Optimal Statistical Inference for Individualized Treatment Effects in High-dimensional Models

Tianxi Cai
Harvard University, Boston, USA
T. Tony Cai
University of Pennsylvania, Philadelphia, USA
Zijian Guo
Rutgers University, Piscataway, USA

Summary. The ability to predict individualized treatment effects (ITEs) based on a given patient's profile is essential for personalized medicine. We propose a hypothesis testing approach to choosing between two potential treatments for a given individual in the framework of high-dimensional linear models. The methodological novelty lies in the construction of a debiased estimator of the ITE and establishment of its asymptotic normality uniformly for an arbitrary future high-dimensional observation, while the existing methods can only handle certain specific forms of observations. We introduce a testing procedure with the type-I error controlled and establish its asymptotic power. The proposed method can be extended to making inference for general linear contrasts, including both the average treatment effect and outcome prediction. We introduce the optimality framework for hypothesis testing from both the minimaxity and adaptivity perspectives and establish the optimality of the proposed procedure. An extension to high-dimensional approximate linear models is also considered. The finite sample performance of the procedure is demonstrated in simulation studies and further illustrated through an analysis of electronic health records data from patients with rheumatoid arthritis.

Keywords: Electronic Health Records; Personalized Medicine; Prediction; General Linear Contrasts; Confidence Intervals; Bias Correction.

1. Introduction

It has been well recognized that the effectiveness and potential risk of a treatment often vary significantly by patient subgroups. The ability to predict individualized treatment effects (ITEs) based on a given covariate profile is essential for precision medicine. Although trial-and-error and one-size-fits-all approaches remain a common practice, much recent focus has been placed on predicting treatment effects at a more individual level (La Thangue and Kerr, 2011; Ong et al., 2012). Genetic mutations and gene-expression profiles are increasingly used to guide treatment selection for cancer patients (Albain et al., 2010; Eberhard et al., 2005). Large scale clinical trials are being conducted to evaluate individualized treatment strategies (Chantrill et al., 2015; Evans and Relling, 2004; Simon et al., 2007). The increasing availability of electronic health records (EHR) systems with detailed patient data promises a new paradigm for translational precision medicine research. Models for predicting ITE can be estimated using real world data and can potentially be deployed more efficiently to clinical practice.

Motivated by the ITE estimation using EHR data with high-dimensional covariates, we consider in this paper efficient estimation and inference procedures for predicting a future patient's ITE given his/her p dimensional covariates when p is potentially much larger than the sample size n. Specifically, we consider high-dimensional linear regression models for the outcomes in the two treatment groups:

$$\mathbf{Y}_k = \mathbb{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \quad k = 1, 2, \tag{1}$$

where $\mathbf{Y}_k = (y_{k,1}, ..., y_{k,n_k})^\intercal$ and $\mathbb{X}_k = (\mathbf{X}_{k,1}, ..., \mathbf{X}_{k,n_k})^\intercal$ are the response and covariates observed independently for the n_k subjects in the treatment group k respectively, $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, ..., \epsilon_{k,n_k})^\intercal$ is the error vector with constant variance $\sigma_k^2 = \text{var}(\epsilon_{k,i})$ and $\boldsymbol{\beta}_k \in \mathbb{R}^p$ is the regression vector for the k^{th} treatment group. For a given patient with covariate vector $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$, we construct point and interval estimators for the ITE $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and consider the hypothesis testing

$$H_0 : \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \leq 0 \quad \text{vs.} \quad H_1 : \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) > 0. \tag{2}$$

1.1. Individualized Treatment Selection

While clinical trials and traditional cohort studies remain critical sources for precision medicine research, they have limitations including the generalizability of study findings and the limited ability to test broader hypotheses. In recent years, due to the increasing adoption of EHR and the linkage of EHR with bio-repositories and other research registries, integrated large datasets now exist as a new source for precision medicine studies. For example, the Partner's Healthcare System (PHS) biobank contains both a wealth of clinical (e.g. diagnoses, treatments, laboratory values) and biological measurements including genomic data (Gainer et al., 2016). These integrated datasets open opportunities for developing EHR-based individualized treatment selection models, which can potentially be fed back to the EHR system for guiding clinical decision making.

To enable EHR for such precision medicine research, different patients cohorts with specific diseases of interest have been constructed at PHS via the efforts of the Informatics for Integrating Biology and the Bedside (i2b2) (Kohane et al., 2012). An example of such disease cohort is rheumatoid arthritis (RA), consisting of 4453 patients identified as having RA using a machine learning algorithm (Liao et al., 2010). A small subset of these patients have their genetic and biological markers measured. The biomarker data integrated with EHR data can be used to derive ITE models for guiding treatment strategies for RA patients. A range of disease modifying treatment options are now available for RA patients, including methotrexate, tumor necrosis factor inhibitors often referred to as anti-TNF, and the combination of the two (Calabrese et al., 2016). The superiority of the combination therapy over monotherapy has been well established (Emery et al., 2008; Breedveld et al., 2006; van der Heijde et al., 2006). Despite its superiority, a significant proportion of patients do not respond to the combination therapy with reported response rates ranging from about 30% to 60%. Due to the high cost and significant side effects including serious infection and malignancy associated with anti-TNF therapy (Bongartz et al., 2006), there is a pressing need to develop ITE models to guide RA treatment selection. We address this need by deriving an ITE model for RA using the biomarker linked EHR data at PHS. The proposed procedures are desirable tools for application since the number of potential predictors is large in this setting.

1.2. Statistical Framework and Contributions

Many statistical and machine learning algorithms have been proposed for estimating the ITEs (Zhou et al., 2017; Zhao et al., 2012; Imai and Ratkovic, 2013; Qian and Murphy, 2011). However,

existing methods largely focused on the low-dimensional settings. In the presence of high dimensional predictors, inference for the ITEs becomes significantly more challenging. Several regularized procedures have been proposed for estimating $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ (Chen et al., 2001; Tibshirani, 1996; Fan and Li, 2001; Candès and Tao, 2007; Sun and Zhang, 2012; Zhang, 2010; Belloni et al., 2011; Moon et al., 2007; Song et al., 2015; Belloni et al., 2014). However, when the goal is to construct confidence intervals (CIs) for Δ_{new}, it is problematic to estimate Δ_{new} by simply plugging in the regularized estimators due to their inherent biases. These biases can accumulate when projecting along the direction of $\mathbf{x}_{\text{new}}$ and result in a significant bias in Δ_{new}; see Table 2 for details.

