JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/01621459.2023.2195977

Taylor & Francis
Taylor & Francis Group

THEORY AND METHODS

Test of Significance for High-Dimensional Thresholds with Application to Individualized Minimal Clinically Important Difference

Huijie Feng[a], Jingyi Duan[a], Yang Ning[a], and Jiwei Zhao[b]

[a]Department of Statistics and Data Science, Cornell University, Ithaca, NY; [b]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

This work is motivated by learning the individualized minimal clinically important difference, a vital concept to assess clinical importance in various biomedical studies. We formulate the scientific question into a high-dimensional statistical problem where the parameter of interest lies in an individualized linear threshold. The goal is to develop a hypothesis testing procedure for the significance of a single element in this parameter as well as of a linear combination of this parameter. The difficulty dues to the high-dimensional nuisance in developing such a testing procedure, and also stems from the fact that this high-dimensional threshold model is nonregular and the limiting distribution of the corresponding estimator is nonstandard. To deal with these challenges, we construct a test statistic via a new bias-corrected smoothed decorrelated score approach, and establish its asymptotic distributions under both null and local alternative hypotheses. We propose a double-smoothing approach to select the optimal bandwidth in our test statistic and provide theoretical guarantees for the selected bandwidth. We conduct simulation studies to demonstrate how our proposed procedure can be applied in empirical studies. We apply the proposed method to a clinical trial where the scientific goal is to assess the clinical importance of a surgery procedure. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2021
Accepted March 2023

KEYWORDS
Bandwidth selection; High-dimensional statistical inference; Kernel method; Nonstandard asymptotics

1. Introduction

1.1. Motivation: Individualized Minimal Clinically Important Difference (iMCID) under High-dimensionality

In clinical studies, instead of statistical significance, the effect of a treatment or intervention is widely assessed through clinical significance. By leveraging patient-reported outcomes (PRO) that are directly collected from the patients without a third party's interpretation, the aim of assessing clinical significance is to provide clinicians and policy makers the clinical effectiveness of the treatment or intervention. For example, in our motivating study, the ChAMP randomized controlled trial (Bisson et al. 2015), the interest is to identify *the smallest WOMAC pain score change* such that the corresponding improvement and beyond can be claimed as clinically significant. In Jaeschke et al. (1989), this concept was first and formally introduced as the minimal clinically important difference (MCID), "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate a change in the patient's management".

There are roughly three approaches to determine the magnitude of MCID (Lassere et al. 2001; Erdogan et al. 2016; Angst et al. 2017; Jayadevappa et al. 2017): distribution-based, opinion-based, and anchor-based. Although adopted in various studies (Wyrwich et al. 1999a, 1999b; Samsa et al. 1999; Bellamy et al. 2001; Norman et al. 2003), the first two approaches are usually criticized (McGlothlin and Lewis 2014) due to, for example, "distribution-based methods are not derived from individual patients" and "expert opinion may not be a valid and reliable way to determine what is important to patients." The third approach, anchor-based, conceptually determines the MCID by incorporating both certainty of effective treatment encoded as a continuous variable and the patient's satisfaction collected from the anchor question. It is clinically evident (Wells et al. 2001) that the magnitude of MCID would depend on various factors such as the demographic variables and the patients' baseline status. For example, in a shoulder pain reduction study (Heald et al. 1997), because of the higher expectation for complete recovery, the healthier patients with mild pain at baseline often deemed greater pain reduction as "meaningful" than the ones who suffered from chronic disease. Therefore, it is of scientific interest to generally estimate the individualized MCID (iMCID) based on each individual patient's clinical profile as well as to quantify the uncertainties of those estimates.

Nowadays, there is an increasing use and advancing development of EHR-based (electronic health records) studies in clinical research. The EHR data are complex, diverse and high-dimensional (Abdullah et al. 2020). The rich information contained in the EHR data could facilitate the determination

CONTACT Jiwei Zhao *jiwei.zhao@wisc.edu* Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726.
The first two authors contribute equally to the article.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

and quantification of iMCID. Therefore, there is a pressing need to develop statistical methods that incorporate the high-dimensional data into both magnitude determination and uncertainty quantification of iMCID.

1.2. Problem Formulation

To facilitate the presentation, we first introduce some notation. Let $X \in \mathbb{R}$ be a continuous variable representing the score change collected from the PRO, for example, the WOMAC pain score change from baseline to one year after surgery in the ChAMP trial. Let $Y = \pm 1$ be a binary variable derived from the patient's response to the anchor question, where $Y = 1$ represents an improved health condition and $Y = -1$ otherwise. We use a d-dimensional vector $\boldsymbol{Z}$ to denote the patient's clinical profile including demographic variables, clinical biomarkers, disease histories, among many others. Suppose the data we observe are n iid samples $\{(x_i, y_i, \boldsymbol{z}_i)\}_{i=1}^n$ of $(X, Y, \boldsymbol{Z})$. We focus on the high-dimensional setting, that is, $d \gg n$.

First, if there were no covariate $\boldsymbol{Z}$, the MCID can be estimated by $\arg\max_\tau \{\mathbb{P}(X \geq \tau | Y = 1) + \mathbb{P}(X < \tau | Y = -1)\}$, which is equivalent to

$$\arg\min_{\tau} \ \mathbb{E}[w(Y) L_{01}\{Y(X - \tau)\}], \tag{1.1}$$

where $L_{01}(u) = \frac{1}{2}\{1 - \text{sign}(u)\}$ is the 0-1 loss, $\text{sign}(u) = 1$ if $u \geq 0$ and -1 otherwise, $w(1) = 1/\pi$, $w(-1) = 1/(1-\pi)$ and $\pi = \mathbb{P}(Y = 1)$. When the high-dimensional covariate $\boldsymbol{Z}$ is available, as the focus of this article, the natural idea is to consider the iMCID with a functional form of $\boldsymbol{Z}$, say $\tau(\boldsymbol{Z})$. In clinical practice, a simple structure, such as linear, is preferred due to its transparency and convenience for interpretation, especially for high-dimensional data. Therefore, we focus on the linear structure $\tau(\boldsymbol{Z}) = \boldsymbol{\beta}^T \boldsymbol{Z}$ in this article. The objective thus becomes

$$\begin{aligned} \boldsymbol{\beta}^* &= \arg\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}), \quad \text{where} \\ R(\boldsymbol{\beta}) &= \mathbb{E}\big[w(Y) L_{01}\{Y(X - \boldsymbol{\beta}^T \boldsymbol{Z})\}\big], \end{aligned} \tag{1.2}$$

and the expectation is with respect to the joint distribution of $(X, Y, \boldsymbol{Z})$. Throughout this article we assume that $\boldsymbol{\beta}^*$ exists and is unique—the existence and uniqueness can be verified under specific models; see Section S3.1 in the supplementary materials for details. Denote $\boldsymbol{\beta}^* = (\theta^*, \boldsymbol{\gamma}^{*T})^T$, where θ^* is an arbitrary one-dimensional component of $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ represents the rest of the parameter which is high-dimensional. In this article, we start from considering the hypothesis testing procedure for the parameter θ^*. With a simple reparameterization, the same procedure can be applied to infer the iMCID $\boldsymbol{c}_0^T \boldsymbol{\beta}^*$ for some fixed and known vector $\boldsymbol{c}_0 \in \mathbb{R}^d$.

It is worthwhile to mention that, although the motivation of this article is to study iMCID, our formulation of this problem can be similarly applied to other scenarios as well, such as the covariate-adjusted Youden index (Xu et al. 2014), one-bit compressed sensing (Boufounos and Baraniuk 2008), linear binary response model (Manski 1975, 1985), and personalized medicine (Wang et al. 2018). Interested readers could refer to Feng et al. (2022) for those examples.

1.3. From Estimation to Inference

Incorporating high-dimensional data in the objective, that is, moving forward from (1.1) to (1.2), is not trivial, even for the purpose of estimation only. Recently, Mukherjee et al. (2019) established the rate of convergence of the (penalized) maximum score estimator for (1.2) in growing dimension that d is allowed to grow with n. In a related work, Feng et al. (2022) proposed a regularized empirical risk minimization framework with a smoothed surrogate loss for estimating the high-dimensional parameter $\boldsymbol{\beta}^*$, and showed the estimation problem is nonregular in that there do not exist estimators of $\boldsymbol{\beta}^*$ with root-n convergence rate uniformly over a proper parameter space.

Under (1.2), developing a valid statistical inference procedure is challenging, even for fixed dimensional setting. Manski (1975, 1985) considered the binary response model $Y = \text{sign}(X - \boldsymbol{Z}^T\boldsymbol{\beta} + \epsilon)$, where ϵ may depend on $(X, \boldsymbol{Z})$ but with $\text{Median}(\epsilon | X, \boldsymbol{Z}) = 0$. It can be shown that the true coefficient $\boldsymbol{\beta}^*$ can be equivalently defined via (1.2) with $w(-1) = w(1) = 1/2$. The maximum score estimator is proposed to estimate $\boldsymbol{\beta}$, and is later shown to have a non-Gaussian limiting distribution (Kim and Pollard 1990). To tackle the challenge of nonstandard limiting distribution of the maximum score estimator, Horowitz (1992) proposed the smoothed maximum score estimator which is asymptotically normal in fixed dimension.

On top of the nonregularity of the problem (1.2), the high-dimensionality of the parameter adds an additional layer of complexity for inference. The reason is that the estimator that minimizes the penalized loss function does not have a tractable standard limiting distribution under high dimensionality, due to the bias induced by the penalty term. For regular models (e.g., generalized linear models), there is a growing literature on correcting the bias from the penalty for valid inference, such as Javanmard and Montanari (2014), Zhang and Zhang (2014), Van de Geer et al. (2014), Belloni et al. (2015), Ning and Liu (2017), Cai and Guo (2017), Fang et al. (2017), Neykov et al. (2018), Feng and Ning (2019), Fang et al. (2020), among others. Their main idea is to first construct a consistent estimator of the high dimensional parameter via proper regularization, and then remove the bias (via debiasing or decorrelation) in order to develop valid inferential statistics. While these methods enjoy great success under regular models, it remains unclear whether they can be applied to conduct valid inference in nonregular models such as the problem we consider in this article. To the best of our knowledge, our work is the first that provides valid inferential tools for nonregular models in high dimension.

1.4. Our Contributions

In this article, we propose a unified hypothesis testing framework for the one-dimensional parameter θ^* as well as for the iMCID encoded as a linear combination of $\boldsymbol{\beta}^*$. We start from considering the hypothesis testing problem $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$, where we treat $\boldsymbol{\gamma}^*$ as a high-dimensional nuisance parameter. Built on the smoothed surrogate estimation framework (Feng et al. 2022), we propose a bias corrected smoothed decorrelated score to form the score test statistic.

There are several new ingredients in the construction of our score statistic. First, the score function is derived based on a smoothed surrogate loss to overcome the nonregularity due to the nonsmoothness of the 0–1 loss. Second, unlike the existing works on high-dimensional inference, the score function from the smoothed surrogate loss is asymptotically biased. By explicitly estimating the bias term, we derive a new bias corrected score. Third, the decorrelation step, developed by Ning and Liu (2017) for regular models, is applied to reduce the uncertainty of estimating high-dimensional nuisance parameters. Compared to Ning and Liu (2017), the adoption of the smoothed loss and the corresponding bias correction step are new, which also make our inference much more challenging than the existing works. Theoretically, we show that under some conditions, the proposed score test statistic converges in distribution to a standard Gaussian distribution under the null hypothesis. We further establish the local asymptotic power of the test statistic when θ^* deviates from 0 in a local neighborhood. In particular, we give the conditions under which the test statistic has asymptotic power one.

When constructing the bias corrected smoothed decorrelated score, we need to specify a bandwidth parameter, whose optimal choice depends on the unknown smoothness of the data distribution. We further propose a double-smoothing approach to select the optimal bandwidth by minimizing the mean squared error (MSE) of the score function. To our knowledge, such bandwidth selection procedures have not been studied for high-dimensional models. We show that under some extra smoothness assumptions, the ratio of the data-driven bandwidth to the theoretically optimal bandwidth converges to one in probability. Moreover, the proposed score test statistic with the data-driven bandwidth still converges in distribution to a standard Gaussian distribution under the null hypothesis.

