α_n via a cross validation procedure is encouraged if enough time and resource for computation are available.

C.3. A model diagnostic heuristic

The monotonicity assumption is the most important and intrinsic feature of the proposed generalized regression model (1.1), and this section provides another heuristic to examine this assumption in real data applications. Figures 3 and 4 provide some empirical illustrations that Y has a monotonically increasing relationship with $\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ for two cis-elements. In this section we further discuss a model diagnostic tool to check the monotonicity assumption of our model, and we apply this tool to the real data example.

For model (1.1), our goal is to verify the assumption that the response Y has a monotonically increasing relationship with the linear term $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}^*$. However, this assumption cannot be directly verified in reality since $\boldsymbol{\beta}^*$ is unknown. Instead we develop a model diagnostic heuristic by replacing $\boldsymbol{\beta}^*$ with $\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ and checking whether Y has a monotonically increasing relationship with $\boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ in a cross-validation procedure. Specifically, we split the response and explanatory variables $(Y, \boldsymbol{X})$ into two parts with equal number of observations: the first part $(Y_1, \boldsymbol{X}_1)$ has the first half of all observations and the second part $(Y_2, \boldsymbol{X}_2)$ has the second half of all observations. We fit RMRCE and obtain $\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ on $(Y_1, \boldsymbol{X}_1)$

TABLE 8

Model diagnostic results for RMRCE in 1,000 studied real datasets. *The percentage of 1,000 adjusted p-values that are smaller than 0.05 for Spearman's rank correlation tests, or equivalently the percentage of 1,000 cis-elements that pass the model diagnostic heuristic tests. Results for different α_n and λ_n are shown.*

	α_n				
λ_n	1	3	5	7	9
0.01	0.796	0.761	0.747	0.732	0.725
0.02	0.759	0.756	0.722	0.709	0.701
0.03	0.775	0.752	0.725	0.716	0.703
0.04	0.772	0.722	0.707	0.727	0.709
0.05	0.775	0.747	0.711	0.696	0.693

and calculate the Spearman's rank correlation between Y_2 and $\boldsymbol{X}_2^\mathsf{T}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$. We use the one-sided Spearman's rank correlation test [71] to test whether the Spearman's rank correlation is significantly positive.

As a demonstration, we apply the model diagnostic heuristic to the real data example of TF prediction. For the protein-binding activity Y and gene expression level $\boldsymbol{X}$ of each of the 1,000 cis-elements, we split $(Y, \boldsymbol{X})$ into $(Y_1, \boldsymbol{X}_1)$ and $(Y_2, \boldsymbol{X}_2)$, fit RMRCE on $(Y_1, \boldsymbol{X}_1)$, and lastly perform the one-sided Spearman's rank correlation test on Y_2 and $\boldsymbol{X}_2^\mathsf{T}\widehat{\boldsymbol{\beta}}^{\mathrm{RMRCE}}$ to examine the monotonicity assumption. 1,000 p-values are obtained and adjusted for multiple testing using Bonferroni method. With an adjusted p-value smaller than 0.05, we reject the null hypothesis that there is none or a negative association. The whole procedure is repeated for different choices of RMRCE tuning parameters α_n and λ_n. Table 8 shows the percentage of the adjusted p-values that are smaller than 0.05. Most of the adjusted p-values are smaller than 0.05.

Appendix D: Proofs

This section collects the proofs of the results in the paper. Recall $Z_{ii'}(\boldsymbol{\beta}) := (\boldsymbol{X}_i - \boldsymbol{X}_{i'})^\mathsf{T}\boldsymbol{\beta}$, and $S_{ii'} := \mathbb{1}(Y_i > Y_{i'})$. In the sequel, we define $p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) := \mathbb{P}(S_{ii'} = 0 | \boldsymbol{X}_i, \boldsymbol{X}_{i'}) = \mathbb{P}(Y_i < Y_{i'} | \boldsymbol{X}_i, \boldsymbol{X}_{i'})$, and $p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) := \mathbb{P}(S_{ii'} = 1 | \boldsymbol{X}_i, \boldsymbol{X}_{i'}) = \mathbb{P}(Y_i > Y_{i'} | \boldsymbol{X}_i, \boldsymbol{X}_{i'})$.

D.1. Proof of Proposition 2.1

Proof. (i) Given the link function $D \circ \Lambda(u, v) = u + v$, we have $Y = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}^* + \epsilon$ but without requiring $\boldsymbol{X}$ and ϵ to be centered. First, by simple calculation, using the fact that $\boldsymbol{X}$ is independent with ϵ, it immediately follows

$$\frac{\mathbb{E}[(Y_1 - Y_2)(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})]}{\sqrt{\mathrm{Cov}(Y_1 - Y_2)}\sqrt{\mathrm{Cov}(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})}} = \frac{\mathbb{E}[\boldsymbol{\beta}^{*\mathsf{T}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\mathsf{T}\boldsymbol{\beta}]}{\sqrt{\mathrm{Cov}(Y_1 - Y_2)}\sqrt{\mathrm{Cov}(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})}}.$$

Noticing that $\boldsymbol{X}_1 - \boldsymbol{X}_2$ is a random vector with mean $\mathbf{0}$ and covariance $2\boldsymbol{\Sigma}$, we have

$$\operatorname*{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d}\Big\{\frac{\mathbb{E}[(Y_1-Y_2)(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}-\boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})]}{\sqrt{\mathrm{Cov}(Y_1-Y_2)}\sqrt{\mathrm{Cov}(\boldsymbol{X}_1^\mathsf{T}\boldsymbol{\beta}-\boldsymbol{X}_2^\mathsf{T}\boldsymbol{\beta})}}\Big\} = \operatorname*{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d}\frac{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}},$$

since $\mathrm{Cov}(Y_1 - Y_2)$ is a constant as a function of $\boldsymbol{\beta}$. Let $\mathbf{a} := \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*$ and $\mathbf{b} := \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2$. Using Cauchy-Schwarz inequality yields that $\max_{\mathbf{b}:\|\mathbf{b}\|_2=1}\mathbf{a}^\mathsf{T}\mathbf{b}$ achieves at $\mathbf{b} = \mathbf{a}/\|\mathbf{a}\|_2$. From the inversibility of $\boldsymbol{\Sigma}$, it then follows $\operatorname{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d}\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sqrt{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ up to a scaling.

(ii) This follows directly from the proof of Theorem 1 in [3].

Hence we complete the proof of the proposition. □

D.2. Proof of Lemma 4.1

Proof. First of all, recall $\mathcal{L}(\boldsymbol{\beta}) = -\mathbb{E}\big(S_{ii'}\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0) + (1 - S_{ii'})\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\big)$ and $\mathcal{L}_n(\boldsymbol{\beta}) = -2\sum_{i<i'}\mathbb{E}\Big\{S_{ii'}F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))+(1-S_{ii'})(1-F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})))\Big\}/\{n(n-1)\}$. By the definition of $p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)$, and $p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)$, after taking conditional expectation of the response given the covariates, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta}) = \frac{2}{n(n-1)}\sum_{i<i'}\mathbb{E}\Big\{&p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big[F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) - \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0)\big]\\ &+ p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big[1 - F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) - \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\big]\Big\}.\end{aligned}$$

According to whether $Z_{ii'}(\boldsymbol{\beta}) > 0$ or not, we rewrite the above expression as follows,

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta})\\ =&\frac{2}{n(n-1)}\sum_{i<i'}\mathbb{E}\Big\{\big[p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) - p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big]F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\\ &+ \big[p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) - p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big]\big[1 - F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big]\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0)\Big\}.\end{aligned}$$

Because of $|p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) - p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)| \le 1$, it yields

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta}) \le \frac{2}{n(n-1)}\sum_{i<i'}\mathbb{E}\Big\{&F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\\ &+ \big[1 - F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big]\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0)\Big\}. \quad \text{(D.1)}\end{aligned}$$