In this paper, we develop the High-dimensional Individualized Treatment Selection (HITS) method that aims to choose between two treatments for a given individual based on the observed high-dimensional covariates. We propose a novel bias-corrected estimator for $\mathbf{x}_{\text{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and establish its asymptotic normality for any given $\mathbf{x}_{\text{new}}$. This is achieved by imposing an additional novel constraint in the construction of the projection direction, which is used to correct the bias of the plug-in estimator. This additional constraint guarantees that the variance of the HITS estimator dominates its bias for any $\mathbf{x}_{\text{new}}$. With this bias-corrected estimator, we construct CIs and carry out hypothesis test for Δ_{new} under the challenging setting where $\mathbf{x}_{\text{new}}$ is of high-dimension and of no special structure. Rigorous justifications are given for the coverage and length properties of the resulted CIs and also for Type I error control and power of the proposed testing procedure. More generally, the HITS method can be adapted for making inference about any linear contrasts $\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k$ for $k = 1, 2$, which are crucial to inference for average treatment effect (ATE) and inference related to prediction; see Sections 2.4 and 5 for details. We have also extended the asymptotic normality results to high-dimensional approximate linear models. We further introduce an optimality framework for hypothesis testing in the high-dimensional sparse linear model and establish the optimality of HITS from two perspectives, minimaxity and adaptivity, where minimaxity captures the difficulty of the testing problem with true sparsity level known a priori and adaptivity is for the more challenging setting with unknown sparsity.

We summarize two key contributions of the current paper below and then compare the present work to existing high dimensional inference literature in Section 1.3.

- To the best of our knowledge, the method proposed in the current paper is the first unified inference procedure with theoretical guarantees for general linear contrasts $\mathbf{x}_{\text{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and $\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k$ for $k = 1, 2$, where no structural assumptions are made on the high-dimensional loading $\mathbf{x}_{\text{new}}$. This is a challenging task as noted in prior literature on inference for linear contrasts in high dimensional regression (Cai and Guo, 2017; Athey et al., 2018; Zhu and Bradic, 2018).
- Optimal detection boundary without knowledge of the exact sparsity level, noted as an open question in Zhu and Bradic (2017), is addressed in the current paper. It is shown that HITS is adaptively optimal for testing the hypotheses (2) over a large class of loadings $\mathbf{x}_{\text{new}}$ with unknown and unconstrained sparsity level.

1.3. Comparisons with High-dimensional Inference Literature

For a single regression coefficient under sparse linear models, Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014) introduced debiasing methods for CI construction. Inference for more general linear contrasts has been investigated recently in Cai and Guo (2017); Athey et al. (2018); Zhu and Bradic (2018). These all require special structure on the loading $\mathbf{x}_{\text{new}}$. Our work is the first to provide valid inference procedures for general contrasts with arbitrary high-dimensional loading $\mathbf{x}_{\text{new}}$ without special structures. More specifically, in the context of constructing

CIs, Cai and Guo (2017) showed a significance difference between sparse and dense $\mathbf{x}_{\text{new}}$. The methods developed for a single regression coefficient can be extended to a sparse $\mathbf{x}_{\text{new}}$ but the construction of a dense $\mathbf{x}_{\text{new}}$ relies on a conservative upper bound and requires the information on sparsity level. Athey et al. (2018) constructed CI for the general linear contrasts for $\mathbf{x}_{\text{new}}$ only if the loading $\mathbf{x}_{\text{new}}$ has a bounded weighted ℓ_2 norm and constructed CI for ATE under the *overlap* assumption; see Section 2.4 for detailed discussion. Zhu and Bradic (2018) constructed a CI for the linear contrast under the condition that the conditional expectation $\mathbb{E}[\mathbf{x}_{\text{new}}^\intercal \mathbf{X}_{1,i} \mid \mathbf{v}_1^\intercal \mathbf{X}_{1,i}, \cdots, \mathbf{v}_{p-1}^\intercal \mathbf{X}_{1,i}]$ is a sparse linear combination of $\mathbf{v}_1^\intercal \mathbf{X}_{1,i}, \cdots, \mathbf{v}_{p-1}^\intercal \mathbf{X}_{1,i}$, where $\{\mathbf{v}_j\}_{1\le j\le p-1}$ span the space orthogonal to $\mathbf{x}_{\text{new}}$. The most significant distinction of the proposed HITS method from the aforementioned literature is a unified uncertainty quantification method for all high-dimensional loadings $\mathbf{x}_{\text{new}}$.

Hypothesis testing for more general functionals has been recently considered in Javanmard and Lee (2017) and Zhu and Bradic (2017). Javanmard and Lee (2017) reduced the testing problem for a general functional to that for the projection of the functional of interest to a given orthogonal basis and then construct a debiased estimator of the corresponding basis expansion. The test statistic is constructed by comparing this debiased estimator and its projection to the null parameter space. This strategy can also be used to construct CIs for a linear contrast, but is only valid if $\mathbf{x}_{\text{new}}$ is sparse. Our proposed method is useful for more general hypothesis testing problems in high dimensions. After the first version of the present paper was released, Javanmard and Lee (2020) adopted our proposed construction for the projection direction in (8) and (9) and extended their testing procedure to handle arbitrary linear contrast; see Remark 2 of Javanmard and Lee (2020) for details. Zhu and Bradic (2017) proposed a general testing procedure by first constructing an estimator by ℓ_1 projection of the penalized estimator to the null parameter space and then debiasing both the penalized and projected estimators. The test is based on the difference between these two debiased estimators and a critical value computed via bootstrap. Although this test, in principle, controls the type I error of (2), the asymptotic power is established only when the true parameter is well separated from the null under the ℓ_∞ norm by $n^{-\frac{1}{4}}$. There are no results on the power if the true parameters in the alternative outside this region. Our approach is distinct; we establish the asymptotic normality of the proposed estimator of Δ_{new}, uniformly over all loadings $\mathbf{x}_{\text{new}}$ and the whole parameter space of approximately sparse regression vectors. As a consequence, 1) we have an asymptotic expression of the power of the proposed test for all Δ_{new}; 2) since the asymptotic power in Zhu and Bradic (2017) is established by inequalities instead of the limiting distribution, the CIs for Δ_{new} by inverting the testing procedure in Zhu and Bradic (2017) can be more conservative than the CI constructed in the present paper. Additionally, we resolved the open question raised in Zhu and Bradic (2017) "the minimax detection rate for this problem without knowledge of the sparsity level is also an open question" in Corollary 4 of the present paper. We provide more technical comparisons to these two approaches in Remark 1.