1.5. Paper Structure and Notation

The organization of this article is as follows. In Section 2, we first provide some background on the estimation of iMCID then introduce the bias corrected smoothed deccorelated score and the associated test statistic. In Section 3, we discuss the theoretical properties of the score test. The data-driven bandwidth selection is addressed in Section 4. The corresponding results for $\boldsymbol{c}_0^T\boldsymbol{\beta}^*$, the linear combination of parameters, are briefly summarized in Section 5. Sections 6 and 7 contain simulation studies and a real data example, respectively. All the technical details and proofs are contained in the supplementary materials.

Throughout the article, we adopt the following notation. For any set $\mathcal{S}$, we write $|\mathcal{S}|$ for its cardinality. For any vector $\boldsymbol{v} \in \mathbb{R}^d$, we use $\boldsymbol{v}_{\mathcal{S}}$ to denote the subvector of $\boldsymbol{v}$ with entries indexed by the set $\mathcal{S}$, and define its ℓ_q norm as $||\boldsymbol{v}||_q = (\sum_{j=1}^d |\boldsymbol{v}_j|^q)^{1/q}$ for some real number $q \geq 0$. For any matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$, we denote $||\boldsymbol{M}||_{\max} = \max_{i,j} |M_{ij}|$. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists some positive constant C such that $a_n \leq Cb_n$ for any n. We let $a_n \asymp b_n$ stand for $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For function $F(\theta, \boldsymbol{\gamma})$, we denote $\nabla_\theta F(\theta, \boldsymbol{\gamma})$ and $\nabla_{\boldsymbol{\gamma}} F(\theta, \boldsymbol{\gamma})$ as the first-order derivatives, and $\nabla^2_{\theta,\theta} F(\theta, \boldsymbol{\gamma})$ the second-order derivative.

2. Methodology

2.1. Review of Penalized Smoothed Surrogate Estimation

Under high dimensionality that $d \gg n$, estimating $\boldsymbol{\beta}^*$ via the empirical risk minimization (1.2) induces challenges from both statistical and computational perspectives. The nonsmoothness of $L_{01}(u)$ would cause the estimator to have a nonstandard convergence rate, which happens even in the fixed low dimensional case (Kim and Pollard 1990). Moreover, minimizing the empirical risk function based on the 0–1 loss is computationally NP-hard and is often very difficult to implement. To tackle these challenges, Feng et al. (2022) considered the following smoothed surrogate risk

$$R_\delta(\boldsymbol{\beta}) = \mathbb{E}\Big[w(Y)L_{\delta,K}\big\{Y(X - \boldsymbol{\beta}^T\boldsymbol{Z})\big\}\Big], \tag{2.1}$$

where $L_{\delta,K}(u) = \int_{u/\delta}^{\infty} K(t)dt$ is a smoothed approximation of $L_{01}(u)$, K is a kernel function defined in Section 3 and $\delta > 0$ is a bandwidth parameter. As the bandwidth δ shrinks to 0, $L_{\delta,K}(u)$ converges pointwisely to $L_{01}(u)$ (for any $u \neq 0$), from which it can be shown that $\boldsymbol{\beta}^*$ also minimizes the smoothed risk $R_\delta(\boldsymbol{\beta})$ up to a small approximation error. They further proposed the following penalized smoothed surrogate estimator

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} R^n_\delta(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{2.2}$$

where $P_\lambda(\boldsymbol{\beta})$ is some sparsity inducing penalty (e.g., Lasso) with a tuning parameter λ, and $R^n_\delta(\boldsymbol{\beta})$ is the corresponding empirical risk

$$R^n_\delta(\boldsymbol{\beta}) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \bar{R}^i_\delta(\boldsymbol{\beta}) = \tfrac{1}{n}\sum_{i=1}^n w(y_i)L_{\delta,K}\Big(y_i(x_i - \boldsymbol{\beta}^T\boldsymbol{z}_i)\Big). \tag{2.3}$$

Computationally, the empirical surrogate risk $R^n_\delta(\boldsymbol{\beta})$ is a smooth function of $\boldsymbol{\beta}$, which renders the optimization more tractable. Statistically, under some conditions, the estimator $\hat{\boldsymbol{\beta}}$ is shown to be rate-optimal, that is, the convergence rate of $\hat{\boldsymbol{\beta}}$ matches the minimax lower bound up to a logarithmic factor. We refer to Feng et al. (2022) for the detailed results.

2.2. Bias Corrected Smoothed Decorrelated Score

While Feng et al. (2022) showed that the penalized smoothed surrogate estimator $\hat{\boldsymbol{\beta}}$ is consistent, it does not automatically equip with a practical inferential procedure for $\boldsymbol{\beta}^*$, mainly because of the sparsity inducing penalty. In practice, how to draw valid statistical inference is often the ultimate goal. In our motivating example, it is of critical importance to quantify the uncertainty of $\boldsymbol{c}_0^T\boldsymbol{\beta}^*$ where $\boldsymbol{c}_0$ represents the realized value of a new patient's clinical profile. In other words, we would like to develop a testing procedure for

$$H_{0L} : \boldsymbol{c}_0^T\boldsymbol{\beta}^* = 0 \text{ versus } H_{1L} : \boldsymbol{c}_0^T\boldsymbol{\beta}^* \neq 0. \tag{2.4}$$

In this section, we focus on a special case of (2.4), the hypothesis test for θ^*,

$$H_0 : \theta^* = 0 \text{ versus } H_1 : \theta^* \neq 0, \tag{2.5}$$

where we treat $\boldsymbol{\gamma}$ as the nuisance parameter. Once the results for (2.5) are clear, we can extend them to (2.4), to be presented in Section 5.

For (2.5), we propose a new bias corrected smoothed decorrelated score test. It is well known that the classical score test is constructed based on the magnitude of the gradient of the log-likelihood, or more generally, the empirical risk function associated with $R(\boldsymbol{\beta})$ in (1.2). However, this construction breaks down in our problem due to the following two reasons.

First, to construct the score statistic, one needs to plug in some estimate of the nuisance parameter $\boldsymbol{\gamma}$ such as $\hat{\boldsymbol{\gamma}}$ obtained by partitioning $\hat{\boldsymbol{\beta}} = (\hat{\theta}, \hat{\boldsymbol{\gamma}}^T)^T$ in (2.2). However, since $\boldsymbol{\gamma}$ is a high-dimensional parameter, the estimation error from $\hat{\boldsymbol{\gamma}}$ may become the leading term in the asymptotic analysis of the score function. To deal with the high-dimensional nuisance parameter, we use the decorrelated score, where the key idea is to project the score of the parameter of interest to a high-dimensional nuisance space (Ning and Liu 2017). On the population level, it takes the form

$$\nabla_\theta R(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_{\boldsymbol{\gamma}} R(\theta, \boldsymbol{\gamma}), \tag{2.6}$$

where the decorrelation vector is $\boldsymbol{\omega}^* = \big(\nabla^2_{\boldsymbol{\gamma},\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*)\big)^{-1} \nabla^2_{\boldsymbol{\gamma},\theta} R(\boldsymbol{\beta}^*)$. When $R(\theta, \boldsymbol{\gamma})$ corresponds to the expected log-likelihood function of the data, the definition of $\boldsymbol{\omega}^*$ coincides with that in Ning and Liu (2017). In general, however, $R(\theta, \boldsymbol{\gamma})$ is not always the log-likelihood function, so we define $\boldsymbol{\omega}^*$ as $\big(\nabla^2_{\boldsymbol{\gamma},\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*)\big)^{-1} \nabla^2_{\boldsymbol{\gamma},\theta} R(\boldsymbol{\beta}^*)$ in order to mitigate the bias from estimating $\boldsymbol{\gamma}$. We refer to the review paper (Neykov et al. 2018) for further discussions.

Second, even if the above decorrelated score approach can successfully remove the effect of the high-dimensional nuisance parameter, one cannot construct the sample based decorrelated score from (2.6), as the sample version of $R(\theta, \boldsymbol{\gamma})$ is non-differentiable, leading to the so called nonstandard inference. To circumvent this issue, we approximate $R(\theta, \boldsymbol{\gamma})$ in (2.6) by the smoothed surrogate risk $R_\delta(\theta, \boldsymbol{\gamma})$ in (2.1), that is

$$\begin{aligned} &\nabla_\theta R(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_{\boldsymbol{\gamma}} R(\theta, \boldsymbol{\gamma}) \\ &\quad = \big\{\nabla_\theta R_\delta(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_{\boldsymbol{\gamma}} R_\delta(\theta, \boldsymbol{\gamma})\big\} - \text{approximation bias}. \end{aligned} \tag{2.7}$$

Since the empirical version of $R_\delta(\theta, \boldsymbol{\gamma})$ is smooth, we define the (empirical) smoothed decorrelated score function as $S_\delta(\theta, \boldsymbol{\gamma}) = \nabla_\theta R^n_\delta(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_{\boldsymbol{\gamma}} R^n_\delta(\theta, \boldsymbol{\gamma})$. With $\boldsymbol{\gamma}$ estimated by $\hat{\boldsymbol{\gamma}}$, the estimated score function is then naturally defined as

$$\hat{S}_\delta(\theta, \hat{\boldsymbol{\gamma}}) = \nabla_\theta R^n_\delta(\theta, \hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\omega}}^T \nabla_{\boldsymbol{\gamma}} R^n_\delta(\theta, \hat{\boldsymbol{\gamma}}), \tag{2.8}$$

where $\hat{\boldsymbol{\omega}}$, to be defined more precisely in Section 2.3, is an estimator of $\boldsymbol{\omega}^*$.

In view of (2.7) and (2.8), the sample version of $\nabla_\theta R_\delta(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_{\boldsymbol{\gamma}} R_\delta(\theta, \boldsymbol{\gamma})$ is given by $\hat{S}_\delta(\theta, \hat{\boldsymbol{\gamma}})$ and therefore, to construct a valid score function, it remains to estimate the approximation bias in (2.7). To proceed, we first analyze the population version of this approximation bias, which is simply $\boldsymbol{v}^{*T} \nabla R_\delta(\boldsymbol{\beta}^*)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, where $\boldsymbol{v}^* = (1, -\boldsymbol{\omega}^{*T})^T$. After some analysis, we can show that the magnitude of the approximation bias depends on the smoothness of $f(x|y, \boldsymbol{z})$, the conditional density of X given Y and $\boldsymbol{Z}$. To obtain an explicit form of the approximation bias, we assume that $f(x|y, \boldsymbol{z})$ is ℓth order differentiable for some $\ell \geq 2$, to be defined more precisely in Section 3. Under this assumption, we can show that as the bandwidth parameter $\delta \to 0$, $\boldsymbol{v}^{*T} \nabla R_\delta(\boldsymbol{\beta}^*) = \delta^\ell \mu^* (1 + o(1))$, where

$$\begin{aligned} \mu^* := \boldsymbol{v}^{*T} \boldsymbol{b}^* &= \boldsymbol{v}^{*T} \Big(\int K(u) \frac{u^\ell}{\ell!} du \Big) \\ &\quad \sum_{y \in \{-1,1\}} w(y) \int \boldsymbol{z} f^{(\ell)}(\boldsymbol{\beta}^{*T} \boldsymbol{z} | y, \boldsymbol{z}) f(y, \boldsymbol{z}) d\boldsymbol{z}, \\ &= \underbrace{\Big(\int K(u) \frac{u^\ell}{\ell!} du \Big)}_{\gamma_{K,\ell}} \boldsymbol{v}^{*T} \\ &\quad \underbrace{\mathbb{E}\Big[w(Y) Y \boldsymbol{Z} f^{(\ell)}(\boldsymbol{\beta}^{*T} \boldsymbol{Z} | Y, \boldsymbol{Z}) \Big]}_{T^{(\ell)}(\boldsymbol{\beta}^*)}, \end{aligned} \tag{2.9}$$

and $f^{(\ell)}(x|y, \boldsymbol{z})$ denotes the ℓth order derivative of $f(x|y, \boldsymbol{z})$ with respect to x.