Following Assumption **(A1)**, we have $Z_{ii'}(\boldsymbol{\beta}) \sim N_1(0, \sigma_{\boldsymbol{\beta}}^2)$ with $\sigma_{\boldsymbol{\beta}}^2 := 2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}$ for all $1 \le i \ne i' \le n$. Hence with the expectation $\mathbb{E}$ in (D.1) taken with respect to $\mathbb{P}_{\boldsymbol{\beta}}(z) \sim N_1(0, \sigma_{\boldsymbol{\beta}}^2)$, it follows

$$\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta}) \le \frac{2}{n(n-1)}\sum_{i<i'}\Big\{\int_0^\infty\{1 - F_{ii'}(\alpha_n z)\}d\mathbb{P}_{\boldsymbol{\beta}}(z) + \int_{-\infty}^0 F_{ii'}(\alpha_n z)d\mathbb{P}_{\boldsymbol{\beta}}(z)\Big\}.$$

Notice that with Assumption **(A2)**, we have $F_{ii'}(-u) = 1 - F_{ii'}(u)$ for arbitrary $u \geq 0$, together with $\mathbb{P}_{\boldsymbol{\beta}}(z) \sim N_1(0, \sigma_{\boldsymbol{\beta}}^2)$, simple calculation immediately yields

$$\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta}) \leq \frac{4}{n(n-1)} \sum_{i<i'} \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_{\boldsymbol{\beta}}} \{1 - F_{ii'}(\alpha_n z)\} \exp\big(-\frac{z^2}{2\sigma_{\boldsymbol{\beta}}^2}\big) dz.$$

Assumption **(A2)** also assumes the existence of some absolute constants $C_1, C_2 > 0$ such that $1 - F_{ii'}(u) \leq C_1 \exp(-C_2 u)$ for any $u > 0$. Thus we further have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta}) &\leq 2 \int_0^\infty \frac{C_1}{\sqrt{2\pi}\sigma_{\boldsymbol{\beta}}} \exp\big(-\frac{z^2}{2\sigma_{\boldsymbol{\beta}}^2} - C_2 \alpha_n z\big) dz \\ &= 2C_1 e^{\sigma_{\boldsymbol{\beta}}^2 C_2^2 \alpha_n^2} (1 - \Phi(\sigma_{\boldsymbol{\beta}} C_2 \alpha_n)) \leq \frac{2C_1}{C_2 \sigma_{\boldsymbol{\beta}} \alpha_n}.\end{aligned}$$

And similarly we can prove $-(\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}_n(\boldsymbol{\beta})) \leq \frac{2C_1}{C_2 \sigma_{\boldsymbol{\beta}} \alpha_n}$. So we finally obtain

$$\sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} |\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \leq \frac{2C_1}{C_2 \alpha_n} \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}}}.$$

This ends the proof. □

D.3. Proof of Lemma 4.3

Proof. We begin by introducing some additional notation. For any function $f : \mathbb{R}^d \to \mathbb{R}$, define $\inf f := \inf_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} f(\boldsymbol{\beta})$. Given $r > 0$ and $\boldsymbol{\beta}^* \in \mathbb{R}^d$, we denote $\inf_r f := \inf_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1, \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq r} f(\boldsymbol{\beta})$. And similarly we define the corresponding versions for "$\sup f$" and "$\sup_r f$". With

$$\boldsymbol{\beta}_{r,\alpha_n}^* = \operatorname*{argmin}_{\boldsymbol{\beta}_1=1, \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq r} \mathcal{L}_n(\boldsymbol{\beta}) \quad \text{and} \quad \boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}_1=1} \mathcal{L}(\boldsymbol{\beta}),$$

we immediately have

$$\mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*) - \mathcal{L}(\boldsymbol{\beta}^*) = |-\inf \mathcal{L} + \mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*)| = |-\inf \mathcal{L} + \inf_r \mathcal{L}_n - \inf_r \mathcal{L}_n + \mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*)|.$$

By the triangular inequality, $\mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*) - \mathcal{L}(\boldsymbol{\beta}^*) \leq |-\inf \mathcal{L} + \inf_r \mathcal{L}_n| + |\mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*) - \inf_r \mathcal{L}_n|$. Since $\boldsymbol{\beta}_{r,\alpha_n}^*$ is the minimizer of $\mathcal{L}_n(\boldsymbol{\beta})$ under the restrictions $\boldsymbol{\beta}_1 = 1$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r$, we further have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*) - \mathcal{L}(\boldsymbol{\beta}^*) &\leq |-\inf_r \mathcal{L}_n + \inf \mathcal{L}| + |\mathcal{L}(\boldsymbol{\beta}_{r,\alpha_n}^*) - \mathcal{L}_n(\boldsymbol{\beta}_{r,\alpha_n}^*)| \\ &\leq |-\inf_r \mathcal{L}_n + \inf \mathcal{L}| + \sup |\mathcal{L} - \mathcal{L}_n|.\end{aligned}$$

Notice that

$$-\inf_r \mathcal{L}_n + \inf \mathcal{L} = \sup_r(-\mathcal{L}_n) - \sup(-\mathcal{L}) = \sup_r(-\mathcal{L}_n + \mathcal{L} - \mathcal{L}) - \sup(-\mathcal{L})$$

$$\leq \sup_r(|\mathcal{L}_n - \mathcal{L}| - \mathcal{L}) - \sup(-\mathcal{L})$$
$$\leq \sup_r |\mathcal{L}_n - \mathcal{L}| + \sup_r(-\mathcal{L}) - \sup(-\mathcal{L}) \leq \sup |\mathcal{L}_n - \mathcal{L}|,$$

and similarly, $\inf_r \mathcal{L}_n - \inf \mathcal{L} \leq \sup |\mathcal{L}_n - \mathcal{L}|$. They lead to $|-\inf_r \mathcal{L}_n + \inf \mathcal{L}| \leq \sup |\mathcal{L}_n - \mathcal{L}|$. Hence we have $\mathcal{L}(\boldsymbol{\beta}^*_{r,\alpha_n}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq 2 \sup |\mathcal{L} - \mathcal{L}_n|$.

Now using Assumptions **(A1)** and **(A2)**, it follows from Lemma 4.1 that

$$\mathcal{L}(\boldsymbol{\beta}^*_{r,\alpha_n}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq 2 \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} |\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \leq \frac{4C_1}{C_2\alpha_n} \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}}. \quad \text{(D.2)}$$

Given Assumptions **(A1)** and **(A3)**, Proposition 4.2 guarantees that for any given positive constants $\gamma := \{\gamma_1, \gamma_2\}$ with $\gamma_2/\gamma_1 - 1$ arbitrarily close to 0, such that for some small enough $r(\gamma) > 0$ only depending on γ, as long as $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r(\gamma)$, we have

$$\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \asymp \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2. \quad \text{(D.3)}$$

Hence when $\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2 \leq r(\gamma)$ holds, combining (D.2) and (D.3) implies

$$\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2^2 \lesssim \alpha_n^{-1} \cdot \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}}.$$

This completes the proof. □

D.4. Proof of Lemma 4.4

Proof. Without loss of generality, assume $M = 1$. First of all, let us recall

$$\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\Big\{ \big[p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) - p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big] F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)$$
$$+ \big[p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) - p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*)\big] \big[1 - F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big] \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0)\Big\}.$$

Under the monotonic transformation model (1.2), we have further derivation,

$$\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\Big\{ \big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*)) - 1\big] F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\Big\}$$
$$+ \mathbb{E}\Big\{ \big[1 - 2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))\big] \big[1 - F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big] \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0)\Big\}.$$

Due to the fact $F_\epsilon(-u) = 1 - F_\epsilon(u)$ and the property $F_{ii'}(-u) = 1 - F_{ii'}(u)$, it follows

$$\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\Big\{ \big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*)) - 1\big] F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta})) \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\Big\}$$
$$+ \mathbb{E}\Big\{ \big[2F_\epsilon(-Z_{ii'}(\boldsymbol{\beta}^*)) - 1\big] \big[F_{ii'}(-\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big] \mathbb{1}(-Z_{ii'}(\boldsymbol{\beta}) < 0)\Big\}. \quad \text{(D.4)}$$