Another intuitive inference method for a general linear contrast is to plug-in the debiased estimator for individual regression coefficients developed in Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014). A numerical comparison of this estimator with the proposed HITS procedure is given in Section 6. The results show that HITS not only is computationally more efficient but also has uniformly better coverage properties than the plug-in estimator.

From another perspective, we compare the optimality results for hypothesis testing established here with those for CIs given in Cai and Guo (2017). The adaptivity framework for hypothesis testing is different from that for CI construction. In addition, the current paper considers a broader classes of loadings than those in Cai and Guo (2017), including the case of decaying loadings and

a more general class of sparse exact loadings.

1.4. Organization of the Paper

The rest of the paper is organized as follows. Section 2 presents the proposed testing and CI procedures for Δ_{new}. Theoretical properties are given in Section 3; Optimality of the testing procedure is discussed in Section 4; The proposed method is extended in Section 5 to quantify uncertainty for prediction in high dimensional linear regression; The numerical performance is investigated in Section 6. In Section 7, we apply the proposed method to infer about ITE of the aforementioned combination therapy over methotrexate alone for treating RA using EHR data from PHS. Discussions are provided in Section 8 and proofs of the main results are given in Section 9. Additional discussions, simulations and proofs are presented in the supplement (Cai et al., 2020).

1.5. Notations

For a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{X}_{i\cdot}$, $\mathbf{X}_{\cdot j}$, and $\mathbf{X}_{ij}$ denote respectively its i^{th} row, j^{th} column, and (i,j) entry. For a vector $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}_{-j}$ denotes the subvector of $\mathbf{x}$ excluding the j^{th} element, $\text{supp}(\mathbf{x})$ denotes the support of $\mathbf{x}$ and the ℓ_q norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_q = (\sum_{j=1}^p |x_j|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|\mathbf{x}\|_0 = |\text{supp}(\mathbf{x})|$ and $\|\mathbf{x}\|_\infty = \max_{1\leq j\leq p}|x_j|$. For a matrix $\mathbb{A}$, we define the spectral norm $\|\mathbb{A}\|_2 = \sup_{\|\mathbf{x}\|_2=1}\|\mathbb{A}\mathbf{x}\|_2$; For a symmetric matrix $\mathbb{A}$, $\lambda_{\min}(\mathbb{A})$ and $\lambda_{\max}(\mathbb{A})$ denote respectively the smallest and largest eigenvalue of $\mathbb{A}$. We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq Cb_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\overline{\lim}_{n\to\infty}\frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$. For a sequence of random variables X_n indexed by n, we use $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{p} X$ to represent that X_n converges to X in distribution and in probability, respectively. For $p = p(n)$, we consider the regime that $p(n) \to \infty$ with $n \to \infty$ and hence write $n \to \infty$ for $\min\{n,p\} \to \infty$.

2. Methodology

In this section, we detail proposed inference procedures for the ITE $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$. We first discuss existing bias correction methods in high-dimensional regression in Section 2.1 and introduce a novel construction of projection direction which adapts to any given loading $\mathbf{x}_{\text{new}}$ in Section 2.2, where throughout we use subscript $k \in \{1,2\}$ to index the treatment group. Then in Section 2.3, we propose point and interval estimators as well as a hypothesis testing procedure for Δ_{new}. In Section 2.4, we extend the proposed method to inference for average treatment effect.

2.1. Existing Method of Bias Correction: Minimize Variance with Bias Constrained

Given the observations $\mathbf{Y}_k \in \mathbb{R}^{n_k}$ and $\mathbb{X}_k \in \mathbb{R}^{n_k\times p}$, $\boldsymbol{\beta}_k$ can be estimated by the Lasso estimator,

$$\widehat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta}_k\in\mathbb{R}^p,\in\mathbb{R}^+}{\arg\min} \frac{\|\mathbf{Y}_k - \mathbb{X}_k\boldsymbol{\beta}_k\|_2^2}{2n_k} + A\sqrt{\frac{\log p}{n_k}}\sum_{j=1}^p W_{k,j}|\boldsymbol{\beta}_{k,j}|, \quad \text{for} \quad k = 1,2, \tag{3}$$

with a pre-specified positive constant $A > 0$ and $W_{k,j} = \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,ij}^2}$ denoting the penalization weight for the j^{th} variable in the k^{th} treatment group. The variance σ_k^2 is then estimated by

$\widehat{\sigma}_k^2 = \frac{1}{n_k}\|\mathbf{Y}_k - \mathbb{X}_k\widehat{\boldsymbol{\beta}}_k\|_2^2$ for $k = 1, 2$. We note that, other initial estimators can also be used, including the Dantzig Selector (Candès and Tao, 2007) and tuning-free penalized estimators, such as the scaled Lasso (Sun and Zhang, 2012), square-root Lasso (Belloni et al., 2011), and iterated Lasso (Belloni et al., 2012), as long as the initial estimators $\widehat{\beta}_k$ and $\widehat{\sigma}_k^2$ satisfy certain consistency properties (Conditions (B1) and (B2)) as stated in Section 3.2.

We discuss the bias correction idea for estimating $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$ and the same approach can be extended to $k = 2$. A natural and simple way to estimate $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$ is to plug in the Lasso estimator $\widehat{\boldsymbol{\beta}}_1$ in (3). However, this plug-in estimator $\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\beta}}_1$ is known to suffer from the bias induced by the penalty in (3). For the special case $\mathbf{x}_{\text{new}} = \mathbf{e}_j$, where $\mathbf{e}_j$ is the j^{th} Euclidean basis vector, various forms of debiased estimators have been introduced in Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014) to correct the bias of the plug-in estimator $\widehat{\boldsymbol{\beta}}_{1,j}$ and then construct CIs centered at the debiased estimators. The idea can be extended to general linear contrasts $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$ for certain class of $\mathbf{x}_{\text{new}}$, where a key step is to estimate the bias $\mathbf{x}_{\text{new}}^\intercal(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$. To this end, we aim to identify an effective projection direction $\mathbf{u} \in \mathbf{R}^p$ to construct a debiased estimator for $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$ as

$$\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\beta}}_1 + \mathbf{u}^\intercal\widehat{\mathbf{E}}_1, \quad \text{where} \quad \widehat{\mathbf{E}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{k,i}\left(Y_{k,i} - \mathbf{X}_{k,i}^\intercal\widehat{\boldsymbol{\beta}}_k\right) \text{ for } k = 1, 2. \tag{4}$$

The error decomposition of the above bias-corrected estimator is

$$(\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\beta}}_1 + \mathbf{u}^\intercal\widehat{\mathbf{E}}_1) - \mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1 = \mathbf{u}^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} + (\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}})^\intercal(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \tag{5}$$

where $\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{k,i}\mathbf{X}_{k,i}^\intercal$ for $k = 1, 2$.