To estimate the approximation bias $\boldsymbol{v}^{*T} \nabla R_\delta(\boldsymbol{\beta}^*)$, it suffices to estimate μ^*. From (2.9), once $f^{(\ell)}(x|y, \boldsymbol{z})$ at $x = \boldsymbol{\beta}^{*T} \boldsymbol{z}$ is estimated, we can construct a plug-in estimator for μ^*. To be specific, assume that a pilot kernel estimator with some kernel function U and bandwidth h is available to estimate $f^{(\ell)}(\boldsymbol{\beta}^{*T} \boldsymbol{z} | y, \boldsymbol{z})$. Then we can estimate μ^* by

$$\hat{\mu} = \gamma_{K,\ell} \hat{\boldsymbol{v}}^T \hat{T}^{(\ell),n}_{h,U}(\hat{\boldsymbol{\beta}}), \tag{2.10}$$

where $\hat{T}^{(\ell),n}_{h,U}(\hat{\boldsymbol{\beta}}) := \frac{1}{n} \sum_{i=1}^n w(y_i) y_i \frac{\boldsymbol{z}_i}{h^{1+\ell}} U^{(\ell)}\big(\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i - x_i}{h}\big)$ and $\hat{\boldsymbol{v}} = (1, -\hat{\boldsymbol{\omega}}^T)^T$.

The last step to construct a valid score test is to find the asymptotic variance of the smoothed decorrelated score $S_\delta(\boldsymbol{\beta}^*)$. Lemma 1 in the next section shows that the asymptotic variance of the standardized decorrelated score $(n\delta)^{1/2} S_\delta(\boldsymbol{\beta}^*)$ is $\sigma^{*2} = \boldsymbol{v}^{*T} \Sigma^* \boldsymbol{v}^*$, where

$$\begin{aligned} \Sigma^* &:= \sum_{y \in \{-1,1\}} w(y)^2 \int \boldsymbol{z}\boldsymbol{z}^T \int K(u)^2 du f(\boldsymbol{\beta}^{*T} \boldsymbol{z} | y, \boldsymbol{z}) f(y, \boldsymbol{z}) d\boldsymbol{z}, \\ &= \underbrace{\Big(\int K(u)^2 du \Big)}_{\tilde{\mu}_K} \underbrace{\mathbb{E}\Big[w(Y)^2 \boldsymbol{Z}\boldsymbol{Z}^T f(\boldsymbol{\beta}^{*T} \boldsymbol{Z} | Y, \boldsymbol{Z}) \Big]}_{H(\boldsymbol{\beta}^*)}, \end{aligned} \tag{2.11}$$

and thus σ^* can be estimated by

$$\hat{\sigma} = \sqrt{\tilde{\mu}_K \hat{\boldsymbol{v}}^T \hat{H}^n_{g,L}(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{v}}}, \tag{2.12}$$

where $\hat{H}^n_{g,L}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^n w^2(y_i) \boldsymbol{z}_i \boldsymbol{z}_i^T \frac{1}{g} L\big(\frac{x_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}{g}\big)$ with some kernel function L and bandwidth g. In Section S4 in the Supplement, we propose an alternative kernel-free estimator of σ^*, which does not require any additional kernel function or bandwidth. We show that the estimator is still consistent for σ^* but may have a slower convergence rate than $\hat{\sigma}$ here.

Equipped with the smoothed decorrelated score $\hat{S}_\delta(\theta, \hat{\boldsymbol{\gamma}})$ in (2.8), the estimate of the approximation bias $\delta^\ell \hat{\mu}$ in (2.10) and the estimate of the asymptotic variance $\hat{\sigma}^2$ in (2.12), we define the bias corrected smoothed decorrelated score statistic as

$$\hat{U}_n = \sqrt{n\delta} \Big(\frac{\hat{S}_\delta(0, \hat{\boldsymbol{\gamma}}) - \delta^\ell \hat{\mu}}{\hat{\sigma}} \Big). \tag{2.13}$$

Remark 1. Compared to the existing decorrelated score approach (Ning and Liu 2017), our methodological innovation is to develop an explicit bias correction step to remove the approximation bias in (2.7) induced by the smoothed surrogate risk. From the theoretical aspect, our test statistic $\hat{U}_n$ is rescaled by $(n\delta)^{1/2}$ rather than the classical $n^{1/2}$ factor, which leads to the nonstandard rate of the decorrelated score not only under the null but also under local alternatives; see Section 3.

2.3. Detailed Implementation

For numerical implementation, we follow the path-following algorithm presented in Feng et al. (2022) to compute the initial estimator $\hat{\boldsymbol{\beta}}$. For the estimator $\hat{\boldsymbol{\omega}}$, recall that $\boldsymbol{\omega}^*$ satisfies $\nabla^2_{\boldsymbol{\gamma},\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*)\boldsymbol{\omega}^* = \nabla^2_{\boldsymbol{\gamma},\theta} R(\boldsymbol{\beta}^*)$. Since $\nabla^2 R(\boldsymbol{\beta}^*)$ can be approximated by the Hessian of the smoothed surrogate loss $\nabla^2 R^n_\delta(\boldsymbol{\beta}^*)$, we consider the following Dantzig type estimator $\hat{\boldsymbol{\omega}}$, where

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} ||\boldsymbol{\omega}||_1 \ \ s.t. \ \ ||\nabla^2_{\boldsymbol{\gamma},\theta} R^n_\delta(\hat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\gamma},\boldsymbol{\gamma}} R^n_\delta(\hat{\boldsymbol{\beta}})\boldsymbol{\omega}||_\infty \le \lambda', \tag{2.14}$$

for some tuning parameter $\lambda' > 0$.

For implementing $\hat{U}_n$, we note that the analysis of the asymptotic distribution of $\hat{U}_n$ is complicated by the dependence between the estimator $\hat{\boldsymbol{\beta}}$ and $S_\delta(\theta, \boldsymbol{\gamma})$. To decouple the dependence and ease theoretical development, we apply the cross-fitting technique to construct the bias corrected smoothed decorrelated score. Specifically, instead of using the same set of samples for estimating $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\omega}}$ and constructing the score function $S_\delta(\theta, \boldsymbol{\gamma})$, we will first estimate $\hat{\boldsymbol{\beta}}$ using one set of samples, and then use the rest of samples for estimating $\hat{\boldsymbol{\omega}}$ and constructing $S_\delta(\theta, \boldsymbol{\gamma})$. We can further switch the samples and aggregate the decorrelated score. Without loss of generality, assume the sample size n is even and we divide the samples into two halves with equal size for this purpose. Formally, denote $\hat{\boldsymbol{\beta}}^{(i)}, \hat{\boldsymbol{\omega}}^{(i)}, i = 1, 2$ as the estimator based on the ith fold of the samples, $\mathcal{N}_i$, and similarly $\nabla R^{n_{(i)}}_\delta(\boldsymbol{\beta}), \nabla^2 R^{n_{(i)}}_\delta(\boldsymbol{\beta})$ as the corresponding gradient and Hessian. Define

$$\hat{S}^{(1)}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(2)}) = \nabla_\theta R^{n_{(1)}}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(2)}) - \hat{\boldsymbol{\omega}}^{(1)T} \nabla_{\boldsymbol{\gamma}} R^{n_{(1)}}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(2)}),$$

and $\hat{S}^{(2)}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(1)})$ in a similar way. The estimated decorrelated score via cross-fitting is

$$\hat{S}_\delta(\theta, \hat{\boldsymbol{\gamma}}) = \tfrac{1}{2}\big(\hat{S}^{(1)}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(2)}) + \hat{S}^{(2)}_\delta(\theta, \hat{\boldsymbol{\gamma}}^{(1)})\big). \tag{2.15}$$

Similarly, we define the cross-fitted estimators $\hat{\mu}$ and $\hat{\sigma}$ as

$$\begin{aligned} \hat{\mu} &= \frac{1}{2}\gamma_{K,\ell}(\hat{\boldsymbol{v}}^{(1)T}\hat{T}^{(\ell),n_{(1)}}_{h,U}(\hat{\boldsymbol{\beta}}^{(2)}) + \hat{\boldsymbol{v}}^{(2)T}\hat{T}^{(\ell),n_{(2)}}_{h,U}(\hat{\boldsymbol{\beta}}^{(1)})), \\ \hat{\sigma}^2 &= \frac{\tilde{\mu}_K}{2}\left[\hat{\boldsymbol{v}}^{(1)T}\hat{H}^{n_{(1)}}_{g,K}(\hat{\boldsymbol{\beta}}^{(2)})\hat{\boldsymbol{v}}^{(1)} + \hat{\boldsymbol{v}}^{(2)T}\hat{H}^{n_{(2)}}_{g,K}(\hat{\boldsymbol{\beta}}^{(1)})\hat{\boldsymbol{v}}^{(2)}\right], \end{aligned} \tag{2.16}$$

where

$$\hat{T}^{(\ell),n_{(1)}}_{h,U}(\hat{\boldsymbol{\beta}}^{(2)}) = \frac{1}{|\mathcal{N}_1|}\sum_{i\in\mathcal{N}_1} w(y_i) y_i \frac{\boldsymbol{z}_i}{h^{1+\ell}} U^{(\ell)}\left(\frac{\hat{\boldsymbol{\beta}}^{(2)T}\boldsymbol{z}_i - x_i}{h}\right),$$

$$\hat{H}^{n_{(1)}}_{g,L}(\hat{\boldsymbol{\beta}}^{(2)}) = \frac{1}{|\mathcal{N}_1|}\sum_{i\in\mathcal{N}_1} w^2(y_i)\boldsymbol{z}_i\boldsymbol{z}_i^T \frac{1}{g} L\left(\frac{x_i - \hat{\boldsymbol{\beta}}^{(2)T}\boldsymbol{z}_i}{g}\right), \tag{2.17}$$

and similarly for $\hat{T}^{(\ell),n_{(2)}}_{h,U}(\hat{\boldsymbol{\beta}}^{(1)}), \hat{H}^{n_{(2)}}_{g,L}(\hat{\boldsymbol{\beta}}^{(1)})$. Given $\hat{S}_\delta(\theta, \hat{\boldsymbol{\gamma}})$ in (2.15) and the above estimators $\hat{\mu}$ and $\hat{\sigma}$, we can form the score test statistic $\hat{U}_n$ in the same way as in (2.13).

3. Theory

3.1. Assumptions

In this article, we consider the following definition of function smoothness.

Definition 1. We say the conditional density $f(x|y,\boldsymbol{z})$ of X given $Y, \boldsymbol{Z}$ is ℓth order smooth, if for any $\boldsymbol{z}$ and $y \in \{-1, 1\}$, the conditional density $f(x|y,\boldsymbol{z})$ is ℓ-times continuously differentiable in x with derivatives $f^{(i)}(x|y,\boldsymbol{z})$ bounded by a constant C, $|f^{(i)}(x|y,\boldsymbol{z})| \le C$ for $i = 1, \ldots, \ell$, and $f^{(\ell)}(x|y,\boldsymbol{z})$ is Hölder continuous with some exponent $0 < \zeta \le 1$, that is, for any $\boldsymbol{z}, \triangle$ and $y \in \{-1, 1\}$, $|f^{(\ell)}(x + \triangle|y,\boldsymbol{z}) - f^{(\ell)}(x|y,\boldsymbol{z})| \le L\triangle^\zeta$, where $L > 0$ is some constant.

Assumption 1. We assume $f(x|y,\boldsymbol{z})$ is ℓth order smooth with some integer $\ell \ge 2$.