Under Assumption **(A1)**, it yields immediately

$$\begin{pmatrix} Z_{ii'}(\boldsymbol{\beta}^*) \\ Z_{ii'}(\boldsymbol{\beta}) \end{pmatrix} \sim -\begin{pmatrix} Z_{ii'}(\boldsymbol{\beta}^*) \\ Z_{ii'}(\boldsymbol{\beta}) \end{pmatrix} \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 2\boldsymbol{\beta}^{*\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta}^* & 2\boldsymbol{\beta}^{*\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta} \\ 2\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta}^* & 2\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta} \end{pmatrix}\right),$$

which straightforwardly implies

$$\begin{aligned} &\mathbb{E}\Big\{\big[2F_\epsilon(-Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\big[F_{ii'}(-\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big]\mathbb{1}(-Z_{ii'}(\boldsymbol{\beta})<0)\Big\} \\ &\quad = \mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\big[F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\big]\mathbb{1}(Z_{ii'}(\boldsymbol{\beta})<0)\Big\}. \end{aligned}$$

Hence (D.4) can be further simplified as

$$\mathcal{L}_n(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta}) = 2\mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\mathbb{1}(Z_{ii'}(\boldsymbol{\beta})<0)\Big\}. \tag{D.5}$$

(1) If we have $\boldsymbol{\beta}^{*\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta}=0$, that is $Z_{ii'}(\boldsymbol{\beta}^*)$ is independent with $Z_{ii'}(\boldsymbol{\beta})$, we then have

$$\mathcal{L}_n(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta}) = 2\mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\Big\}\mathbb{E}\Big\{F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}))\mathbb{1}(Z_{ii'}(\boldsymbol{\beta})<0)\Big\}.$$

Since $Z_{ii'}(\boldsymbol{\beta}^*)$ has the same distribution as $-Z_{ii'}(\boldsymbol{\beta}^*)$, with

$$2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1 = 1-2F_\epsilon(-Z_{ii'}(\boldsymbol{\beta}^*)),$$

it follows

$$\mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\Big\} = \mathbb{E}\Big\{\big[2F_\epsilon(-Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\Big\} = -\mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\Big\},$$

which implies $\mathbb{E}\Big\{\big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*))-1\big]\Big\}=0$. Hence we obtain $\mathcal{L}_n(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta})=0$.

(2) With $\boldsymbol{\Sigma}=\mathbf{I}$, we have

$$\begin{pmatrix} Z_{ii'}(\boldsymbol{\beta}^*) \\ Z_{ii'}(\boldsymbol{\beta}) \end{pmatrix} \sim \mathrm{N}_2\left(\mathbf{0}, 2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \text{with } \rho = \boldsymbol{\beta}^{*\intercal}\boldsymbol{\beta}.$$

Using (D.5), with the above joint normal distribution, it follows

$$\begin{aligned} W(\rho) := \mathcal{L}_n(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta}) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{0} &\big[2F_\epsilon(u)-1\big]F_{ii'}(\alpha_n v) \\ &\frac{1}{4\pi\sqrt{1-\rho^2}}e^{-\frac{u^2+v^2-2\rho uv}{4(1-\rho^2)}}dvdu. \end{aligned}$$

After further simplification, we deduce

$$\begin{aligned} W(\rho) = 2\int_{0}^{\infty}\int_{0}^{\infty} &\big[2F_\epsilon(u)-1\big]\big[1-F_{ii'}(\alpha_n v)\big]\frac{1}{4\pi\sqrt{1-\rho^2}} \\ &\Big\{e^{-\frac{u^2+v^2+2\rho uv}{4(1-\rho^2)}} - e^{-\frac{u^2+v^2-2\rho uv}{4(1-\rho^2)}}\Big\}dvdu. \end{aligned}$$

Now suppose that the noise term ϵ follows a normal distribution with standard deviation σ. It follows that we have $F_\epsilon(u) = \Phi(u/(\sqrt{2}\sigma))$, which renders

$$W(\rho) = 2\int_0^\infty \int_0^\infty \big[2\Phi(\frac{u}{\sqrt{2}\sigma}) - 1\big]\big[1 - F_{ii'}(\alpha_n v)\big] \frac{1}{4\pi\sqrt{1-\rho^2}} \Big\{ e^{-\frac{u^2+v^2+2\rho uv}{4(1-\rho^2)}} - e^{-\frac{u^2+v^2-2\rho uv}{4(1-\rho^2)}} \Big\} dvdu.$$

For sigmoid, Gaussian CDF, and double exponential CDF approximations, we conclude that the corresponding $W(\rho)$ is an increasing function of $|\rho|$ by employing numeric integrations.

(3) With $F_\epsilon(u) = \Phi(u/(\sqrt{2}\sigma))$ and $F_{ii'}(\alpha_n v) = 1/(1 + e^{-\alpha_n v})$, we have

$$\begin{aligned}
&\mathcal{L}_n(\boldsymbol{\beta}^*) - \mathcal{L}(\boldsymbol{\beta}^*) = 2\mathbb{E}\Big\{ \big[2F_\epsilon(Z_{ii'}(\boldsymbol{\beta}^*)) - 1\big] F_{ii'}(\alpha_n Z_{ii'}(\boldsymbol{\beta}^*)) \mathbb{1}(Z_{ii'}(\boldsymbol{\beta}^*) < 0) \Big\} \\
=& 2\int_{-\infty}^0 \big[2F_\epsilon(u) - 1\big] F_{ii'}(\alpha_n u) \frac{1}{2\sqrt{\pi}} e^{-\frac{u^2}{4}} du \\
=& 2\int_0^\infty \big[1 - 2\Phi(\frac{u}{\sqrt{2}\sigma})\big] \frac{1}{1 + e^{\alpha_n u}} \frac{1}{2\sqrt{\pi}} e^{-\frac{u^2}{4}} du.
\end{aligned}$$

Based on this, numerical integrations show $\mathcal{L}_n(\boldsymbol{\beta}^*) - \mathcal{L}(\boldsymbol{\beta}^*) \asymp \alpha_n^{-2}$. And together with the result in Item (2), we have for any $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$, $|\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \lesssim \alpha_n^{-2}$. This completes the proof. □

D.5. Proof of Proposition 4.5

Proof. We aim to validate the following results for the i.i.d. noise terms $\{\epsilon_i, i = 1, 2, \cdots, n\}$ with Gaussian or Cauchy distribution, i.e.,

$$\int_{-\infty}^0 f_\epsilon(x) \exp(-x^2/(2b_n^2)) dx = C b_n (1 + o(1)) \quad \text{as} \quad \mathrm{b_n} \to 0,$$

where $f_\epsilon(\cdot)$ represents the PDF of $\epsilon_2 - \epsilon_1$.