To correct the bias of each individual regression coefficient (that is $\mathbf{x}_{\text{new}} = \mathbf{e}_j$), Zhang and Zhang (2014); Javanmard and Montanari (2014) proposed the foundational idea of constructing a projection direction for bias correction via minimization of variance with the bias constrained. Specifically, in (5), the projection direction is identified such that the variance of $\mathbf{u}^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i}$ is minimized while the bias component $(\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}})^\intercal(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$ is constrained. This idea is generalized in Cai and Guo (2017) to deal with sparse $\mathbf{x}_{\text{new}}$ via the following algorithm

$$\widetilde{\mathbf{u}}_1 = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\left\{\mathbf{u}^\intercal\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} : \ \|\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2\lambda_1\right\} \tag{6}$$

where $\lambda_1 \asymp \sqrt{\log p/n_1}$. Here, $\mathbf{u}^\intercal\widehat{\boldsymbol{\Sigma}}_1\mathbf{u}$ measures the variance of $\mathbf{u}^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i}$ and the constraint on $\|\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty$ further controls the bias term $(\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}})^\intercal(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$ in (5) via the inequality

$$|(\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}})^\intercal(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)| \le \|\widehat{\boldsymbol{\Sigma}}_1\mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_1. \tag{7}$$

The bias corrected estimator $\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\beta}}_1 + \widetilde{\mathbf{u}}_1^\intercal\widehat{\mathbf{E}}_1$ and its variant have been studied in the literature (Cai and Guo, 2017; Tripuraneni and Mackey, 2019; Athey et al., 2018). Cai and Guo (2017) and Athey et al. (2018) considered inference for $\mathbf{x}_{\text{new}}$ of specific structures. It was also shown that $\widetilde{\mathbf{u}}_1$ is effective for bias-correction when $\mathbf{x}_{\text{new}}$ is of certain sparse structure (Cai and Guo, 2017) and when $\mathbf{x}_{\text{new}}$ is of a bounded weighted ℓ_2 norm (Athey et al., 2018). Tripuraneni and Mackey (2019) focused exclusively on its estimation error instead of confidence interval construction for $\mathbf{x}_{\text{new}}^\intercal\beta$. In fact, this

type of projection direction (6), used in Cai and Guo (2017); Athey et al. (2018) and Tripuraneni and Mackey (2019), is not always effective for bias correction or the subsequent confidence interval construction. Proposition 3 of Section 8 shows that if the loading $\mathbf{x}_{\text{new}}$ is of certain dense structure, then the projection direction $\widetilde{\mathbf{u}}_1$ is zero and hence the "bias-corrected" estimator using $\widetilde{\mathbf{u}}_1$ is reduced to the plug-in estimator $\mathbf{x}_{\text{new}}^\intercal \widehat{\boldsymbol{\beta}}_1$. As shown in Table 3, the plug-in estimator can have reasonable estimation error but is not suitable for confidence interval construction due to its large bias. These existing inference procedures cannot automatically adapt to arbitrary structure of $\mathbf{x}_{\text{new}}$.

2.2. New Projection Direction: Minimize variance and Constrain Bias and Variance

To effectively debias for an arbitrary $\mathbf{x}_{\text{new}}$, we propose to identify the projection direction $\widehat{\mathbf{u}}_k$ for the k^{th} treatment as

$$\widehat{\mathbf{u}}_k = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\, \mathbf{u}^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} \quad \text{subject to } \|\widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2 \lambda_k \tag{8}$$

$$|\mathbf{x}_{\text{new}}^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2| \le \|\mathbf{x}_{\text{new}}\|_2^2 \lambda_k, \tag{9}$$

where $\lambda_k \asymp \sqrt{\log p / n_k}$. Our proposed bias corrected estimator for $\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_k$ is then

$$\widehat{\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_k} = \mathbf{x}_{\text{new}}^\intercal \widehat{\boldsymbol{\beta}}_k + \widehat{\mathbf{u}}_k^\intercal \widehat{\mathbf{E}}_k \quad \text{for} \quad k = 1, 2. \tag{10}$$

The main difference between $\widehat{\mathbf{u}}_1$ and $\widetilde{\mathbf{u}}_1$ in (6) is the additional constraint (9). As mentioned earlier, a general strategy for bias correction is to minimize the variance component while constraining the bias (Zhang and Zhang, 2014; Javanmard and Montanari, 2014). However, to develop a unified inference method for all $\mathbf{x}_{\text{new}}$, the optimization strategy developed in the current paper is a triplet, minimizing the variance, constraining the bias and constraining the variance. In particular, the additional constraint (9) is imposed to constrain the variance such that it is dominating the bias component, which is essential for CI construction. We refer to the construction defined in (8) and (9) as "Variance-enhancement Projection Direction". Such a general triplet optimization strategy can be of independent interest and applied to other inference problems.

We shall remark that, the further constraint (9) on the variance is not as intuitive as the other constraints, in the sense that the error decomposition of the bias-corrected estimator in (5) shows that it is sufficient to obtain an accurate estimator of $\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_1$ as long as we adopt the existing idea of minimizing variance under the bias constraint; see the detailed discussion between (5) and (7). In the error decomposition (5), it seems that the additional constraint (9) is not needed. However, (9) is the key component to construct a valid inference procedure uniformly over all $\mathbf{x}_{\text{new}}$. For statistical inference, one not only needs an accurate estimator, but also an accurate assessment of the uncertainty of the estimator. This is the main reason for adding the additional constraint.