Assumption 1 concerns the smoothness of $f(x|y,\boldsymbol{z})$. To see why the smoothness condition is important, notice that the gradient functions of (2.1) and (1.2) are

$$\begin{aligned} \nabla R_\delta(\boldsymbol{\beta}) &= \sum_{y\in\{-1,1\}} w(y) \\ &\qquad \int yz\Big[\int \frac{1}{\delta} K\Big(\frac{y(x - \boldsymbol{\beta}^T\boldsymbol{z})}{\delta}\Big) f(x|y,\boldsymbol{z})dx\Big] f(y,\boldsymbol{z})d\boldsymbol{z} \\ \nabla R(\boldsymbol{\beta}) &= \sum_{y\in\{-1,1\}} w(y) \int yz f(\boldsymbol{\beta}^T\boldsymbol{z}|y,\boldsymbol{z}) f(y,\boldsymbol{z})d\boldsymbol{z}, \end{aligned} \tag{3.1}$$

from which we can see that $f(\boldsymbol{\beta}^T\boldsymbol{z}|y,\boldsymbol{z})$ in $\nabla R(\boldsymbol{\beta})$ is substituted by its kernel approximation $\int \frac{1}{\delta} K(\frac{y(x-\boldsymbol{\beta}^T\boldsymbol{z})}{\delta}) f(x|y,\boldsymbol{z})dx$, and thus the difference between $\nabla R_\delta(\boldsymbol{\beta})$ and $\nabla R(\boldsymbol{\beta})$ naturally depends on the smoothness of $f(x|y,\boldsymbol{z})$.

Notice that our smoothness condition in Definition 1 is slightly stronger than the standard Hölder smoothness condition in the nonparametric literature (Tsybakov 2009). In particular, we require that $f^{(\ell)}(x|y,\boldsymbol{z})$ is Hölder continuous with some exponent $0 < \zeta \le 1$. This additional assumption is essential to show the rate of the bias estimator $\hat{\mu}$ in (2.10). The Hölder class condition in Assumption 1 can be relaxed to a variation of Nikol'ski class condition (Tsybakov 2009); see Section S3.2 in the supplementary materials for details.

Assumption 2. We assume $K(t)$ is a kernel function with bounded support that satisfies: $K(t) = K(-t)$, $|K(t)| \le K_{\max} < \infty$ $\forall\ t \in \mathbb{R}$, $\int K(t)dt = 1$, $\int K^2(t)dt < \infty$, and $|K'| < \infty$. We also assume that K degenerates at the boundaries. A kernel is said to be of order $\ell \ge 1$ if it satisfies $\int t^j K(t)dt = 0$, $\forall\ j = 1, \ldots, \ell - 1$, $\int t^\ell K(t)dt \ne 0$, and $\int |t|^q |K(t)|dt$ are bounded by a constant for any $q \in [\ell, \ell + 1]$.

Assumption 2 above is about the kernel function $K(t)$ that we first introduced in the surrogate risk $R_\delta(\boldsymbol{\beta})$ in (2.1). We provide a list of commonly seen second-order, fourth-order, and sixth-order kernel functions in Section S3.3 of the supplementary materials.

We now impose regularity conditions on $(X, Y, \boldsymbol{Z})$.

Assumption 3. There exists a constant $c > 0$ such that $c \leq \mathbb{P}(Y = 1) \leq 1 - c$ and the weight function $w(\cdot)$ is positive and upper bounded by a constant.

Assumption 4. We assume $\max_{1\leq j\leq d} |Z_j| \leq M_n$ for some M_n that possibly depends on n, where $M_n^2 \leq C\sqrt{n\delta/\log(d)}$ for some constant $C > 0$. We also assume that $\mathbb{E}[|Z_j|^4|Y = y]$ is bounded by a constant for $y \in \{1, -1\}$.

Assumption 5. We assume $\sigma^* = \sqrt{\boldsymbol{v}^{*T}\Sigma^*\boldsymbol{v}^*}$ is bounded away from 0 and infinity by some constants, and $|\mu^*| = |\boldsymbol{v}^{*T}\boldsymbol{b}^*|$ is also upper bounded by a constant.

Assumption 4 requires the boundedness of $\boldsymbol{Z}$ and the fourth order moment. Notice that if each component of $\boldsymbol{Z}$ is sub-Gaussian with bounded sub-Gaussian norm, Assumption 4 is satisfied with high probability with $M_n \asymp \sqrt{\log d}$ providing $(\log d)^3/(n\delta) = O(1)$ which is a mild assumption. For binary covariates $Z_j \in \{0, 1\}$, it holds that $M_n = 1$. Assumption 5 ensures that the asymptotic variance of the smoothed decorrelated score σ^* does not degenerate and the approximation bias μ^* is bounded. In Section S3.4 in the supplementary materials, we verify that under mild conditions, Assumptions 1–5 hold under the binary response model. Finally, we impose the following assumption on the estimators of $\boldsymbol{\beta}^*$ and $\boldsymbol{\omega}^*$.

Assumption 6. Assume there are estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{v}} = (1, -\hat{\boldsymbol{\omega}}^T)^T$ with

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \lesssim \eta_1(n) \quad \text{and} \quad ||\hat{\boldsymbol{v}} - \boldsymbol{v}^*||_1/||\boldsymbol{v}^*||_1 \lesssim \eta_2(n),$$

for some nonrandom sequences $\eta_1(n), \eta_2(n)$ converging to 0 as $n \to \infty$.

It is shown by Feng et al. (2022) that, under some conditions, the estimator $\hat{\boldsymbol{\beta}}$ in (2.2) achieves the (near) minimax-optimal rate $\eta_1(n) = \sqrt{s}(\frac{s\log(d)}{n})^{\ell/(2\ell+1)}$ where $s = ||\boldsymbol{\beta}^*||_0$. For $\hat{\boldsymbol{v}} = (1, -\hat{\boldsymbol{\omega}}^T)^T$, we assume $||\hat{\boldsymbol{v}} - \boldsymbol{v}^*||_1 \lesssim ||\boldsymbol{v}^*||_1\eta_2(n)$. Notice that the term $||\boldsymbol{v}^*||_1$ is not absorbed into $\eta_2(n)$ only for notational simplicity. In Lemma S7 in the Supplement, we show that a Dantzig type estimator $\hat{\boldsymbol{v}}$ could attain the fast rate $\eta_2(n)$.

3.2. Theoretical Results

We start from the following lemma which characterizes the asymptotic distribution of the decorrelated score function evaluated at the true parameter $\boldsymbol{\beta}^*$.

Lemma 1. Under Assumptions 1–5, if $(||\boldsymbol{v}^*||_1M_n)^3/(n\delta)^{1/2} = o(1)$ and $\delta = o(1)$, then

$$\sqrt{n\delta}\frac{\boldsymbol{v}^{*T}(\nabla R_\delta^n(\boldsymbol{\beta}^*) - \nabla R_\delta(\boldsymbol{\beta}^*))}{\sqrt{\boldsymbol{v}^{*T}\Sigma^*\boldsymbol{v}^*}} \xrightarrow{d} N(0, 1), \tag{3.2}$$

where

$$\boldsymbol{v}^{*T}\nabla R_\delta(\boldsymbol{\beta}^*) = \delta^\ell \boldsymbol{v}^{*T}\boldsymbol{b}^*(1 + o(1)). \tag{3.3}$$

Asymptotically, the bias and standard deviation of $\boldsymbol{v}^{*T}\nabla R_\delta^n(\boldsymbol{\beta}^*)$ can be seen from this lemma. Since $\mu^* = \boldsymbol{v}^{*T}\boldsymbol{b}^*$ and $\sigma^* = \sqrt{\boldsymbol{v}^{*T}\Sigma^*\boldsymbol{v}^*}$ are both bounded by constants, the asymptotic bias and standard deviation are of order δ^ℓ and $(n\delta)^{-1/2}$, respectively. Thus, choosing $\delta = cn^{-1/(2\ell+1)}$ for any constant $c > 0$ attains the optimal bias and variance tradeoff. Note that in this lemma we require $(||\boldsymbol{v}^*||_1M_n)^3/(n\delta)^{1/2} = o(1)$ to verify the Lindeberg condition in the central limit theorem, which holds as long as δ does not shrink to zero too fast.

Our first main theorem characterizes the asymptotic normality of the decorrelated score under the null hypothesis with nuisance parameters $\boldsymbol{\gamma}^*$ and $\boldsymbol{\omega}^*$ estimated by those in Assumption 6.

Theorem 1. Under Assumptions 1–6, if $(||\boldsymbol{v}^*||_1M_n)^3/(n\delta)^{1/2} = o(1)$, $\frac{\log(d)}{n\delta^3} = o(1)$, $n\delta^{2\ell+1} = O(1)$, and

$$(n\delta)^{1/2}||\boldsymbol{v}^*||_1\left(\frac{\eta_1(n)}{\delta} \vee \eta_2(n)\right)\left(\sqrt{\frac{\log(d)}{n\delta}} \vee \delta^\ell \vee M_n^2\eta_1(n)\right) = o(1), \tag{3.4}$$

then under $H_0 : \theta^* = 0$, it holds that $\frac{\sqrt{n\delta}\hat{S}_\delta(0,\hat{\boldsymbol{\gamma}}) - \sqrt{n\delta^{2\ell+1}}\mu^*}{\sigma^*} \xrightarrow{d} N(0, 1)$.

Theorem 1 implies that the decorrelated score with some high-dimensional plug-in estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}$ has the same asymptotic distribution as in Lemma 1. Several conditions are needed to show this result. The first condition $(||\boldsymbol{v}^*||_1M_n)^3/(n\delta)^{1/2} = o(1)$ is from Lemma 1, and the second condition $\frac{\log(d)}{n\delta^3} = o(1)$ is also mild as long as δ does not go to zero too fast. The third condition $n\delta^{2\ell+1} = O(1)$ guarantees that the higher order bias of the decorrelated score can be ignored and therefore it suffices to only correct for the leading bias term in (3.3).

We now elaborate the condition (3.4). Roughly speaking, the term $\sqrt{\frac{\log(d)}{n\delta}} \vee \delta^\ell \vee M_n^2\eta_1(n)$ comes from the bound for $||\nabla R_\delta^{n_{(1)}}(\theta, \hat{\boldsymbol{\gamma}}^{(2)}) - \nabla R(\theta, \boldsymbol{\gamma})||_\infty$. Indeed, the cross-fitting technique guarantees the independence between $\hat{\boldsymbol{\gamma}}^{(2)}$ and $\nabla R_\delta^{n_{(1)}}(\theta, \boldsymbol{\gamma})$, which plays a key role in the analysis. Condition (3.4) simply means that this bound interacting with the estimation error of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}$ is sufficiently small. We can further simplify the condition (3.4) by plugging the order of $\eta_1(n)$ derived in Feng et al. (2022) and $\eta_2(n) = s'(\log(d)/n)^{(\ell-1)/(2\ell+1)}$ derived from Lemma S7 in the supplementary materials where $s' = ||\boldsymbol{\omega}^*||_0$.

Recall that in our score statistic $\hat{U}_n$ in (2.13), we plug in the estimators $\hat{\mu}$ and $\hat{\sigma}$ for μ^* and σ^*. In Lemmas S8 and S9 in the supplementary materials, we establish the rate of convergence of $\hat{\mu}$ and $\hat{\sigma}$. Under the assumption that $|\hat{\mu} - \mu^*| = o_p(1)$ and $|\hat{\sigma} - \sigma^*| = o_p(1)$, the Slutsky's theorem implies that the bias corrected decorrelated score statistic $\hat{U}_n \xrightarrow{d} N(0, 1)$ under the null hypothesis.

Accordingly, given the desired significance level α, we define the test function as

$$T_{\text{DS}} = I(|\hat{U}_n| > \Phi^{-1}(1 - \alpha/2)),$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the cdf of the standard normal distribution. Thus, our result shows that the Type I error of the test T_{DS} converges to α asymptotically, that is, $\mathbb{P}(T_{\mathrm{DS}} = 1|H_0) \to \alpha$.

Now denote $\nabla^2_{\theta|\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*) = \nabla^2_{\theta\theta} R(\boldsymbol{\beta}^*) - \nabla^2_{\theta\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*)(\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*))^{-1}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\theta}} R(\boldsymbol{\beta}^*)$. Our second main theorem characterizes the limiting behavior of $\hat{U}_n$ under the local alternative hypothesis $H_1 : \theta^* = \tilde{C}n^{-\phi}$ for some constants $\tilde{C} \neq 0$ and $\phi > 0$.