For the noise term $\epsilon_i \sim \mathrm{N}(\mu, \sigma^2)$, we have $f_\epsilon(x) = e^{-x^2/(4\sigma^2)}/(2\sqrt{\pi}\sigma)$. With bounded σ^2, it yields

$$\int_{-\infty}^0 f_\epsilon(x) e^{-\frac{x^2}{2b_n^2}} dx = \frac{1}{\sqrt{2}\sigma\sqrt{\frac{1}{2\sigma^2} + \frac{1}{b_n^2}}} \Phi(0) = \frac{b_n}{\sqrt{b_n^2 + 2\sigma^2}} = \frac{b_n(1 + o(1))}{\sqrt{2\sigma^2}},$$

as $b_n \to 0$, where $\Phi(\cdot)$ is the CDF of standard normal distribution. For the noise term $\epsilon_i \sim \mathrm{Cauchy}(\mu, \gamma)$, it is easy to see $f_\epsilon(x) = 1/\{2\pi\gamma(1 + (x/(2\gamma))^2)\}$. By simple calculation, we deduce

$$\int_{-\infty}^0 f_\epsilon(x) e^{-\frac{x^2}{2b_n^2}} dx = \frac{1}{\pi} \int_{-\infty}^0 \frac{1}{(1 + x^2)} e^{-\frac{x^2}{2(b_n/(2\gamma))^2}} dx$$

$$= \frac{1}{\pi}\frac{\pi}{2} e^{\frac{1}{2(b_n/(2\gamma))^2}} [1 - \Phi(\sqrt{2}/\sqrt{2(b_n/(2\gamma))^2})]$$
$$= \frac{1}{2} e^{2\gamma^2/b_n^2} [1 - \Phi(2\gamma/b_n)]. \tag{D.6}$$

With $\phi(\cdot)$ the PDF of standard normal distribution, using the fact

$$\phi(x)/(x + 1/x) < 1 - \Phi(x) < \phi(x)/x \text{ for any } x > 0,$$

we have

$$\frac{1}{\sqrt{2\pi}} e^{-2\gamma^2/b_n^2}/[2\gamma/b_n + b_n/(2\gamma)] < 1 - \Phi(2\gamma/b_n) < \frac{1}{\sqrt{2\pi}} e^{-2\gamma^2/b_n^2}/[2\gamma/b_n]. \tag{D.7}$$

Combining (D.6) and (D.7) leads to

$$\frac{\gamma b_n}{\sqrt{2\pi}(4\gamma^2 + b_n^2)} < \int_{-\infty}^{0} f_\epsilon(x) e^{-\frac{x^2}{2b_n^2}} dx < \frac{b_n}{4\sqrt{2\pi}\gamma}.$$

Having bounded γ, we immediately have $\int_{-\infty}^{0} f_\epsilon(x) e^{-\frac{x^2}{2b_n^2}} dx$ $= b_n(1 + o(1))/(4\sqrt{2\pi}\gamma)$ as $b_n \to 0$. Hence we complete the proof of the proposition for both Gaussian and Cauchy distributed noises. □

D.6. Proof of Lemma 4.6

Proof. Recall $\mathcal{L}(\boldsymbol{\beta}) = -\mathbb{E}\big(S_{ii'}\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) > 0) + (1 - S_{ii'})\mathbb{1}(Z_{ii'}(\boldsymbol{\beta}) < 0)\big)$. Under the monotonic transformation model (1.2), i.e., $Y = G(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}^* + \epsilon)$, we have

$$p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) = \mathbb{P}(Y_i < Y_{i'}|\boldsymbol{X}_i, \boldsymbol{X}_{i'}) = \mathbb{P}\big\{G(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}^* + \epsilon_i) < G(\boldsymbol{X}_{i'}^\mathsf{T}\boldsymbol{\beta}^* + \epsilon_{i'})\big|\boldsymbol{X}_i, \boldsymbol{X}_{i'}\big\},$$

and

$$p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) = \mathbb{P}(Y_i > Y_{i'}|\boldsymbol{X}_i, \boldsymbol{X}_{i'}) = \mathbb{P}\big\{G(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}^* + \epsilon_i) > G(\boldsymbol{X}_{i'}^\mathsf{T}\boldsymbol{\beta}^* + \epsilon_{i'})\big|\boldsymbol{X}_i, \boldsymbol{X}_{i'}\big\}.$$

By the monotonicity of $G(\cdot)$, we can write

$$p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\} := p_0(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) = \mathbb{P}\big\{Z_{ii'}(\boldsymbol{\beta}^*) < \epsilon_{i'} - \epsilon_i\big|\boldsymbol{X}_i, \boldsymbol{X}_{i'}\big\}, \tag{D.8}$$
$$p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\} := p_1(\boldsymbol{X}_i, \boldsymbol{X}_{i'}; \boldsymbol{\beta}^*) = \mathbb{P}\big\{Z_{ii'}(\boldsymbol{\beta}^*) > \epsilon_{i'} - \epsilon_i\big|\boldsymbol{X}_i, \boldsymbol{X}_{i'}\big\}. \tag{D.9}$$

Notice in $\mathcal{L}(\boldsymbol{\beta})$ we have both of the two terms $Z_{ii'}(\boldsymbol{\beta}^*)$ and $Z_{ii'}(\boldsymbol{\beta})$ involved. Using Assumption **(A1)**, we immediately have the following joint distribution for $(Z_{ii'}(\boldsymbol{\beta}^*), Z_{ii'}(\boldsymbol{\beta}))^\mathsf{T}$:

$$\begin{pmatrix} Z_{ii'}(\boldsymbol{\beta}^*) \\ Z_{ii'}(\boldsymbol{\beta}) \end{pmatrix} \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 2\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^* & 2\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} \\ 2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}^* & 2\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta} \end{pmatrix}\right), \tag{D.10}$$

which is independent of the indices i and i'. For convenience, let's define

$$\rho := \frac{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sqrt{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}}. \tag{D.11}$$

After taking conditional expectations with respect to the response given the covariates, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta}^*)=&\mathbb{E}\left\{p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\left[\mathbb{1}\{Z_{ii'}(\boldsymbol{\beta}^*)>0\}-\mathbb{1}\{Z_{ii'}(\boldsymbol{\beta})>0\}\right]\right\}\\&+\mathbb{E}\left\{p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\left[\mathbb{1}\{Z_{ii'}(\boldsymbol{\beta}^*)<0\}-\mathbb{1}\{Z_{ii'}(\boldsymbol{\beta})<0\}\right]\right\}.\end{aligned}$$

In order to further simplify the above expression, we define the following regions

$$\begin{aligned}&\mathcal{A}_1=\{Z_{ii'}(\boldsymbol{\beta}^*)>0;Z_{ii'}(\boldsymbol{\beta})>0\};\ \mathcal{A}_2=\{Z_{ii'}(\boldsymbol{\beta}^*)>0;Z_{ii'}(\boldsymbol{\beta})<0\};\\&\mathcal{A}_3=\{Z_{ii'}(\boldsymbol{\beta}^*)<0;Z_{ii'}(\boldsymbol{\beta})>0\};\ \mathcal{A}_4=\{Z_{ii'}(\boldsymbol{\beta}^*)<0;Z_{ii'}(\boldsymbol{\beta})<0\}.\end{aligned}$$

Simple calculation then leads to

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta})-\mathcal{L}(\boldsymbol{\beta}^*)\\=&\mathbb{E}\left\{\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)+\left[p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_3)\right\}\\=&\mathbb{E}\left\{\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)\right\}+\mathbb{E}\left\{\left[p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_3)\right\}.\end{aligned}$$

With the monotonic transformation model (1.2), using monotonicity and the assumption of i.i.d. noise terms $\{\epsilon_i, i=1,2,\cdots,n\}$ assumed in Assumption **(A1)**, we have

$$\begin{aligned}&\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)\geq 0\\\text{and}\quad&\left[p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_3)\geq 0.\end{aligned}\tag{D.12}$$

According to the simplified notation in (D.8) and (D.9) under Model (1.2), with the monotonicity of G, it follows

$$\begin{aligned}&p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\\=&\mathbb{P}\big\{Z_{ii'}(\boldsymbol{\beta}^*)>\epsilon_{i'}-\epsilon_i\big|\boldsymbol{X}_i,\boldsymbol{X}_{i'}\big\}-\mathbb{P}\big\{Z_{ii'}(\boldsymbol{\beta}^*)<\epsilon_{i'}-\epsilon_i\big|\boldsymbol{X}_i,\boldsymbol{X}_{i'}\big\}\\=&1-2\mathbb{P}\big\{Z_{ii'}(\boldsymbol{\beta}^*)<\epsilon_{i'}-\epsilon_i\big|\boldsymbol{X}_i,\boldsymbol{X}_{i'}\big\}=1-2F_\epsilon\big\{-Z_{ii'}(\boldsymbol{\beta}^*)\big\},\end{aligned}$$

where F_ϵ is the CDF for $\epsilon_i-\epsilon_{i'}$ with $i\neq i'$. Recall for any $p>0$ and any random variable $U\geq 0$,

$$\mathbb{E}U^p=\int_0^\infty pu^{p-1}\mathbb{P}(U>u)du.$$

With (D.12), applying the above formula to the random variable

$$U=\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)$$

gives us that

$$\begin{aligned}&\mathbb{E}\left\{\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}-p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)\right\}\\=&\int_0^\infty\mathbb{P}\left\{1-2F_\epsilon\big\{-Z_{ii'}(\boldsymbol{\beta}^*)\big\}>a,\mathcal{A}_2\right\}da\\=&\int_0^\infty\mathbb{P}\left\{-Z_{ii'}(\boldsymbol{\beta}^*)<F_\epsilon^{-1}\big(\frac{1-a}{2}\big),-Z_{ii'}(\boldsymbol{\beta})>0\right\}da.\end{aligned}\tag{D.13}$$