An equivalent way of constructing the projection direction defined in (8) and (9) is,

$$\widehat{\mathbf{u}}_k = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\, \mathbf{u}^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} \quad \text{subject to } \sup_{\mathbf{w}\in\mathcal{C}} |\langle \mathbf{w}, \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \mathbf{x}_{\text{new}} \rangle| \le \|\mathbf{x}_{\text{new}}\|_2 \lambda_k \tag{11}$$

where $\mathcal{C} = \{\mathbf{e}_1, \cdots, \mathbf{e}_p, \mathbf{x}_{\text{new}}/\|\mathbf{x}_{\text{new}}\|_2\}$ with $\mathbf{e}_i$ denoting the ith standard Euclidean basis vector. The feasible set in (11) ensures that the projected values $\langle \mathbf{w}, \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \mathbf{x}_{\text{new}} \rangle$ of $\widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \mathbf{x}_{\text{new}}$ to any of the $p+1$ vectors in $\mathcal{C}$ are small. In contrast, as motivated in (7), the existing debiased procedures only constrain that the projected values to all the standard Euclidean basis vectors, $\max_{1\le j\le p} |\langle \mathbf{e}_j, \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} -$

$\mathbf{x}_{\text{new}}\rangle|$, are small. In the case where $\mathbf{x}_{\text{new}} = e_i$, these two constraints are the same; however, in the case where $\mathbf{x}_{\text{new}}$ is of complicated structures, the additional direction $\mathbf{x}_{\text{new}}/\|\mathbf{x}_{\text{new}}\|_2$ contained in $\mathcal{C}$ is the key component to conduct the bias correction; see more discussion in Section 8.

2.3. Statistical Inference for Individualized Treatment Effect

Combining $\widehat{\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_1}$ and $\widehat{\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_2}$, we propose to estimate Δ_{new} as

$$\widehat{\Delta_{\text{new}}} = \widehat{\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_1} - \widehat{\mathbf{x}_{\text{new}}^\intercal \boldsymbol{\beta}_2}. \tag{12}$$

Compared to the plug-in estimator $\mathbf{x}_{\text{new}}^\intercal(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)$, the main advantage of $\widehat{\Delta_{\text{new}}}$ is that the variance of $\widehat{\Delta_{\text{new}}}$ is dominating the bias of $\widehat{\Delta_{\text{new}}}$ while the bias of the plug-in estimator is usually as large as its variance. (See Table 2 for the numerical illustration.) Such a rebalance of bias and variance is useful for inference as the variance component is much easier to characterize than the bias component.

To conduct HITS, it remains to quantify the variability of $\widehat{\Delta_{\text{new}}}$. The asymptotic variance of $\widehat{\Delta_{\text{new}}}$ is

$$\mathrm{V} = \frac{\sigma_1^2}{n_1}\widehat{\mathbf{u}}_1^\intercal \widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 + \frac{\sigma_2^2}{n_2}\widehat{\mathbf{u}}_2^\intercal \widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2, \tag{13}$$

which can be estimated by $\widehat{\mathrm{V}} = \frac{\widehat{\sigma}_1^2}{n_1}\widehat{\mathbf{u}}_1^\intercal \widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 + \frac{\widehat{\sigma}_2^2}{n_2}\widehat{\mathbf{u}}_2^\intercal \widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2$. With $\widehat{\Delta_{\text{new}}}$ and the variance estimator $\widehat{\mathrm{V}}$, we may construct a CI for the ITE Δ_{new} as

$$\mathrm{CI} = \left(\widehat{\Delta_{\text{new}}} - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}}, \quad \widehat{\Delta_{\text{new}}} + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}}\right) \tag{14}$$

and conduct HITS based on

$$\phi_\alpha = \mathbf{1}\left(\widehat{\Delta_{\text{new}}} - z_\alpha\sqrt{\widehat{\mathrm{V}}} > 0\right), \tag{15}$$

where z_α is the upper α quantile for the standard normal distribution. That is, if $\widehat{\Delta_{\text{new}}} > z_\alpha\sqrt{\widehat{\mathrm{V}}}$, we would recommend subjects with $\mathbf{x}_{\text{new}}$ to receive treatment 1 over treatment 2, vice versa.

It is worth noting that the proposed HITS procedure utilizes the $\mathbf{x}_{\text{new}}$ information in the construction of the projection direction $\widehat{\mathbf{u}}_k$, where both the constraints in (8) and (9) depend on the observation $\mathbf{x}_{\text{new}}$. For observations with different $\mathbf{x}_{\text{new}}$, the corresponding projection directions can be quite different. Second, the method is computationally efficient as the bias correction step only requires the identification of two projection directions instead of performing debiased of $\widehat{\boldsymbol{\beta}}_k$ coordinate by coordinate.

2.4. Application to Inference for Average Treatment Effect

In addition to the individualized treatment selection, the proposed method can also be applied to study the average treatment effect (ATE). We follow the setting of Athey et al. (2018) where $k = 1$ corresponds to the control group and $k = 2$ corresponds to the treatment group. The average treatment over the treatment group is defined as $\bar{\mathbf{X}}_2^\intercal(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$ where $\bar{\mathbf{X}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}$. The statistical inference for the ATE $\bar{\mathbf{X}}_2^\intercal(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$ can be viewed as a special case of $x_{\text{new}}^\intercal(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$ with $x_{\text{new}} = \bar{\mathbf{X}}_2$. Both the point estimator (12) and interval estimator (14) can be directly applied here to construct point and interval estimators for the ATE by taking $x_{\text{new}} = \bar{\mathbf{X}}_2$.

These estimators are different from those proposed in Athey et al. (2018). The main distinction is the additional constraint (9) proposed in the current paper. Without (9), Athey et al. (2018)

requires either the *Bounded Loading* condition (Lemma 1 of Athey et al. (2018)) or the *Overlap* condition (Assumption 5 of Athey et al. (2018)) to guarantee the asymptotic limiting distribution of the corresponding ATE estimators. We state both conditions in the terminology of the current paper, 1) *Bounded Loading.* $\bar{\mathbf{X}}_2\Sigma_1^{-1}\bar{\mathbf{X}}_2$ is assumed to be bounded; 2) *Overlap.* There exists a constant $\eta > 0$ such that $\eta \leq e(x) \leq 1-\eta$ for all $x \in \mathbb{R}^p$ where $e(x)$ is the probability of receiving the treatment for an individual with covariates x. Both conditions are actually limiting applications of the developed method to practical settings. As $\bar{\mathbf{X}}_2\Sigma_1^{-1}\bar{\mathbf{X}}_2$ is of the order $\sqrt{p/n}$, the bounded loading condition is not realistic in the high-dimensional setting $p \gg n$. In addition, if $e(x)$ is the commonly used logit or probit probability function, then the overlap condition only holds if the support of the probability function is bounded.

3. Theoretical Properties

3.1. Model Assumptions and Initial Estimators

We assume the following conditions on the random designs and the regression errors.