Theorem 2. Assume the conditions in Theorem 1 and in Lemmas S8 and S9 of the supplementary materials, and further

$$||\boldsymbol{v}^*||_1^2 M_n^4 n^{1-4\phi}/\delta = o(1),$$
$$(n\delta)^{1/2}||\boldsymbol{v}^*||_1(\eta_1(n) \vee \eta_2(n))M_n n^{-\phi} = o(1), \tag{3.5}$$

and that $\hat{\mu}, \hat{\sigma}$ are consistent estimators of μ^*, σ^*. Then, by choosing the optimal bandwidth $\delta \asymp n^{-1/(2\ell+1)}$, the following results hold under the local alternative hypothesis $H_1 : \theta^* = \tilde{C}n^{-\phi}$.

1. When $\phi = \frac{\ell}{2\ell+1}$, it holds that $\hat{U}_n \xrightarrow{d} N(-\xi, 1)$, where $\xi = \tilde{C}\nabla^2_{\theta|\boldsymbol{\gamma}} R(\boldsymbol{\beta}^*)/\sigma^*$ is assumed to be a constant.
2. When $\phi < \frac{\ell}{2\ell+1}$, it holds that for any fixed t, $\lim_{n\to\infty} \mathbb{P}(|\hat{U}_n| > t) = 1$.

In addition to the conditions imposed in Theorem 1 and Lemmas S8 and S9, we further require two additional conditions involving the magnitude of θ^* in (3.5). The first condition $||\boldsymbol{v}^*||_1^2 M_n^4 n^{1-4\phi}/\delta = o(1)$ is imposed to ensure the local asymptotic normality (LAN) in terms of the parameter θ^*. The second condition in (3.5) is similar to (3.4), which controls the magnitude of $||\nabla R(0, \boldsymbol{\gamma}^*)||_\infty$ and $||\nabla^2 R_\delta(\theta^*, \boldsymbol{\gamma}^*) - \nabla^2 R_\delta(0, \boldsymbol{\gamma}^*)||_{\max}$ under the alternative hypothesis.

This theorem implies that the proposed test converges in distribution to a normal distribution with mean $-\xi$, when the contiguous alternatives approach the null hypothesis at a rate $n^{-\frac{\ell}{2\ell+1}}$. In addition, if the alternatives deviate from the null hypothesis in the magnitude larger than $n^{-\frac{\ell}{2\ell+1}}$ (i.e., $\phi < \frac{\ell}{2\ell+1}$), the asymptotic power of our test is 1. In other words, our test can successfully detect the nonzero θ^* whose magnitude exceeds the order of $n^{-\ell/(2\ell+1)}$. In contrast, for regular models, the local alternative that is detectable is of the standard parametric rate $n^{-1/2}$.

Remark 2. By choosing the optimal bandwidth $\delta \asymp n^{-1/(2\ell+1)}$, all the conditions in Theorem 2 can be simplified and summarized as

$$M_n^2 \le n^{\ell/(2\ell+1)}(\log d)^{-1/2},$$
$$\log d = o(n^{(2\ell-2)/(2\ell+1)}),$$
$$||\boldsymbol{v}^*||_1 M_n^2 = o(n^{2\phi-(\ell+1)/(2\ell+1)}), \quad \text{and}$$
$$||\boldsymbol{v}^*||_1 M_n = o(n^{\ell/(6\ell+3)} \wedge n^{\phi-\ell/(2\ell+1)}\{\eta_1(n) \vee \eta_2(n)\}^{-1}),$$

where $\eta_1(n) = s^{(4\ell+1)/(4\ell+2)}(\log d/n)^{\ell/(2\ell+1)} = o(1)$, $\eta_2(n) = s'(\log d/n)^{(\ell-1)/(2\ell+1)} = o(1)$ and $||\boldsymbol{v}^*||_1\{(\log d)^{1/2} \vee n^{\ell/(2\ell+1)}M_n^2\eta_1(n)\}\{n^{1/(2\ell+1)}\eta_1(n) \vee \eta_2(n)\} = o(1)$. Consider the extreme case $\ell \to \infty$, it can be verified that $\log d = n^{1/5}$, $s = s' = n^{1/10}$, $||\boldsymbol{v}^*||_1 = n^{1/20}$, $M_n = n^{1/50}$ would satisfy all of these conditions when $\phi = 1/2$.

In addition, these conditions could be further simplified if one is willing to assume $||\boldsymbol{v}^*||_1 = O(1)$ and $M_n = O(1)$. If that is the case, condition (3.4) becomes

$$s^{(4\ell+1)/(4\ell+2)}(s^{(4\ell+1)/(4\ell+2)} \vee s')$$
$$n^{-(\ell-1)/(2\ell+1)}(\log d)^{(4\ell-1)/(4\ell+2)} = o(1)$$

when taking $\delta \asymp n^{-1/(2\ell+1)}$. If we consider the extreme case with $\ell \to \infty$, it suffices to have $s(s \vee s')\log d = o(n^{1/2})$ in order to satisfy all of the conditions when $\phi = 1/2$.

4. Data-Driven Bandwidth Selection

In the previous section, we establish the theoretical property of the bias corrected smoothed decorrelated score when the underlying conditional density $f(x|y, \boldsymbol{z})$ is ℓth order smooth. However, this smoothness parameter ℓ is typically unknown in practice, leading to the following two complications. First, in Assumption 2, a kernel function K of the same order is applied, which implicitly requires the knowledge on the smoothness parameter ℓ. In practice, the choice of kernel functions is often determined by the user's preference rather than the theory. Since high order kernels may exacerbate the problem of variability, choosing low order kernels of 2 or 4 is often recommended (even if the density is more smooth); see Härdle et al. (1992). Second, the optimal bandwidth $\delta \asymp n^{-1/(2\ell+1)}$ that balances the asymptotic bias and variance of the decorrelated score in Lemma 1 also depends on the unknown ℓ. It is well known from the nonparametric literature that the choice of bandwidth is an extremely important problem of both theoretical and practical values (Bowman 1984; Silverman 1986; Sheather and Jones 1991; Hall et al. 1992; Jones et al. 1996).

In this section, we focus on how to choose the bandwidth δ in a data-driven manner. In view of the above discussion on the kernels, we assume that a low order kernel K is chosen (for simplicity, we still denote its order by ℓ) and meanwhile the underlying conditional density has a higher order smoothness parameter.

Assumption 7. We assume that the kernel K is of order ℓ and $f(x|y, \boldsymbol{z})$ is $(\ell + r)$th order smooth for some $\ell \ge 2$ and $r > 0$.

We define the optimal bandwidth δ^* as the one that minimizes the MSE of the smoothed decorrelated score:

$$\delta^* = \arg\min_\delta M(\delta), \quad \text{where} \quad M(\delta) = \mathbb{E}[(\boldsymbol{v}^{*T}\nabla R_\delta^n(\boldsymbol{\beta}^*))^2]. \tag{4.1}$$

A direct bias-variance decomposition of $M(\delta)$ gives

$$M(\delta) = \frac{1}{n}\mathbb{E}[(\boldsymbol{v}^{*T}\nabla \bar{R}_\delta^1(\boldsymbol{\beta}^*))^2] + \frac{n-1}{n}(\boldsymbol{v}^{*T}\nabla R_\delta(\boldsymbol{\beta}^*))^2$$
$$:= \frac{1}{n}V(\delta) + \frac{n-1}{n}SB(\delta), \tag{4.2}$$

where $\nabla \bar{R}_\delta^1(\boldsymbol{\beta}^*)$ is defined in (2.3), $V(\delta) = \mathbb{E}[(\boldsymbol{v}^{*T}\nabla \bar{R}_\delta^1(\boldsymbol{\beta}^*))^2]$ is used as a proxy for the variance, and $SB(\delta) = (\boldsymbol{v}^{*T}\nabla R_\delta(\boldsymbol{\beta}^*))^2$ denotes the squared error. To estimate δ^*, our main idea is to construct estimators $\hat{V}(\delta)$ and $\hat{SB}(\delta)$ for $V(\delta)$ and $SB(\delta)$ and then estimate δ^* by

$$\hat{\delta} = \arg\min_\delta \hat{M}(\delta) \quad \text{where} \quad \hat{M}(\delta) = \frac{1}{n}\hat{V}(\delta) + \frac{n-1}{n}\hat{SB}(\delta).$$

From the proof of Lemma 1, we can show that $SB(\delta) = (\delta^\ell \mu^*)^2(1+o(1))$ and $V(\delta) = \delta^{-1}\sigma^{*2}(1+o(1))$ as $\delta \to 0$, and thus σ^{*2}/δ and $(\delta^\ell\mu^*)^2$ are the asymptotic versions of $V(\delta)$ and $SB(\delta)$, respectively. As a result, one may attempt to estimate the optimal bandwidth by minimizing the asymptotic MSE $\hat{\sigma}^2/\delta + (\delta^\ell\hat{\mu})^2$ with the plug-in estimators $\hat{\sigma}$ and $\hat{\mu}$ developed in the previous section. However, the asymptotic MSE depends on the unknown smoothness ℓ and therefore is not appropriate for bandwidth selection in practice.

Instead, we propose to estimate $V(\delta)$ and $SB(\delta)$ using a different strategy. Our estimates are still in the cross-fitting fashion, but when we estimate the bias $B(\delta)$ that dues to the approximation using the kernel function $K(\cdot)$ of order ℓ with bandwidth δ, we will have to use a new pilot kernel function $J(\cdot)$ of order r with bandwidth b. Essentially when the target function has higher order smoothness than the kernel function applied for estimation, a different kernel smoothing procedure has to be used for estimating the bias. This is motivated by the "double-smoothing" technique in nonparametric statistics (Härdle et al. 1992; Neumann 1995), and is also related to the "smoothed cross validation" approach (Hall et al. 1992).

To be specific, we estimate $V(\delta)$ with the following moment estimator

$$\hat{V}(\delta) = \tfrac{1}{2}\big(\hat{\boldsymbol{v}}^{(1)T}\hat{\Gamma}^{(1)}(\delta)\hat{\boldsymbol{v}}^{(1)} + \hat{\boldsymbol{v}}^{(2)T}\hat{\Gamma}^{(2)}(\delta)\hat{\boldsymbol{v}}^{(2)}\big), \tag{4.3}$$

where $\hat{\Gamma}^{(1)}(\delta) = \frac{1}{|\mathcal{N}_1|}\sum_{i\in\mathcal{N}_1}\nabla\bar{R}^i_\delta(\hat{\boldsymbol{\beta}}^{(2)})\nabla\bar{R}^i_\delta(\hat{\boldsymbol{\beta}}^{(2)})^T$ and similarly for $\hat{\Gamma}^{(2)}$. To estimate the squared bias $SB(\delta)$, note that the bias term $B(\delta)$ can be written as

$$\begin{aligned} B(\delta) &= \boldsymbol{v}^{*T}\nabla R_\delta(\boldsymbol{\beta}^*) = \boldsymbol{v}^{*T}\underbrace{(\nabla R_\delta(\boldsymbol{\beta}^*) - \nabla R(\boldsymbol{\beta}^*))}_{A(\boldsymbol{\beta}^*,\delta)} \\ &= \int_u K(u)\Big\{\boldsymbol{v}^{*T}(\nabla R(u\delta,\boldsymbol{\beta}^*) - \nabla R(\boldsymbol{\beta}^*))\Big\}du, \end{aligned} \tag{4.4}$$

where we use $\nabla R(u\delta,\boldsymbol{\beta}^*) = \sum_y w(y)\int_{\boldsymbol{z}} \boldsymbol{z}yf(u\delta + \boldsymbol{\beta}^{*T}\boldsymbol{z}|y,\boldsymbol{z})f(y,\boldsymbol{z})d\boldsymbol{z}$ to denote the population gradient with a bias induced by $u\delta$, with a bit abuse of notation. We propose to estimate $B(\delta)$ by

$$\begin{aligned} \hat{B}(\delta) = \frac{1}{2}\Bigg(&\hat{\boldsymbol{v}}^{(1)T}\frac{1}{|\mathcal{N}_1|}\sum_{i\in\mathcal{N}_1}A_i(\hat{\boldsymbol{\beta}}^{(2)},\delta) \\ &+ \hat{\boldsymbol{v}}^{(2)T}\frac{1}{|\mathcal{N}_2|}\sum_{i\in\mathcal{N}_2}A_i(\hat{\boldsymbol{\beta}}^{(1)},\delta)\Bigg), \end{aligned} \tag{4.5}$$

where

$$A_i(\hat{\boldsymbol{\beta}},\delta) = \int_u K(u)w(y_i)\frac{\boldsymbol{z}_i y_i}{b}\Big[J\big(\frac{x_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{z}_i - u\delta}{b}\big) - J\big(\frac{x_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{z}_i}{b}\big)\Big]du, \tag{4.6}$$

and $J(\cdot)$ is the aforementioned new pilot kernel function of order r with bandwidth b. Essentially, we substitute $\nabla R(u\delta,\boldsymbol{\beta}^*)$ and $\nabla R(\boldsymbol{\beta}^*)$ in (4.4) with their corresponding second smoothers through kernel function $J(\cdot)$. We now estimate the squared bias by $\hat{SB}(\delta) = \hat{B}(\delta)^2$.