By simple calculation, with $\sigma_1^2 := \sigma_{\boldsymbol{\beta}^*}^2 = 2\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*$ and $\sigma_2^2 := \sigma_{\boldsymbol{\beta}}^2 = 2\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}$, according to the joint distribution specified in (D.10), we have

$$\begin{aligned} H(x;\rho) &:= \mathbb{P}\left\{-Z_{ii'}(\boldsymbol{\beta}^*) < x, -Z_{ii'}(\boldsymbol{\beta}) > 0\right\} \\ &= \int_{-\infty}^{x}\int_{-\infty}^{0} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{u^2}{\sigma_1^2}+\frac{v^2}{\sigma_2^2}+2\rho\frac{uv}{\sigma_1\sigma_2}\right]\right\} dudv \\ &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}}\int_{-\infty}^{x} e^{-\frac{u^2}{2\sigma_1^2}}\left[\int_{-\infty}^{0}\exp\left\{-\frac{(v+\rho u/\sigma_1)^2}{2(1-\rho^2)}\right\}dv\right]du \\ &= \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{u^2}{2\sigma_1^2}}\Phi(\frac{\rho u}{\sigma_1\sqrt{1-\rho^2}})du, \end{aligned} \tag{D.14}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Combining (D.13) and (D.14), we deduce

$$\begin{aligned} K(\rho) := \mathbb{E}\left\{\left[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\} - p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\right]\mathbb{1}(\mathcal{A}_2)\right\} &= \int_0^{\infty} H\left\{F_\epsilon^{-1}(\frac{1-a}{2});\rho\right\}da \\ &= 2\int_{-\infty}^{F_\epsilon^{-1}(1/2)} H(x;\rho)f_\epsilon(x)dx. \end{aligned} \tag{D.15}$$

Via exchange of taking derivative and integral, we have the first derivative of the above function $K(\rho)$ is of the form

$$K'(\rho) = 2\int_{-\infty}^{F_\epsilon^{-1}(1/2)} f_\epsilon(x)\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{u^2}{2\sigma_1^2}}\phi(\frac{\rho u}{\sigma_1\sqrt{1-\rho^2}})\frac{u}{\sigma_1(1-\rho^2)^{3/2}}dudx,$$

where $\phi(\cdot)$ represents the PDF for the standard normal distribution. With simple calculation, we can simplify $K'(\rho)$ further as follows:

$$\begin{aligned} K'(\rho) &= \frac{1}{\pi\sigma_1^2(1-\rho^2)^{3/2}}\int_{-\infty}^{F_\epsilon^{-1}(1/2)} f_\epsilon(x)\int_{-\infty}^{x} ue^{-\frac{u^2}{2\sigma_1^2(1-\rho^2)}}dudx \\ &= \frac{1}{\pi\sigma_1^2(1-\rho^2)^{3/2}}\int_{-\infty}^{F_\epsilon^{-1}(1/2)} f_\epsilon(x)\left\{-\sigma_1^2(1-\rho^2)e^{-\frac{x^2}{2\sigma_1^2(1-\rho^2)}}\right\}dx \\ &= \frac{-1}{\pi\sqrt{1-\rho^2}}\int_{-\infty}^{F_\epsilon^{-1}(1/2)} f_\epsilon(x)e^{-\frac{x^2}{2\sigma_1^2(1-\rho^2)}}dx. \end{aligned} \tag{D.16}$$

Due to $F_\epsilon^{-1}(1/2) = 0$ since F_ϵ is the CDF for $\epsilon_2 - \epsilon_1$, it yields immediately from (D.16) that

$$K'(\rho) = \frac{-1}{\pi\sqrt{1-\rho^2}}\int_{-\infty}^{0} f_\epsilon(x)e^{-\frac{x^2}{2\sigma_1^2(1-\rho^2)}}dx.$$

By Assumption **(A3')**, it follows

$$\int_{-\infty}^{0} f_\epsilon(x)e^{-\frac{x^2}{2\sigma_1^2(1-\rho^2)}}dx = C\sigma_1\sqrt{1-\rho^2}(1+o(1)) \quad \text{as} \quad \rho\to 1.$$

It yields

$$K'(\rho) = -C(\sigma_1/\pi)(1+o(1)) \quad \text{as} \quad \rho \to 1.$$

When $\rho = 1$, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, which leads to $K(1) = 0$. Applying Taylor expansion at $\rho = 1$ gives us

$$K(\rho) = K(1) + K'(\rho)(\rho-1) + o(\rho-1) = K'(\rho)(\rho-1) + o(\rho-1), \text{as } \rho \to 1.$$

Hence by the definition of $K(\rho)$, we have

$$\mathbb{E}\big\{\big[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\} - p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\big]\mathbb{1}(\mathcal{A}_2)\big\} = K(\rho) = C(\sigma_1/\pi)(1-\rho)(1+o(1)), \text{as } \rho \to 1.$$

Similarly, by symmetry, we conclude

$$\mathbb{E}\big\{\big[p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\} - p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\big]\mathbb{1}(\mathcal{A}_3)\big\} = C(\sigma_1/\pi)(1-\rho)(1+o(1)), \text{as } \rho \to 1.$$

In summary, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \\
=&\mathbb{E}\big\{\big[p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\} - p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\}\big]\mathbb{1}(\mathcal{A}_2)\big\} + \mathbb{E}\big\{\big[p_0\{Z_{ii'}(\boldsymbol{\beta}^*)\} - p_1\{Z_{ii'}(\boldsymbol{\beta}^*)\}\big]\mathbb{1}(\mathcal{A}_3)\big\} \\
=&c(1-\rho)(1+o(1)), \text{as } \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \to 0,
\end{aligned}$$

for some absolute constant $c > 0$. We can then pick positive constant set $\gamma := \{\gamma_1, \gamma_2\}$ with $\gamma_2/\gamma_1 - 1$ arbitrarily close to 0, such that for some small enough $r(\gamma) > 0$ only depending on γ, as long as $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le r(\gamma)$, we have

$$\gamma_1 \cdot \Big(1 - \frac{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sqrt{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}}\Big) \le \mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) \le \gamma_2 \cdot \Big(1 - \frac{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}{\sqrt{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}}\Big),$$

due to the definition of ρ in (D.11). □

D.7. Proof of Theorem 4.8

Proof. The proof is split to three steps.