(A1) For $k = 1, 2$, $\mathbf{X}_{k,i}$ are i.i.d. p-dimensional sub-gaussian random vectors with $\mathbf{\Sigma}_k = \mathbb{E}(\mathbf{X}_{k,i}\mathbf{X}_{k,i}^\intercal)$ satisfying $c_0 \leq \lambda_{\min}(\mathbf{\Sigma}_k) \leq \lambda_{\max}(\mathbf{\Sigma}_k) \leq C_0$ for positive constants $C_0 \geq c_0 > 0$. For $k = 1, 2$, the error vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, ..., \epsilon_{k,n_k})^\intercal$ is sub-gaussian and satisfies the moment conditions $\mathbb{E}(\epsilon_{k,i} \mid \mathbf{X}_{k,i}) = 0$ and $\mathbb{E}(\epsilon_{k,i}^2 \mid \mathbf{X}_{k,i}) = \sigma_k^2$ for some unknown positive constant $0 < \sigma_k^2 < \infty$. The errors $\{\epsilon_{1,i}\}_{1\leq i\leq n_1}$ are independent of $\{\epsilon_{2,i}\}_{1\leq i\leq n_2}$.

(A2) For $k = 1, 2$, the error vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, ..., \epsilon_{k,n_k})^\intercal$ is independent of $\mathbb{X}_k$ and follows Gaussian distribution with mean zero and covariance $\sigma_k^2 \cdot \mathrm{I}_{n_k}$.

The assumption (*A*1) is standard for the design and the regression error in the high-dimension literature. The condition (A2) on the error vectors is stronger but is only needed to establish the distributional results. This assumption is further relaxed in Section 3.4.

We consider the capped-ℓ_1 sparse regression vectors with

$$\sum_{j=1}^{p} \min\{|\boldsymbol{\beta}_{k,j}|/\sigma_k\lambda_0, 1\} \leq s_k \quad \text{for} \quad k = 1, 2. \tag{16}$$

where $\lambda_0 = \sqrt{2\log p/n}$. As remarked in Zhang and Zhang (2014), the capped-ℓ_1 condition in (16) holds if $\boldsymbol{\beta}_k$ is ℓ_0 sparse with $\|\boldsymbol{\beta}_k\|_0 \leq s_k$ or ℓ_q sparse with $\|\boldsymbol{\beta}_k\|_q^q/(\sigma_k\lambda_0)^q \leq s_k$ for $0 < q \leq 1$. We introduce the following general conditions on the initial estimators.

(B1) With probability larger than $1 - g(n_1, n_2)$ where $g(n_1, n_2) \to 0$ as $\min\{n_1, n_2\} \to \infty$, $\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \lesssim s_k\sqrt{\log p/n_k}$ for $k = 1, 2$,

(B2) $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ satisfy $\max_{k=1,2}\left|\widehat{\sigma}_k^2/\sigma_k^2 - 1\right| \xrightarrow{p} 0$ as $\min\{n_1, n_2\} \to \infty$.

A variety of estimators proposed in the high-dimensional regression literature for estimating the regression vector and the regression error variance are known to satisfy the above conditions under various conditions. See Sun and Zhang (2012); Belloni et al. (2011); Bickel et al. (2009); Bühlmann and van de Geer (2011) and the reference therein for more details.

3.2. Asymptotic Normality

Before stating the theorem, we present some intuitions for the estimation error of the proposed estimator, which relies on the following error decompositions of $\widehat{\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k}$,

$$\widehat{\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k} - \mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k = \widehat{\mathbf{u}}_k^{\intercal}\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{k,i}\boldsymbol{\epsilon}_{k,i} + (\widehat{\boldsymbol{\Sigma}}_k\widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}})^{\intercal}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k). \tag{17}$$

This decomposition (17) reflects the bias and variance decomposition of $\widehat{\mathbf{x}_{\text{new}}^{\intercal}\boldsymbol{\beta}_k}$, where the first error term $\widehat{\mathbf{u}}_k^{\intercal}\frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{X}_{k,i}\boldsymbol{\epsilon}_{k,i}$ is the variance while the second error term $(\widehat{\boldsymbol{\Sigma}}_k\widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}})^{\intercal}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ is the remaining stochastic bias. A similar bias and variance decomposition for the estimator $\widehat{\Delta_{\text{new}}}$ can be established by applying (17) with $k = 1, 2$. The following theorem establishes the rate of convergence for $\widehat{\Delta_{\text{new}}}$.

THEOREM 1. *Suppose that the conditions* (A1) *and* (B1) *hold and* $s_k \leq cn_k/\log p$ *for* $k = 1, 2$, *then with probability larger than* $1 - p^{-c} - g(n_1, n_2) - \frac{1}{t^2}$ *for some* $t > 1$,

$$\left|\widehat{\Delta_{\text{new}}} - \Delta_{\text{new}}\right| \lesssim t\|\mathbf{x}_{\text{new}}\|_2\left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}}\right) + \|\mathbf{x}_{\text{new}}\|_2\left(\frac{\|\boldsymbol{\beta}_1\|_0\log p}{n_1} + \frac{\|\boldsymbol{\beta}_2\|_0\log p}{n_2}\right). \tag{18}$$

One of the terms on the right hand side of (18), $\|\mathbf{x}_{\text{new}}\|_2\left(\|\boldsymbol{\beta}_1\|_0\log p/n_1 + \|\boldsymbol{\beta}_2\|_0\log p/n_2\right)$, is an upper bound for the remaining bias of the proposed debiased estimator while $\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1} + 1/\sqrt{n_2}\right)$ is an upper bound for the corresponding variance. The following theorem shows that under the additional condition (A2) and stronger sparsity conditions, the proposed estimator has an asymptotic normal distribution.

THEOREM 2. *Suppose that the conditions* (A1), (A2) *and* (B1) *hold and* $s_k \leq c\sqrt{n_k}/\log p$ *for* $k = 1, 2$, *then as* $\min\{n_1, n_2\} \to \infty$,

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\Delta_{\text{new}}} - \Delta_{\text{new}}\right) \xrightarrow{d} N(0, 1) \tag{19}$$

where V *is defined in* (13).

A key component of establishing the limiting distribution for $\widehat{\Delta_{\text{new}}}$ is to show that the standard error $\sqrt{\mathrm{V}}$ dominates the upper bound for the bias term in (18). We present this important component in the following Lemma, which characterizes the magnitude of V in (13).

LEMMA 1. *Suppose that the condition* (A1) *holds, then with probability larger than* $1 - p^{-c}$,

$$c_1\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1} + 1/\sqrt{n_2}\right) \leq \sqrt{\mathrm{V}} \leq C_1\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1} + 1/\sqrt{n_2}\right), \tag{20}$$

for some positive constants $c_1, C_1 > 0$.