To analyze theoretical properties of the estimates, let's define $\Delta = [q_1 n^{-1+\epsilon_1}, q_2 n^{-1+\epsilon_2}]$ as the range of bandwidth δ for some constants $0 < q_1 \le q_2$ and $0 < \epsilon_1 < \epsilon_2 < 1$. Since the optimal bandwidth δ^* is of order $n^{-1/(2\ell+1)}$, we can guarantee $\delta^* \in \Delta$ for some suitable ϵ_1 and ϵ_2. Under some conditions, the uniform convergence rates of $\hat{V}(\delta)$ and $\hat{SB}(\delta)$ are given by

$$|\hat{V}(\delta) - V(\delta)| \lesssim \frac{\psi_1(n,\delta)}{\delta}, \quad |\hat{SB}(\delta) - SB(\delta)| \lesssim \delta^{2\ell}\psi_2(n,\delta),$$

uniformly over all $\delta \in \Delta$, where

$$\begin{aligned} \psi_1(n,\delta) &= ||\boldsymbol{v}^*||_1^2\Bigg(\eta_2(n) \vee \sqrt{\frac{\log(n\vee d)}{n\delta}} \vee M_n\eta_1(n)\Bigg), \\ \psi_2(n,\delta) &= ||\boldsymbol{v}^*||_1^2\Bigg(\sqrt{\frac{\log(n\vee d)}{nb^{2\ell+1}}} \vee (\delta\vee b)^r \\ &\qquad \vee M_n\eta_1(n)(1\vee\frac{M_n\eta_1(n)}{\delta^\ell}) \vee \eta_2(n)\Bigg). \end{aligned}$$

We refer to Lemmas S11 and S12 in Section S2 in the supplementary materials for the formal statement of the results and further interpretations of the rates. With these two lemmas, we further establish the convergence rate of the data-driven bandwidth $\hat{\delta}$, that is, $\frac{\hat{\delta}-\delta^*}{\delta^*} \lesssim C_{n,\delta^*}$, where $C_{n,\delta^*} = \psi_1(n,\delta^*)\vee\psi_2(n,\delta^*)$, see Theorem S2 in the supplementary materials. That is, our data-driven bandwidth $\hat{\delta}$ is consistent.

For notational simplicity, let $\hat{U}_n(\delta)$ denote the bias corrected smoothed decorrelated score test statistic with a pre-specified bandwidth parameter δ. The main result in this section shows that $\hat{U}_n(\hat{\delta})$ with the data-driven bandwidth $\hat{\delta}$ is still asymptotically normal under the null hypothesis.

Theorem 3. Under the conditions in Theorem S3 in the supplementary materials and $H_0 : \theta^* = 0$, it holds that

$$\hat{U}_n(\hat{\delta}) \xrightarrow{d} N(0,1).$$

There are two major challenges in the analysis of $\hat{U}_n(\hat{\delta})$. First, the estimator $\hat{\delta}$ and the decorrelated score statistic are generally dependent with each other, which prevents the direct use of many concentration inequalities such as Bernstein's. To decouple the dependence, similar to Section 2.3, we carefully design a cross-fitting approach by splitting the data into three folds. Due to the space constraint, we leave the detailed algorithm to Section S2 in the supplementary materials. Second, different from Theorem 1 which presents the asymptotic normality of $\hat{U}_n(\delta)$ with a fixed δ, the uncertainty of $\hat{\delta}$ needs to be incorporated in the proof of Theorem 3. In particular, we use concentration inequalities to take care of the higher order error terms in the Taylor expansion of $\hat{U}_n(\hat{\delta})$ with respect to both δ and $\boldsymbol{\beta}$. This leads to much more involved analysis than that in Theorem 1.

5. Hypothesis Testing for the Linear Combination

Our results presented thus far are for the hypothesis testing (2.5), where θ is simply a single element in $\boldsymbol{\beta}$. In applications, researchers may be more interested in the hypothesis testing (2.4), a linear combination of the parameter $\boldsymbol{\beta}$. For example, as we mentioned at the beginning of Section 2.2, in the study of inferring iMCID, it is of interest to test $\boldsymbol{c}_0^T\boldsymbol{\beta}^* = 0$ where $\boldsymbol{c}_0$ represents the realized value of a new patient's clinical profile and we assume $c_{01} \neq 0$ without loss of generality. The methods

and results for (2.5) are essential. Below, we show that, in a parallel manner, all of them can be developed for (2.4) which is more applicable in scientific applications. We also point out, technically, our methods and results can be further extended to a more general null hypothesis $H_{0M}: \mathbf{M}\boldsymbol{\beta}^* = \mathbf{0}$ where $\mathbf{M} \in \mathbb{R}^{m\times d}$ and m is a fixed integer.

To test the hypothesis (2.4), consider the one to one reparameterization $(\theta, \boldsymbol{\gamma}) \to (\xi, \boldsymbol{\gamma})$, where $\xi = \boldsymbol{c}_0^T\boldsymbol{\beta}$. Under this new set of parameters, the null hypothesis can be written as $H_{0L}: \xi^* = 0$, and the smoothed surrogate loss reduces to $R_\delta^n(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma})$, where we write $\boldsymbol{c}_0 = (c_{01}, \boldsymbol{c}_{02}^T)^T$ with $\boldsymbol{c}_{02} \in \mathbb{R}^{d-1}$. Define $\boldsymbol{C} = \begin{bmatrix} \frac{1}{c_{01}} & 0 \\ \frac{-\boldsymbol{c}_{02}}{c_{01}} & \boldsymbol{I}_{d-1} \end{bmatrix} \in \mathbb{R}^{d\times d}$. From the chain rule, we can show that

$$\nabla_{(\xi,\boldsymbol{\gamma})} R_\delta^n(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma}) = \boldsymbol{C}$$
$$\nabla R_\delta^n(\theta, \boldsymbol{\gamma}), \ \nabla^2_{(\xi,\boldsymbol{\gamma}),(\xi,\boldsymbol{\gamma})} R_\delta^n(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma}) = \boldsymbol{C}\nabla^2 R_\delta^n(\theta, \boldsymbol{\gamma})\boldsymbol{C}^T, \tag{5.1}$$

and similarly for $R_\delta(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma})$ and $R(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma})$. Therefore, following the same idea as in Section 2.2, we define the smoothed decorrelated score as

$$S_\delta^L(\xi, \boldsymbol{\gamma}, \boldsymbol{\omega}_L^*) = \nabla_\xi R_\delta^n(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma}) - \boldsymbol{\omega}_L^{*T}\nabla_{\boldsymbol{\gamma}} R_\delta^n(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma}), \tag{5.2}$$

where $\boldsymbol{\omega}_L^* = \left[\nabla^2_{\boldsymbol{\gamma},\boldsymbol{\gamma}} R(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma})\right]^{-1} \nabla^2_{\boldsymbol{\gamma},\xi} R(\frac{\xi - \boldsymbol{c}_{02}^T\boldsymbol{\gamma}}{c_{01}}, \boldsymbol{\gamma})$.

We write $\boldsymbol{v}_L^* = (1, \boldsymbol{\omega}_L^{*T})^T$ and denote μ_L^*, σ_L^* as the (scaled) asymptotic bias and standard deviation of the score function $S_\delta^L(\xi, \boldsymbol{\gamma}, \boldsymbol{\omega}_L^*)$, that is,

$$\mu_L^* = \boldsymbol{v}_L^{*T}\boldsymbol{C}\boldsymbol{b}^*, \qquad \sigma_L^* = \sqrt{\boldsymbol{v}_L^{*T}\boldsymbol{C}\Sigma^*\boldsymbol{C}^T\boldsymbol{v}_L^*}, \tag{5.3}$$

where $\boldsymbol{b}^*$ and Σ^* are defined in (2.9) and (2.11), respectively. From above we can see that the estimation methods for $\boldsymbol{\omega}^*, \mu^*$ and σ^* proposed in Section 2.2 can be easily extended to obtain corresponding estimators for $\boldsymbol{\omega}_L^*, \mu_L^*$, and σ_L^*. Given these estimators, we define the test statistics for H_{0L} as

$$\hat{U}_n^L = (n\delta)^{1/2}\frac{S_\delta^L(0, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\omega}}_L) - \delta^\ell\hat{\mu}_L}{\hat{\sigma}_L}.$$

Accordingly, all the parallel results presented in Sections 3 and 4 can be developed. In the interest of space, we only present the following result that characterizes the asymptotic distribution of $\hat{U}_n^L$ under the null hypothesis H_{0L}. All other parallel results are omitted.

Theorem 4. If Assumptions 1–6 hold with $\mu^*, \sigma^*, \boldsymbol{v}^*, \hat{\boldsymbol{v}}$ substituted by $\mu_L^*, \sigma_L^*, \boldsymbol{v}_L^*, \hat{\boldsymbol{v}}_L$, and in addition $\hat{\mu}_L$ and $\hat{\sigma}_L$ are consistent estimators of μ_L^* and σ_L^*, respectively, then under the same conditions as in Theorem 1 and the null hypothesis $H_{0L}: \xi^* = 0$, it holds that $\hat{U}_n^L \xrightarrow{d} N(0, 1)$.

6. Simulation Studies

In this section, we evaluate the empirical performance of the proposed methods. Although many models can be formulated as special cases of our problem (1.2), here we mainly consider the following binary response model

$$Y = \text{sign}(X - \boldsymbol{\beta}^{*T}\boldsymbol{Z} + \epsilon), \tag{6.1}$$

where ϵ possibly depends on X and $\boldsymbol{Z}$ but the median of ϵ given X and $\boldsymbol{Z}$ is 0.

6.1. Experiments with Prespecified Bandwidth

In the first set of experiments, we evaluate the performance of the proposed test statistic with prespecified bandwidth. We use Gaussian kernel K of order 2 with bandwidth prespecified at $\delta = n^{-1/5}$. The choices of other tuning parameters are detailed in Section S5 in the supplementary materials. Throughout this section, we consider sample size $n = 800$, dimension $d = 100, 500, 1000$ and generate $\beta_2^*, \ldots, \beta_s^*$ by sampling from a uniform distribution within $[1, 2]$ for $s = 3, 10$. The first coordinate β_1^* would vary depending on the purpose of the experiment, and the rest coordinates of $\boldsymbol{\beta}^*$ are all set to 0. After that, the coefficient vector is then normalized such that $||\boldsymbol{\beta}^*||_2 = 1$. We generate $X \sim N(0, 1)$ and $\boldsymbol{Z} \sim N(0, \Sigma_\rho)$, where $(\Sigma_\rho)_{jk} = \rho^{|j-k|}$ with $\rho = 0.2, 0.5, 0.7$. For all cases, the simulations are repeated 250 times.