(1) In the first step, we show

$$\widehat{\boldsymbol{\Delta}} := \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \boldsymbol{\theta}^*). \tag{D.17}$$

By Assumption **(B3)** and simple algebra, for any $\boldsymbol{\Delta} \in \mathbb{R}^d$, we have

$$P(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - P(\boldsymbol{\theta}^*) \ge P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}}) - P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}}) - 2P(\boldsymbol{\theta}^*_{\mathcal{M}^{\perp}}). \tag{D.18}$$

In addition, due to the convex differentiability of $L_n(\cdot)$ (Assumption **(B2)**), we have

$$L_n(\boldsymbol{\theta}^* + \widehat{\boldsymbol{\Delta}}) - L_n(\boldsymbol{\theta}^*) \ge -|\langle \nabla L_n(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\Delta}} \rangle|.$$

Holder's inequality then yields

$$L_n(\boldsymbol{\theta}^* + \widehat{\boldsymbol{\Delta}}) - L_n(\boldsymbol{\theta}^*) \ge -P^*(\nabla L_n(\boldsymbol{\theta}^*)) P(\widehat{\boldsymbol{\Delta}}).$$

Using Assumption **(B5)**, we further have

$$L_n(\boldsymbol{\theta}^* + \widehat{\boldsymbol{\Delta}}) - L_n(\boldsymbol{\theta}^*) \geq -\frac{\lambda_n}{2}(P(\widehat{\boldsymbol{\Delta}}_{\overline{\mathcal{M}}}) + P(\widehat{\boldsymbol{\Delta}}_{\overline{\mathcal{M}}^\perp})). \tag{D.19}$$

Combining (D.18) and (D.19), and using the fact $\widehat{\boldsymbol{\theta}}$ minimizes $L_n(\boldsymbol{\theta}) + \lambda_n P(\boldsymbol{\theta})$, we have

$$0 \geq \frac{\lambda_n}{2}\Big(P(\widehat{\boldsymbol{\Delta}}_{\overline{\mathcal{M}}^\perp}) - 3P(\widehat{\boldsymbol{\Delta}}_{\overline{\mathcal{M}}}) - 4P(\boldsymbol{\theta}^*_{\mathcal{M}^\perp})\Big).$$

This then proves (D.17).

(2). Let's define

$$\mathcal{F}(\boldsymbol{\Delta}) := L_n(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) + \lambda_n P(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - L_n(\boldsymbol{\theta}^*) - \lambda_n P(\boldsymbol{\theta}^*).$$

In the second step, we proceed to prove the following assertion: if for all $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \boldsymbol{\theta}^*) \cap \{\|\boldsymbol{\Delta}\|_2 = \gamma\}$ we have $\mathcal{F}(\boldsymbol{\Delta}) > 0$, then $\|\widehat{\boldsymbol{\Delta}}\|_2 \leq \gamma$. To this end, let's assume $\|\widehat{\boldsymbol{\Delta}}\|_2 > \gamma$. Then because $\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \boldsymbol{\theta}^*)$ is star-shaped (by Assumption **(B1)**), we can always find $t^* \in (0, 1)$ such that

$$t^*\widehat{\boldsymbol{\Delta}} \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \boldsymbol{\theta}^*) \cap \{\|\boldsymbol{\Delta}\|_2 = \gamma\}.$$

However, by Assumption **(B2)** we have

$$\mathcal{F}(t^*\widehat{\boldsymbol{\Delta}}) \leq t^*\mathcal{F}(\widehat{\boldsymbol{\Delta}}) \leq 0.$$

Therefore, we have a contradiction, and accordingly $\|\widehat{\boldsymbol{\Delta}}\|_2 \leq \gamma$.

(3). In the end, let's prove the main theorem. For all $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \boldsymbol{\theta}^*) \cap \{\|\boldsymbol{\Delta}\|_2 = \gamma\}$, using Assumption **(B4)**, we have

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{\Delta}) = L_n(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - L_n(\boldsymbol{\theta}^*) + \lambda_n(P(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - P(\boldsymbol{\theta}^*)) \\
\geq& -|\langle \nabla L_n(\boldsymbol{\theta}^*), \boldsymbol{\Delta}\rangle| + \kappa_L\|\boldsymbol{\Delta}\|_2^2 - \delta_L\|\boldsymbol{\Delta}\|_2 - \tau_L^2(\boldsymbol{\theta}^*) + \lambda_n(P(\boldsymbol{\theta}^* + \boldsymbol{\Delta}) - P(\boldsymbol{\theta}^*)) \\
\geq& -\frac{\lambda_n}{2}P(\boldsymbol{\Delta}) + \kappa_L\|\boldsymbol{\Delta}\|_2^2 - \delta_L\|\boldsymbol{\Delta}\|_2 - \tau_L^2(\boldsymbol{\theta}^*) + \lambda_n\Big(P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}) - P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}}) - 2P(\boldsymbol{\theta}^*_{\mathcal{M}^\perp})\Big) \\
\geq& -\frac{3}{2}\lambda_n P(\boldsymbol{\Delta}_{\overline{\mathcal{M}}}) - 2\lambda_n P(\boldsymbol{\theta}^*_{\mathcal{M}^\perp}) + \kappa_L\|\boldsymbol{\Delta}\|_2^2 - \delta_L\|\boldsymbol{\Delta}\|_2 - \tau_L^2(\boldsymbol{\theta}^*).
\end{aligned}$$

Finally, using Assumption **(B5)**, we derive

$$\mathcal{F}(\boldsymbol{\Delta}) \geq -(\frac{3}{2}\lambda_n\Psi(\overline{\mathcal{M}}) + \delta_L)\|\boldsymbol{\Delta}\|_2 + \kappa_L\|\boldsymbol{\Delta}\|_2^2 - \tau_L^2(\boldsymbol{\theta}^*) - 2\lambda_n P(\boldsymbol{\theta}^*_{\mathcal{M}^\perp}).$$

Hence, by picking

$$\gamma^2 = (2\lambda_n\Psi(\overline{\mathcal{M}}) + \delta_L)^2/\kappa_L^2 + 2(\tau_L^2(\boldsymbol{\theta}^*) + 2\lambda_n P(\boldsymbol{\theta}^*_{\mathcal{M}^\perp}))/\kappa_L,$$

we have, for all $\boldsymbol{\Delta} \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \boldsymbol{\theta}^*) \cap \{\|\boldsymbol{\Delta}\|_2 = \gamma\}$,

$$\mathcal{F}(\boldsymbol{\Delta}) \geq 0.$$

This completes the proof. □

D.8. Proof of Lemma 4.10

Proof. In the following we only consider the constrainted version of $\widehat{\mathcal{L}}_n$ that takes value infinity outside of a small ball of $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$. First, with Assumptions **(A1)** and **(A2)**, Lemma 4.1 gives us

$$\sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} |\mathcal{L}_n(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta})| \leq \frac{2C_1}{C_2\alpha_n} \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}_1=1} \frac{1}{\sqrt{2\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}}. \tag{D.20}$$

Secondly, using Assumptions **(A1)** and **(A3)**, Equation (9) in [22] and Proposition 4.2 implies that for $\boldsymbol{\Delta}$ small enough with $\gamma_2/\gamma_1 - 1$ close to 0,

$$\begin{aligned} &\gamma_1\lambda_{\min}(\boldsymbol{\Gamma})\|\boldsymbol{\Delta}\|_2^2 \leq \mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*) \leq \gamma_2\lambda_{\max}(\boldsymbol{\Gamma})\|\boldsymbol{\Delta}\|_2^2 \\ &\text{and} \quad \mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*) = \boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta}(1/4 + o(1)). \end{aligned} \tag{D.21}$$

Thirdly, given Assumptions **(A1)**-**(A3)**, Lemma 4.3 shows, under further Assumption **(A6)**,

$$\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2 \lesssim \alpha_n^{-1/2}. \tag{D.22}$$

Combining (D.20), (D.21), and (D.22) as well as Assumption **(A6)** yields

$$\mathcal{L}_n(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*) \geq C_1'\|\boldsymbol{\Delta}\|_2^2 - C_2'/\alpha_n,$$

and

$$\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \geq C_3'\|\boldsymbol{\Delta}\|_2^2 - C_4'/\alpha_n - C_5'\alpha_n^{-1/2}\|\boldsymbol{\Delta}\|_2.$$

In fact, the first one is trivial and the second one can be derived in detail as follows. First note

$$\begin{aligned} &\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \\ =&\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) + \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \\ &+ \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) - \mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}), \end{aligned}$$

where the first two terms and the last two terms can be lower bounded through (D.20),

$$\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) + \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) - \mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \geq -C_4'/\alpha_n.$$

Hence it immediately follows

$$\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \geq -C_4'/\alpha_n + \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}). \tag{D.23}$$