We shall highlight that such a characterization of the variance, mainly the lower bound of (20), is only achieved through incorporating the novel additional constraint (9) to identify the projection direction. Without this additional constraint, as shown in Proposition 3, the variance level can be exactly zero and hence the lower bound in (20) does not hold.

3.3. Hypothesis Testing and Confidence Interval

We discuss two corollaries of Theorem 2, one for the hypothesis testing problem (2) related to the individualized treatment selection and the other for construction of CIs for Δ_{new}. Regarding the testing problem, we consider the following parameter space

$$\Theta(s) = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \mathbf{B}_1, \boldsymbol{\Sigma}_1 \\ \mathbf{B}_2, \boldsymbol{\Sigma}_2 \end{pmatrix} : \sum_{j=1}^{p} \min\left(\frac{|\boldsymbol{\beta}_{k,j}|}{\sigma_k \lambda_0}, 1 \right) \leq s,\ 0 < \sigma_k \leq M_0,\ c_0 \leq \lambda_{\min}(\boldsymbol{\Sigma}_k) \leq \lambda_{\max}(\boldsymbol{\Sigma}_k) \leq C_0,\ \text{for } k = 1, 2 \right\},$$

where $\mathbf{B}_k = (\boldsymbol{\beta}_k^{\intercal}, \sigma_k)^{\intercal}$ for $k = 1, 2$ and $M_0 > 0$ and $C_0 \geq c_0 > 0$ are some positive constants. Then we define the null hypothesis parameter space as

$$\mathcal{H}_0(s) = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \mathbf{B}_1, \boldsymbol{\Sigma}_1 \\ \mathbf{B}_2, \boldsymbol{\Sigma}_2 \end{pmatrix} \in \Theta(s) : \mathbf{x}_{\text{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \leq 0 \right\} \tag{21}$$

and the local alternative parameter space as

$$\mathcal{H}_1(s, \delta_0) = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \mathbf{B}_1, \boldsymbol{\Sigma}_1 \\ \mathbf{B}_2, \boldsymbol{\Sigma}_2 \end{pmatrix} \in \Theta(s) : \mathbf{x}_{\text{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \delta_0 \|\mathbf{x}_{\text{new}}\|_2 (1/\sqrt{n_1} + 1/\sqrt{n_2}) \right\}, \tag{22}$$

for $\delta_0 > 0$. The following corollary provides the theoretical guarantee for the individualized treatment selection, where the type I error of the proposed HITS procedure in (15) is controlled and the asymptotic power curve is also established.

COROLLARY 1. *Suppose that the conditions* (A1), (A2) *and* (B1), (B2) *hold and* $s_k \leq c\sqrt{n_k}/\log p$ *for* $k = 1, 2$ *and some positive constant* $c > 0$*, then for any* $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$*, the type I error of the proposed test* ϕ_α *defined in* (15) *is controlled as,* $\overline{\lim}_{\min\{n_1, n_2\} \to \infty} \sup_{\boldsymbol{\theta} \in \mathcal{H}_0} \mathbb{P}_{\boldsymbol{\theta}}(\phi_\alpha = 1) \leq \alpha$*. For any given* $\boldsymbol{\theta} \in \mathcal{H}_1(\delta_0)$ *and any* $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$*, the asymptotic power of the test* ϕ_α *is*

$$\lim_{\min\{n_1, n_2\} \to \infty} \mathbb{P}_{\boldsymbol{\theta}}(\phi_\alpha = 1) = 1 - \Phi^{-1}\left(z_\alpha - \frac{\delta_0}{\sqrt{\mathrm{V}}} \|\mathbf{x}_{\text{new}}\|_2 (1/\sqrt{n_1} + 1/\sqrt{n_2}) \right). \tag{23}$$

Together with Lemma 1, we observe that the proposed test is powerful with $\delta_0 \to \infty$, where δ_0 controls the local alternative defined in (22). The main massage for real applications is that the individualized treatment selection would be easier if the collected data set has larger sample sizes n_1 and n_2 and also the future observation of interest has a smaller ℓ_2 norm. This phenomenon is especially interesting for the individualized treatment selection with high-dimensional covariates, where the corresponding norm $\|\mathbf{x}_{\text{new}}\|_2$ can be of different orders of magnitude; see Section 6 for numerical illustrations.

In addition to the hypothesis testing procedure, we also establish the coverage of the proposed CI in (14) for ITE Δ_{new}:

COROLLARY 2. *Suppose that* (A1), (A2) *and* (B1), (B2) *hold and* $s_k \leq c\sqrt{n_k}/\log p$ *for* $k = 1, 2$ *and some positive constant* $c > 0$*. Then the* CI *defined in* (14) *satisfies*

$$\underline{\lim}_{\min\{n_1, n_2\} \to \infty} \mathbb{P}_{\boldsymbol{\theta}}(\Delta_{\text{new}} \in \text{CI}) \geq 1 - \alpha \quad \textit{for any } \mathbf{x}_{\text{new}} \in \mathbb{R}^p.$$

Another important perspective of CI construction is the precision of the CIs, which can be measured by the length. It follows from Lemma 1 that the length of the constructed CI in (14) is controlled at the order of magnitude $\|\mathbf{x}_{\text{new}}\|_2(1/\sqrt{n_1}+1/\sqrt{n_2})$, which means that the length depends on both the sample sizes n_1 and n_2 and also the ℓ_2 norm of the future observation $\|\mathbf{x}_{\text{new}}\|_2$. For observations with different $\mathbf{x}_{\text{new}}$, the lengths of the corresponding CIs for ITE Δ_{new} can be quite different, where the length is determined by $\|x_{\text{new}}\|_2$ and the numerical illustration is present in Section 6.

REMARK 1. It is helpful to compare some of the technical details with the related work Zhu and Bradic (2017); Javanmard and Lee (2017). The type I error control in Theorem 1 of Zhu and Bradic (2017) required stronger model complexity assumptions than Corollary 1. Specifically, using our notation, Zhu and Bradic (2017) required $\log p = o(n^{1/8})$ and $s_k \ll n^{\frac{1}{4}}/\sqrt{\log p}$ while we only need $s_k \ll \sqrt{n}/\log p$. More fundamentally, Zhu and Bradic (2017) did not establish the asymptotic limiting distribution as Theorem 2 in the present paper. Instead of using the asymptotic limiting distribution, Theorems 2 and 4 in Zhu and Bradic (2017) used inequalities to show that the asymptotic powers are close to 1 if the parameters in the alternative are well separated from the null under the ℓ_∞ norm by $n^{-\frac{1}{4}}$; as a consequence, the power function for the local neighborhood cannot be established as in (23). Javanmard and Lee (2017) required the loading to be sparse and particularly, in Lemma 2.4 , the loading $\mathbf{x}_{\text{new}}$ is required to satisfy $\mu\|u_1\|_1 < 1$ where, using our notation, $\mu = \|\mathbf{x}_{\text{new}}\|_2\sqrt{\log p/n}$ and $u_1 = \mathbf{x}_{\text{new}}$. This required the condition $\|\mathbf{x}_{\text{new}}\|_1\|\mathbf{x}_{\text{new}}\|_2 < \sqrt{n/\log p}$ on the loading $\mathbf{x}_{\text{new}}$ (see more discussion after Proposition 3 in Section 8). This stringent condition has been removed in the published version Javanmard and Lee (2020) by applying our proposed projection direction in (8) and (9); see Remark 2 of Javanmard and Lee (2020) for details.