In the first scenario, we let $\epsilon \sim N(0, 0.2^2(1+2(X-\boldsymbol{\beta}^{*T}\boldsymbol{Z})^2))$, which is referred to as Heteroscedastic Gaussian scenario later on. We compare the proposed smoothed decorrelated score test (SDS) with the decorrelated score test method (DS) (Ning and Liu 2017) and Honest confidence region method (Honest) (Belloni et al. 2016) from the "hdm" package. We fix the significance level at 0.05 and first evaluate the performance of the tests under the null hypothesis $H_0: \beta_1^* = 0$. In this case, we set $\beta_1^* = 0$. Note that the R code for the DS and Honest approaches is tailored for the high-dimensional logistic regression, which differs from the above data-generating process.

Table 1 reports the empirical Type I error rate under the first scenario. The error rate from the SDS method is generally close to the nominal significance level 0.05, which empirically verifies the theoretical results in Theorem 1. For both Honest and DS methods, the empirical Type I error rate seems to be consistently higher or lower than the nominal level. This is expected as

Table 1. The empirical Type I error rate of the tests under the Heteroscedastic Gaussian scenario from SDS, DS and Honest methods.

		$s = 3$			$s = 10$		
d	Method	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$
100	SDS	5.6%	5.0%	6.4%	4.8%	4.8%	5.2%
	DS	1.2%	2.0%	2.0%	2.0%	1.8%	1.8%
	Honest	5.2%	5.6%	7.6%	5.4%	5.2%	6.8%
500	SDS	4.8%	4.4%	5.6%	5.6%	5.0%	4.8%
	DS	0.2%	0.4%	0.4%	0.2%	0.0%	0.4%
	Honest	7.0%	10.8%	7.6%	8.2%	6.8%	7.2%
1000	SDS	4.4%	6.0%	5.6%	5.0%	5.4%	5.0%
	DS	0.0%	0.4%	0.2%	0.0%	0.0%	0.4%
	Honest	10.0%	10.4%	12.6%	12.4%	6.4%	15.2%

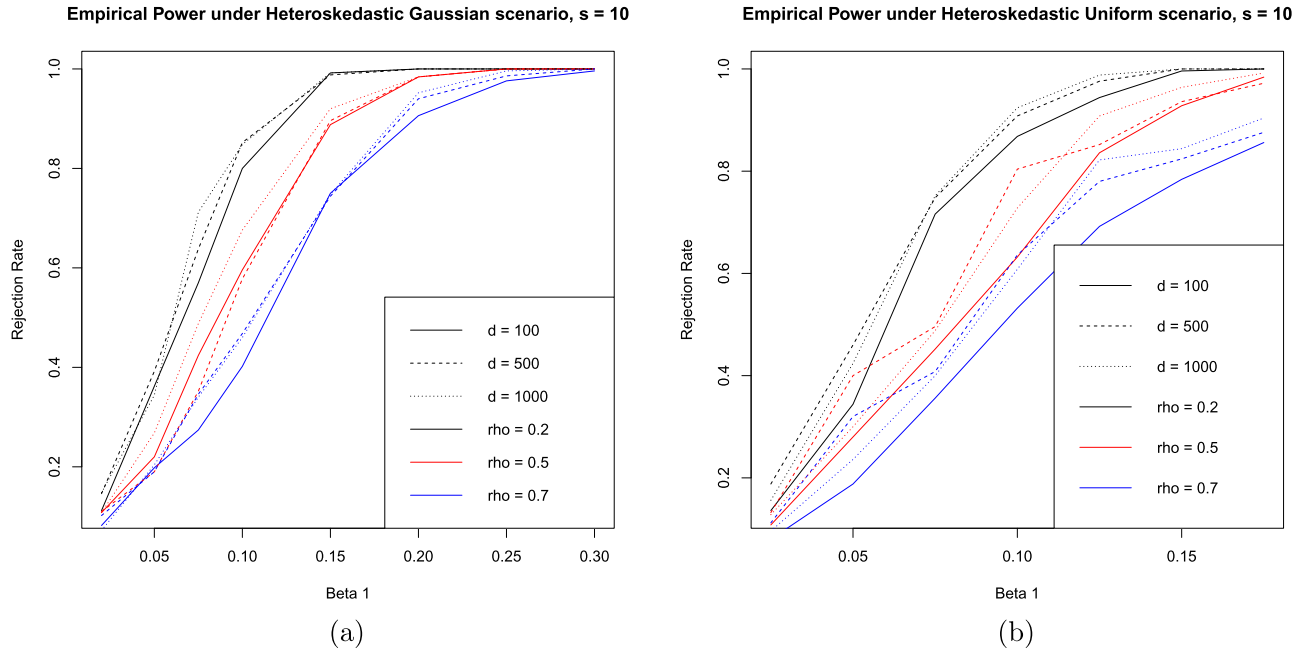

Figure 1. Empirical rejection rate of the proposed test under both scenarios with $s = 10$, $d = 100, 500, 1000$ and $\rho = 0.2, 0.5, 0.7$.

Table 2. The empirical Type I error rate of the tests under the Heteroscedastic Uniform scenario from SDS, DS, and Honest methods.

		$s = 3$			$s = 10$		
d	method	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$
100	SDS	6.0%	5.6%	6.8%	7.2%	6.8%	7.2%
	DS	2.0%	0.4%	0.8%	1.6%	0.8%	1.2%
	Honest	9.6%	17.6%	19.2%	9.6%	18.4%	21.6%
500	SDS	6.4%	6.8%	6.4%	6.0%	7.6%	7.2%
	DS	1.6%	0.8%	0.0%	0.8%	1.2%	0.4%
	Honest	11.2%	15.6%	20.0%	13.6%	17.2%	22.8%
1000	SDS	8.4%	7.6%	8.8%	9.2%	7.6%	8.0%
	DS	0.0%	0.4%	0.4%	1.2%	0.8%	0.0%
	Honest	13.6%	15.2%	22.4%	16.0%	19.2%	20.8%

these two methods only work for the logistic regression. By taking a closer look at the Normal Q-Q plot of the test statistics, we observe that the distribution of the test statistics from the Honest and DS methods deviate substantially from Gaussian, as opposed to those yield by the proposed SDS method. Please see Section S6.2 in the supplementary materials for more details.

In the second scenario, we let $\epsilon \sim 0.2 \cdot \text{Unif}(-G(X, \boldsymbol{Z}), G(X, \boldsymbol{Z}))$, where $G(x, \boldsymbol{z}) = \sqrt{1 + 2(x - \boldsymbol{\beta}^{*T}\boldsymbol{z})^2}$. In other words, the error ϵ follows a uniform distribution such that its range depends on the covariates $X, \boldsymbol{Z}$ (we will call it Heteroscedastic Uniform scenario). Similar to the Heteroscedastic Gaussian case, we compare SDS method with DS and Honest methods and study the empirical Type I error rate. From Table 2 we can see that the proposed method yields Type I error close to the nominal level as opposed to the other two. The Normal QQ-plots in Section S6.2 in the supplementary materials further confirm the asymptotic normality of our SDS test statistics. The above results suggest that in practice, if the underlying data generating process is the binary response model, our proposed approach provides valid inferential results while the existing approaches fail.

Next, we investigate the empirical power of the SDS method. We use the same data-generating processes as in the above two scenarios, but instead of setting $\beta_1^* = 0$, we vary β_1^* in the grid $\{0.02, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30\}$ for the Heteroscedastic Gaussian case, and $\{0.025, 0.05, 0.075, 0.10, 0.125, 0.15, 0.175\}$ for the Heteroscedastic Uniform case. Similarly, we consider $s = 3, 10$, $d = 100, 500, 1000$ and $\rho = 0.2, 0.5, 0.7$. Figure 1 shows the empirical rejection rate of the SDS method when $s = 10$ (see Section S6.3 for the results when $s = 3$). Note that we do not compare with the DS and Honest methods for the empirical power, because these two tests do not maintain the desired Type I error in our scenarios. We can see that for all considered cases, the empirical power converges to 1 as the magnitude of the signal β_1^* becomes larger, which agrees with Theorem 2. In addition, we find that the dimension d has minor effects on the empirical power, which is reasonable as Theorem 2 only depends on $\log d$ via the condition (3.5). Finally, we note that the power of the test deteriorates as the correlation of the design increases.

6.2. Experiments with Data-Driven Bandwidth

In the next set of experiments, we study the empirical performance of the data-driven bandwidth selection approach. We first study the Type I error and power of our SDS method for testing $H_0 : \beta_1^* = 0$ versus $H_1 : \beta_1^* \neq 0$ with data-driven bandwidth. We consider the same data generating processes as in Section 6.1 with $n = 800, d = 100, s = 3, 10$ and $\rho = 0.2, 0.5, 0.7$. We seek for the minimizer of the estimated MSE over $\delta \in [0.1, 1.2]$ and each experiment is repeated 250 times. After $\hat{\delta}$ is obtained, we plug-in it into the test statistic and estimate the bias and variance as discussed in Section S5 in the supplementary materials. With the same implementations, we also evaluate the empirical power of the test by varying β_1^* in the same grid as in Section 6.1.

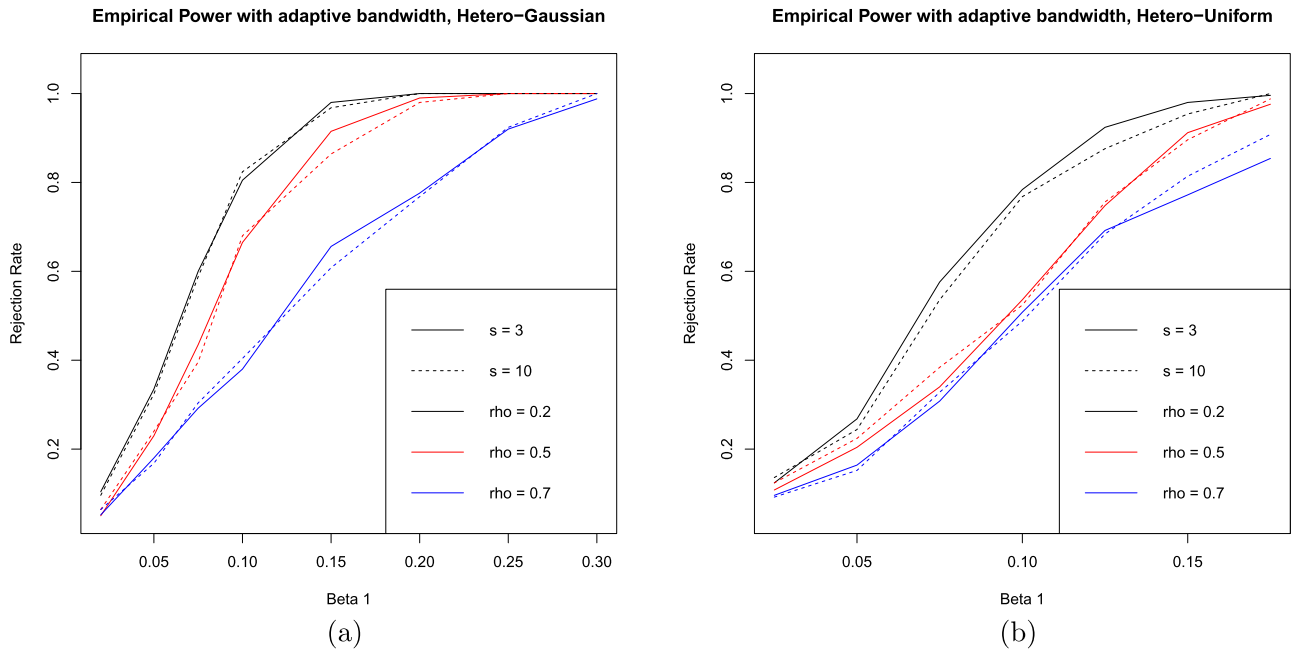

Figure 2. Empirical power of the tests under the Heteroscedastic Gaussian and Uniform scenarios with data-driven bandwidth $\hat{\delta}$.

Table 3. The empirical Type I error rate of the tests under the Heteroscedastic Gaussian and Uniform scenarios with data-driven bandwidth $\hat{\delta}$.