Notice that by adding and subtracting same terms, we have

$$\mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) = \mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}^*) - (\mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) - \mathcal{L}(\boldsymbol{\beta}^*)). \tag{D.24}$$

In (D.24), following from (D.21), we have

$$\mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta})-\mathcal{L}(\boldsymbol{\beta}^*)=0.25(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta}-\boldsymbol{\beta}^*)^{\mathsf{T}}\boldsymbol{\Gamma}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta}-\boldsymbol{\beta}^*)(1+o(1)),$$
$$\mathcal{L}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})-\mathcal{L}(\boldsymbol{\beta}^*)=0.25(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}-\boldsymbol{\beta}^*)^{\mathsf{T}}\boldsymbol{\Gamma}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}-\boldsymbol{\beta}^*)(1+o(1)). \quad \text{(D.25)}$$

Combining (D.23), (D.24), and (D.25) implies

$$\begin{aligned}&\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta})-\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})\\ \geq&-C_4'/\alpha_n+0.25(\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta}+2\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\Gamma}(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}-\boldsymbol{\beta}^*))(1+o(1)).\end{aligned}$$

Further with Cauchy inequality, (D.22), and Assumption **(A3)**, it follows

$$\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta})-\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})\geq-C_4'/\alpha_n-C_5'\alpha_n^{-1/2}\|\boldsymbol{\Delta}\|_2+C_3'\|\boldsymbol{\Delta}\|_2^2.$$

By the definition of κ_n, it then follows

$$\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta})-\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})\geq C_3'\|\boldsymbol{\Delta}\|_2^2-C_4'/\alpha_n-C_5'\alpha_n^{-1/2}\|\boldsymbol{\Delta}\|_2-O_P(\kappa_n),$$

and accordingly

$$\begin{aligned}\delta\widehat{\mathcal{L}}_n(\boldsymbol{\Delta},\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})&:=\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}+\boldsymbol{\Delta})-\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})-\langle\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}),\boldsymbol{\Delta}\rangle\\ &\geq C_3'\|\boldsymbol{\Delta}\|_2^2-C_4'/\alpha_n-O_P(\kappa_n)-C_5'\alpha_n^{-1/2}\|\boldsymbol{\Delta}\|_2-\langle\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}),\boldsymbol{\Delta}\rangle.\end{aligned}$$

We then determine the value of $\langle\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}),\boldsymbol{\Delta}\rangle$. By simple algebra, we have

$$\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta})=-\frac{2}{n(n-1)}\sum_{i<i'}\widetilde{S}_{ii'}\alpha_n\widetilde{\boldsymbol{X}}_{ii'}F_{ii'}'\{\widetilde{S}_{ii'}\alpha_nZ_{ii'}(\boldsymbol{\beta})\}.$$

Now note $\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta})$ is a U-statistic of order two, written as

$$\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}):=-\frac{2}{n(n-1)}\sum_{i<i'}h(\{\boldsymbol{X}_i,\epsilon_i\},\{\boldsymbol{X}_{i'},\epsilon_{i'}\}),$$

with $h(\{\boldsymbol{X}_i,\epsilon_i\},\{\boldsymbol{X}_{i'},\epsilon_{i'}\})=\widetilde{S}_{ii'}\alpha_n\widetilde{\boldsymbol{X}}_{ii'}F_{ii'}'\{\widetilde{S}_{ii'}\alpha_nZ_{ii'}(\boldsymbol{\beta})\}$. In addition, let $\|\cdot\|_{\psi_2}$ be the subgaussian norm defined in [72]:

$$\|X\|_{\psi_2}:=\sup_{p\geq1}\frac{1}{\sqrt{p}}\Big(\mathbb{E}|X|^p\Big)^{1/p}.$$

Combined with Assumption **(A4)** that $\sup_{u\in\mathbb{R}}|F_{ii'}'(u)|\leq C_3$, we have, for arbitrary $j\in\{1,\ldots,d\}$,

$$\Big\|\widetilde{S}_{ii'}\alpha_n[\widetilde{\boldsymbol{X}}_{ii'}]_jF_{ii'}'\{\widetilde{S}_{ii'}\alpha_nZ_{ii'}(\boldsymbol{\beta})\}\Big\|_{\psi_2}\leq C_3\alpha_n\Big\|[\widetilde{\boldsymbol{X}}_{ii'}]_j\Big\|_{\psi_2}$$

is upper bounded by an absolute constant. Therefore, for any pair (i,i'), we have

$$h(\{\boldsymbol{X}_i,\epsilon_i\},\{\boldsymbol{X}_{i'},\epsilon_{i'}\})$$

is subgaussian. Moreover, it is easy to see that

$$\mathbb{E}\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) = \nabla\mathbb{E}\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) = \nabla\mathcal{L}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) = 0.$$

Therefore, $\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})$ is a U-statistic of centered subgaussian distributed elements. Employing the standard Hoeffding's decoupling technique and Bonferroni's adjustment then yields

$$\|\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})\|_\infty \overset{\mathbb{P}}{\lesssim} \alpha_n\sqrt{\log d/n}.$$

Cauchy inequality then gives us

$$|\langle\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}), \boldsymbol{\Delta}\rangle| \le \|\nabla\widehat{\mathcal{L}}_n(\boldsymbol{\beta}^*_{r(\gamma),\alpha_n})\|_\infty\|\boldsymbol{\Delta}\|_1,$$

which yields

$$\delta\widehat{\mathcal{L}}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}) \overset{\mathbb{P}}{\gtrsim} C_3'\|\boldsymbol{\Delta}\|_2^2 - C_4'/\alpha_n - C_5'\alpha_n^{-1/2}\|\boldsymbol{\Delta}\|_2 - O_P(\kappa_n) - C_6'\alpha_n\sqrt{\log d/n}\|\boldsymbol{\Delta}\|_1.$$

Then under Assumption **(A5)** and Assumption **(A0)** that for some α_n large enough,

$$\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_0 \le s_n,$$

we have, letting $\widehat{\boldsymbol{\Delta}} := \widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$,

$$\|\widehat{\boldsymbol{\Delta}}\|_1 \le 4\|\widehat{\boldsymbol{\Delta}}_S\|_1 \le 4\sqrt{s_n}\|\widehat{\boldsymbol{\Delta}}\|_2.$$

And by Theorem 4.8, setting $P(\boldsymbol{\theta}) = \sum_{j=2}^d |\boldsymbol{\theta}_j|$ and $\mathcal{A} := \{\boldsymbol{\theta} \in \mathbb{R}^d : \boldsymbol{\theta}_1 = 1, \|\boldsymbol{\theta} - \boldsymbol{\beta}^*\|_2 \le r\}$, we have, when $\lambda_n \gtrsim \alpha_n\sqrt{\log d/n}$,

$$\|\widehat{\boldsymbol{\Delta}}\|_2^2 \overset{\mathbb{P}}{\lesssim} s_n\lambda_n^2 + \alpha_n^{-1} + \alpha_n^2 s_n\log d/n + \kappa_n,$$

which implies, when $\lambda_n \asymp \alpha_n\sqrt{\log d/n}$,

$$\|\widehat{\boldsymbol{\Delta}}\|_2^2 \overset{\mathbb{P}}{\lesssim} \alpha_n^2 s_n\log d/n + \alpha_n^{-1} + \kappa_n.$$

This completes the proof. □

D.9. Proof of Theorem 4.11

Proof. Picking $\alpha_n \asymp (n/(s_n\log d))^{1/3}$, Lemma 4.10 then yields

$$\|\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_2 \overset{\mathbb{P}}{\lesssim} \left(\frac{s_n\log d}{n}\right)^{1/6} + \kappa_n^{1/2}. \tag{D.26}$$

Due to Lemma 4.3, we have $\|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2^2 \overset{\mathbb{P}}{\lesssim} \alpha_n^{-1} \asymp (n/(s_n\log d))^{-1/3}$, which together with (D.26) leads to

$$\|\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2 \le \|\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*_{r(\gamma),\alpha_n}\|_2 + \|\boldsymbol{\beta}^*_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2 \overset{\mathbb{P}}{\lesssim} \Big(\frac{s_n \log d}{n}\Big)^{1/6} + \kappa_n^{1/2}.$$

Thus as long as $s_n \log d/n \to 0$ and $\kappa_n \to 0$, we have $\|\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n} - \boldsymbol{\beta}^*\|_2 \overset{\mathbb{P}}{\to} 0$.

Finally, noticing that for n large enough, by Lemma 4.3 and the above result, $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$ and $\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n}$ are both within the ball of $\{\boldsymbol{\beta} : \beta_1 = 1, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le r(\gamma)\}$. Since $r(\gamma)$ only depends on γ, it could be picked as an absolute constant by fixing γ_1, γ_2, with $\gamma_2/\gamma_1 = 1.01$ for example. Then we have $\boldsymbol{\beta}^*_{r(\gamma),\alpha_n}$ and $\widehat{\boldsymbol{\beta}}_{r(\gamma),\alpha_n}$ are indeed local minima for n large enough. This completes the proof. □

Acknowledgements

We thank the Editor, AE, and two anonymous referees for their helpful comments. We are grateful to Dr. Peter Radchenko and Dr. Haileab Hilafu for providing codes to implement their methods.

The work of Fang Han is supported by NSF grant DMS-1712536 and a UW faculty start-up grant. The work of Hongkai Ji and Zhicheng Ji is supported by NIH grants R01HG006841 and R01HG006282. The work of Honglang Wang is supported by an IUPUI faculty start-up grant.

References

[1] Peter J Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2011. MR2488795

[2] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric Regression*. Cambridge University Press, 2003. MR1998720

[3] Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2):303–316, 1987. MR0903188

[4] Peter J Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[5] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.

[6] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[7] Christopher Cavanagh and Robert P Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–381, 1998. MR1630210

[8] Joel L Horowitz. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64(1):103–137, 1996. MR1366143

[9] Jianming Ye and Naihua Duan. Nonparametric $n^{-1/2}$-consistent estimation for the general transformation models. *The Annals of Statistics*, 25(6):2682–2717, 1997. MR1604440

[10] Songnian Chen. Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697, 2002. MR1929984

[11] Peng-Jie Dai, Qing-Zhao Zhang, and Zhi-Hua Sun. Variable selection of generalized regression models based on maximum rank correlation. *Acta Mathematicae Applicatae Sinica, English Series*, 30(3):833–844, 2014. MR3285979

[12] Xingjie Shi, Jin Liu, Jian Huang, Yong Zhou, Yang Xie, and Shuangge Ma. A penalized robust method for identifying gene–environment interactions. *Genetic Epidemiology*, 38(3):220–230, 2014.

[13] Hyungtaik Ahn, Hidehiko Ichimura, and James L Powell. Simple estimators for monotone index models. Technical report, Department of Economics, UC Berkeley, 1996.

[14] Hisatoshi Tanaka. Semiparametric least squares estimation of monotone single index models and its application to the iterative least squares estimation of binary choice models. Technical report, Citeseer, 2008.

[15] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[16] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *The Annals of Operations Research*, 46(1):157–178, 1993. MR1260016

[17] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. MR2968857

[18] Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531, 1992. MR1162997

[19] Shuangge Ma and Jian Huang. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.

[20] Joel L Horowitz. Bootstrap methods for median regression models. *Econometrica*, 66(6):1327–1351, 1998. MR1654307

[21] Junyi Zhang, Zhezhen Jin, Yongzhao Shao, and Zhiliang Ying. Statistical inference on transformation models: a self-induced smoothing approach. *arXiv preprint arXiv:1302.6651*, 2013.

[22] Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61(1):123–137, 1993. MR1201705

[23] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics (in press)*, 2015. MR3650403

[24] Jianqing Fan, Quefeng Li, and Yuyan Wang. Robust estimation of high-dimensional mean regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 79(1):247–265, 2017. MR3664833

[25] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory*, 2009.

[26] Jared C Foster, Jeremy MG Taylor, and Bin Nan. Variable selection in monotone single-index models via the adaptive lasso. *Statistics in Medicine*, 32(22):3944–3954, 2013. MR3102450

[27] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference*, 6(1):1–40, 2017. MR3636866

[28] Xinyang Yi, Zhaoran Wang, Constantine Caramanis, and Han Liu. Optimal linear estimation under unknown nonlinear transform. *arXiv preprint arXiv:1505.03257*, 2015.

[29] Peter Radchenko. High dimensional single index models. *Journal of Multivairate Analysis*, 139:266–282, 2015. MR3349492

[30] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. MR1137117

[31] Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. MR1209564

[32] R Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[33] R Dennis Cook. Principal Hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94, 1998. MR1614584

[34] Xiangrong Yin and Haileab Hilafu. Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 77(4):879–892, 2015. MR3382601

[35] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–252, 1964. MR0192611

[36] John D Kalbfleisch and Ross L Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, 2011. MR0570114

[37] Gangadharrao S Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, 1986. MR0799154

[38] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958. MR0090462

[39] Bo E Honoré and James Powell. Pairwise difference estimators for nonlinear models. In *Identification and Inference for Econometric Models*, pages 520–553. Cambridge University Press, 1997. MR2232156

[40] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. MR0138175

[41] Roger B Nelsen. *An Introduction to Copulas*. Springer, 2013. MR2197664

[42] Chris AJ Klaassen and Jon A Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997. MR1466545

[43] Marc Hallin and Davy Paindaveine. Semiparametrically efficient rank-based inference for shape. i. optimal rank-based tests for sphericity. *The Annals of Statistics*, 34(6):2707–2756, 2006. MR2329465

[44] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–147, 1974. MR0356377

[45] BM Brown and You-Gan Wang. Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92(1):149–158, 2005. MR2158616

[46] Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008. MR2396809

[47] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

[48] Aurélie C Lozano, Nicolai Meinshausen, and Eunho Yang. Minimum distance estimation for robust high-dimensional regression. *Electronic Journal of Statistics*, 10(1):1296–1340, 2016. MR3504182

[49] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. MR1379242

[50] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkv1176, 2015.

[51] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology*, 26(11):1293–1300, 2008.

[52] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010. MR2719855

[53] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012. MR3025133

[54] T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012. MR3097607

[55] Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004. MR2108039

[56] Ulf Grenander. *Abstract Inference*. Wiley New York, 1981. MR0599175

[57] Xiaotong Shen, Jian Shi, and Wing Hung Wong. Random sieve likelihood and general regression models. *Journal of the American Statistical Association*, 94(447):835–846, 1999. MR1723339

[58] Deborah Nolan and David Pollard. U-processes: rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987. MR0888439

[59] Miguel A Arcones and Evarist Gine. Limit theorems for U-processes. *The Annals of Probability*, 21(3):1494–1542, 1993. MR1235426

[60] Robert P Sherman. Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459, 1994. MR1272092

[61] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000. MR1766124

[62] Jana Jurečková, Pranab Kumar Sen, and Jan Picek. *Methodology in Robust and Nonparametric Statistics*. CRC Press, 2012. MR2963549

[63] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. MR1946581

[64] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. MR2604701

[65] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998. MR1646710

[66] Shikai Luo and Subhashis Ghosal. Forward selection and estimation in high dimensional single index model. *Statistical Methodology*, 33:172–179, 2016. MR3582782

[67] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989. MR1015136

[68] Robert E Thurman, Eric Rynes, Richard Humbert, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[69] Karen Kapur, Yi Xing, Zhengqing Ouyang, and Wing Hung Wong. Exon arrays provide accurate assessments of gene expression. *Genome Biolology*, 8(5):R82, 2007.

[70] Hong-Mei Zhang, Hu Chen, Wei Liu, Hui Liu, Jing Gong, Huili Wang, and An-Yuan Guo. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(D1):D144–D149, 2012.

[71] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, 2013. MR3221959

[72] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012. MR2963170