	$s = 3$			$s = 10$		
Data generating process	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$
Heteroscedastic Gaussian	6.8%	7.2%	5.6%	8.4%	7.2%	6.4%
Heteroscedastic Uniform	8.8%	8.0%	7.6%	8.4%	6.8%	5.2%

Table 4. The three significant variables (with p-value $< 0.05/d$) identified by the proposed SDS method, and their corresponding p-values obtained from the DS and Honest methods.

p-value	KQOL_6wk	flex_inj_pre	KSymp_3mo
SDS	3.583e-07	5.575e-05	4.482e-05
DS	0.0168	0.0340	0.0213
Honest	0.0591	0.0241	0.4383

Table 3 shows the empirical Type I error rate over 250 repetitions when $\beta_1^* = 0$, and Figure 2 shows the empirical power for different $\beta_1^* \neq 0$ in these scenarios. Similar to the case when the bandwidth δ is prespecified, the empirical Type I errors are generally close to the nominal level 0.05, and the empirical power converges to 1 as β_1^* becomes larger. We refer to Section S6 in the supplementary materials for further numerical results.

From all of these numerical results, we recommend using the data-driven bandwidth selection approach in practice.

7. Analysis of ChAMP Trial

In this section we analyze the ChAMP (Chondral Lesions And Meniscus Procedures) trial (Bisson et al. 2017), which contains clinical information about $n = 138$ patients undergoing arthroscopic partial meniscectomy (APM), a knee surgery for meniscal tears. The response variable is $Y = 1$ if the patient is healthy/satisfactory and -1 otherwise, obtained from the SF-36 survey. The continuous measurement X encodes the WOMAC pain score change from the baseline to one-year after the surgery. The dataset also contains $d = 160$ additional variables from the patient's clinical profile, denoted by $\boldsymbol{Z}$. The scientific question is to determine the iMCID, defined as a linear combination of the covariates $\boldsymbol{\beta}^T\boldsymbol{Z}$, such that the treatment of debriding chondral lesions can be claimed as clinically significant by comparing the WOMAC pain score change with this individualized threshold. As we can see, this application naturally fits into our formulation (1.2) with weight function $w(y) = 1/\mathbb{P}(Y = y)$. The goal of the analysis is to address this question by providing valid inferential results for each component of $\boldsymbol{\beta}$.

We apply the proposed SDS test for $H_{0j} : \beta_j = 0$ versus $H_{1j} : \beta_j \neq 0$, where $1 \leq j \leq d$. We use the same tuning parameter setting for estimating $\boldsymbol{\beta}^*, \omega^*$ and the asymptotic bias and variance of the score function following Section S5. For comparison, we also apply the DS and Honest methods discussed in Section 6.1.

Table 4 lists the three significant variables (i.e., those with p-value $< 0.05/d$ =3.125e-04) from the proposed SDS approach. Interestingly, all of them are clinically relevant and can provide meaningful implications for iMCID. The significance of the variable KQOL_6wk, which represents the KOOS score for quality of life at 6-week, definitely indicates how the patients recover at a relatively early stage after the surgery. The variable flex_inj_pre means the degree of flexion right before the surgery. Its significance recommends that the baseline disease severity would affect the magnitude of iMCID—this similar phenomenon was also discovered in the clinical literature for other types of diseases, such as the shoulder pain reduction study (Heald et al. 1997). The third variable KSymp_3mo is the KOOS score for other symptoms at 3-month. In some previous analysis of ChAMP trial where only estimate is available but without inference results, this variable has the second largest coefficient (Feng et al. 2022).

The results from the DS and the Honest methods are different. First, the DS method only yields 1 significant variable and the Honest method yields 13. From Table 4, the three significant variables identified by the proposed SDS method cannot be identified by either DS or Honest. In general, compared to the proposed SDS method, DS identifies fewer significant variables while Honest identifies more. This phenomenon is also evident from the simulation results in Section 6.1. On the other hand, the results from the three methods do not completely contradict with each other. For instance, the two significant variables, KQOL_6wk and KSymp_3mo, identified from the proposed SDS method, has the fourth and fifth smallest p-values in the DS method. Please refer to Section S6.7 in the supplementary materials for more results of this analysis.

In general, recall that DS and Honest methods are devised for the logistic regression, while our proposed SDS method can produce valid inference results under the binary response model (6.1), which is more flexible since the distribution of ϵ is left unspecified. Therefore, we expect that the significant variables identified by SDS are potentially more reliable and clinically more relevant. Our results presented in this session echo this rationale.

Supplemental Materials

The supplementary materials include the technical proofs, some more detailed theoretical results and discussions, and additional numerical results.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor, and two reviewers for their insightful comments which have helped improve the manuscript substantially.

Funding

Ning is supported in part by National Science Foundation (NSF) CAREER award DMS-1941945 and NSF award DMS-1854637. Zhao is supported in part by NSF award DMS-2122074 and a startup grant from the University of Wisconsin-Madison.

References

Abdullah, S. S., Rostamzadeh, N., Sedig, K., Garg, A. X., and McArthur, E. (2020), "Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records," *Informatics*, 7, 17. [1]

Angst, F., Aeschlimann, A., and Angst, J. (2017), "The Minimal Clinically Important Difference Raised the Significance of Outcome Effects above the Statistical Level, with Methodological Implications for Future Studies," *Journal of Clinical Epidemiology*, 82, 128–136. [1]

Bellamy, N., Carr, A., Dougados, M., Shea, B., and Wells, G. (2001), "Towards a Definition of "difference" in Osteoarthritis," *The Journal of Rheumatology*, 28, 427–430. [1]

Belloni, A., Chernozhukov, V., and Kato, K. (2015), "Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems," *Biometrika*, 102, 77–94. [2]

Belloni, A., Chernozhukov, V., and Wei, Y. (2016), "Post-Selection Inference for Generalized Linear Models with Many Controls," *Journal of Business & Economic Statistics*, 34, 606–619. [9]

Bisson, L. J., Kluczynski, M. A., Wind, W. M., Fineberg, M. S., Bernas, G. A., Rauh, M. A., Marzo, J. M., and Smolinski, R. J. (2015), "Design of a Randomized Controlled Trial to Compare Debridement to Observation of Chondral Lesions Encountered during Partial Meniscectomy: The ChAMP (Chondral Lesions And Meniscus Procedures) Trial," *Contemporary Clinical Trials*, 45, 281–286. [1]

Bisson, L. J., Kluczynski, M. A., Wind, W. M., Fineberg, M. S., Bernas, G. A., Rauh, M. A., Marzo, J. M., Zhou, Z., and Zhao, J. (2017), "Patient Outcomes after Observation versus Debridement of Unstable Chondral Lesions during Partial Meniscectomy: The Chondral Lesions and Meniscus Procedures (ChAMP) Randomized Controlled Trial," *The Journal of Bone and Joint Surgery*, 99, 1078–1085. [11]

Boufounos, P. T., and Baraniuk, R. G. (2008), "1-bit Compressive Sensing," in *2008 42nd Annual Conference on Information Sciences and Systems*, IEEE, pp. 16–21. [2]

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353–360. [7]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [2]

Erdogan, B. D., Leung, Y. Y., Pohl, C., Tennant, A., and Conaghan, P. G. (2016), "Minimal Clinically Important Difference as Applied in Rheumatology: An OMERACT Rasch Working Group Systematic Review and Critique," *The Journal of Rheumatology*, 43, 194–202. [1]

Fang, E. X., Ning, Y., and Li, R. (2020), "Test of Significance for High-Dimensional Longitudinal Data," *Annals of Statistics*, 48, 2622–2645. [2]

Fang, E. X., Ning, Y., and Liu, H. (2017), "Testing and Confidence Intervals for High Dimensional Proportional Hazards Models," *Journal of the Royal Statistical Society*, Series B, 79, 1415–1437. [2]

Feng, H., and Ning, Y. (2019), "High-Dimensional Mixed Graphical Model with Ordinal Data: Parameter Estimation and Statistical Inference," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 654–663. [2]

Feng, H., Ning, Y., and Zhao, J. (2022), "Nonregular and Minimax Estimation of Individualized Thresholds in High Dimension with Binary Responses," *The Annals of Statistics*, 50, 2284–2305. [2,3,5,6,11]

Hall, P., Marron, J., and Park, B. U. (1992), "Smoothed Cross-Validation," *Probability Theory and Related Fields*, 92, 1–20. [7,8]

Härdle, W., Hall, P., and Marron, J. (1992), "Regression Smoothing Parameters that are not far from their Optimum," *Journal of the American Statistical Association*, 87, 227–233. [7,8]

Heald, S. L., Riddle, D. L., and Lamb, R. L. (1997), "The Shoulder Pain and Disability Index: The Construct Validity and Responsiveness of a Region-Specific Disability Measure," *Physical Therapy*, 77, 1079–1089. [1,11]

Horowitz, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica: Journal of the Econometric Society*, 60, 505–531. [2]

Jaeschke, R., Singer, J., and Guyatt, G. H. (1989), "Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference," *Controlled Clinical Trials*, 10, 407–415. [1]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [2]

Jayadevappa, R., Cook, R., and Chhatre, S. (2017), "Minimal Important Difference to Infer Changes in Health-Related Quality of Life—A Systematic Review," *Journal of Clinical Epidemiology*, 89, 188–198. [1]

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407. [7]

Kim, J., and Pollard, D.. (1990), "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219. [2,3]

Lassere, M., van der Heijde, D., and Johnson, K. R. (2001), "Foundations of the Minimal Clinically Important Difference for Imaging," *The Journal of Rheumatology*, 28, 890–891. [1]

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228. [2]

——— (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333. [2]

McGlothlin, A. E., and Lewis, R. J. (2014), "Minimal Clinically Important Difference: Defining What Really Matters to Patients," *Journal of the American Medical Association*, 312, 1342–1343. [1]

Mukherjee, D., Banerjee, M., and Ritov, Y. (2019), "Non-Standard Asymptotics in High Dimensions: Manski's Maximum Score Estimator Revisited," arXiv preprint arXiv:1903.10063. [2]

Neumann, M. H. (1995), "Automatic Bandwidth Choice and Confidence Intervals in Nonparametric Regression," *The Annals of Statistics*, 23, 1937–1959. [8]

Neykov, M., Ning, Y., Liu, J. S., and Liu, H. (2018), "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations," *Statistical Science*, 33, 427–443. [2,4]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [2,3,4,5,9]

Norman, G. R., Sloan, J. A., and Wyrwich, K. W. (2003), "Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation," *Medical Care*, 41, 582–592. [1]

Samsa, G., Edelman, D., Rothman, M. L., Williams, G. R., Lipscomb, J., and Matchar, D. (1999), "Determining Clinically Important Differences in Health Status Measures," *Pharmacoeconomics*, 15, 141–155. [1]

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Series B, 53, 683–690. [7]

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis* (Vol. 26), Boca Raton, FL: CRC Press. [7]

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, New York: Springer. [5]

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2]

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018), "Quantile-Optimal Treatment Regimes," *Journal of the American Statistical Association*, 113, 1243–1254. [2]

Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L., Strand, V., Brooks, P., and Tugwell, P. (2001), "Minimal Clinically Important Differences: Review of Methods," *The Journal of Rheumatology*, 28, 406–412. [1]

Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., and Wolinsky, F. D. (1999a), "Linking Clinical Relevance and Statistical Significance in Evaluating Intra-Individual Changes in Health-Related Quality of Life," *Medical Care*, 37, 469–478. [1]

Wyrwich, K. W., Tierney, W. M., and Wolinsky, F. D. (1999b), "Further Evidence Supporting an SEM-based Criterion for Identifying Meaningful Intra-Individual Changes in Health-Related Quality of Life," *Journal of Clinical Epidemiology*, 52, 861–873. [1]

Xu, T., Wang, J., and Fang, Y. (2014), "A Model-Free Estimation for the Covariate-Adjusted Youden Index and its Associated Cut-Point," *Statistics in Medicine*, 33, 4963–4974. [2]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [2]