3.4. Further Extensions: Approximately Linear Models and Non-Gaussianity

The inference results established under model (1) can be further extended to approximate linear models (Belloni et al., 2011, 2012),

$$\mathbf{Y}_k = \mathbb{X}_k\boldsymbol{\beta}_k + \mathbf{r}_k + \boldsymbol{\epsilon}_k,\ k = 1, 2, \tag{24}$$

where the high-dimensional vector $\boldsymbol{\beta}_k \in \mathbb{R}^p$ satisfies the capped ℓ_1 sparsity (16) and the approximation error $\mathbf{r}_k = (r_{k,1}, \cdots, r_{k,n_k})^\intercal \in \mathbb{R}^{n_k}$ is defined with $r_{k,i} = \mathbb{E}\left(Y_{k,i} \mid \mathbf{X}_{k,i}\right) - \mathbf{X}_{k,i}^\intercal\boldsymbol{\beta}_k$. We also relax the Gaussian error assumption (A2) through modifying construction of the projection direction as follows,

$$\begin{aligned}\widehat{\mathbf{u}}_k = \ \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\, \mathbf{u}^\intercal\widehat{\boldsymbol{\Sigma}}_k\mathbf{u} \quad \text{subject to } &\|\widehat{\boldsymbol{\Sigma}}_k\mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2\lambda_k \\ &|\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\Sigma}}_k\mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2| \le \|\mathbf{x}_{\text{new}}\|_2^2\lambda_k, \\ &\|\mathbf{X}_k\mathbf{u}\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2\tau_k,\end{aligned} \tag{25}$$

where $\lambda_k \asymp \sqrt{\log p/n_k}$ and $\sqrt{\log n_k} \lesssim \tau_k \ll \min\{\sqrt{n_1}, \sqrt{n_2}\}$. The additional constraint $\|\mathbf{X}_k\mathbf{u}\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2\tau_k$ has been proposed in Javanmard and Montanari (2014) to relax the Gaussian error assumption for establishing asymptotic normality for a single regression coefficient with $\mathbf{x}_{\text{new}} = \mathbf{e}_j$. The following result establishes the asymptotic normality of HITS under the general model (24) with a small approximation error $\mathbf{r}_k$ and non-Gaussian error $\boldsymbol{\epsilon}_k$.

PROPOSITION 1. *Suppose Condition* (A1) *holds and for* $k = 1, 2$, $s_k \le c\sqrt{n_k}/\log p$, $\|\mathbf{r}_k\|_2 \xrightarrow{p} 0$ *as* $n_k \to \infty$ *and* $\max_{k=1,2}\max_{1\le i\le n}\mathbb{E}(\epsilon_{k,i}^{2+\nu} \mid \mathbf{X}_{k,i}) \le M_0$ *for some constants* $\nu > 0$ *and* $M_0 > 0$.

For the initial estimator $\widehat{\boldsymbol{\beta}}_k$ *given in* (3) *and the projection direction* $\widehat{\mathbf{u}}^k$ *in* (25)*, the estimator* $\widehat{\Delta_{\text{new}}}$ *given in* (12) *satisfies the asymptotic limiting distribution* (19) *under the model* (24)*.*

A few remarks are in order. Firstly, the asymptotic normality in Proposition 1 holds for a broad class of estimators satisfying the condition (B1), with the Lasso estimator $\widehat{\boldsymbol{\beta}}_k$ in (3) as an example. Other examples include the iterated Lasso estimator (Belloni et al., 2012), which is tuning free and shown to satisfy this condition by Theorem 1 of Belloni et al. (2012) and Proposition 9.7 of Javanmard and Lee (2017) under the model (24) with exact sparse $\boldsymbol{\beta}_k$. Secondly, to make the effect of the approximation errors $\mathbf{r}_k$ negligible for estimating $\boldsymbol{\beta}_k$ under the model (24), the requirement is $\|\mathbf{r}_k\|_2/\|\boldsymbol{\beta}_k\|_0 \xrightarrow{p} 0$ (Belloni et al., 2012) as $n_k \to \infty$; a stronger condition $\|\mathbf{r}_k\|_2 \xrightarrow{p} 0$ is imposed to guarantee the asymptotic normality, which is needed to show that the approximation error in estimating Δ_{new} is negligible in comparison to the main term that is the asymptotically normal. Thirdly, this limiting distribution result does not require independence between $\boldsymbol{\epsilon}_k$ and $\mathbb{X}_k$ and the conditional moment conditions are sufficient for establishing the asymptotic normality. Lastly, this result can be used to construct CIs and conduct hypothesis testing as in (14) and (15), respectively, and the theoretical properties for hypothesis testing and confidence interval analogous to those in Corollaries 1 and 2 can be established.

4. Optimality for Hypothesis Testing

We establish in this section the optimality of the proposed procedure in the hypothesis testing framework from two perspectives, minimaxity and adaptivity. To simplify the presentation, we present the optimality results for the case $n_1 \asymp n_2$, denoted by n and $s_1 \asymp s_2$, denoted by s, the results can be extended to the case where n_1 (or s_1) is of a different order from n_2 (or s_2).

4.1. Optimality Framework for Hypothesis Testing: Minimaxity and Adaptivity

The performance of a testing procedure can be evaluated in terms of its size and its power. For a given null parameter space $\mathcal{H}_0(s)$, we define a set of testing procedures ϕ with the corresponding size asymptotically controlled at α, that is,

$$\mathcal{I}(s,\alpha) = \left\{ \phi : \boldsymbol{\alpha}(s,\phi) = \sup_{\theta \in \mathcal{H}_0(s)} \mathbb{E}_\theta \phi \le \alpha(1+o(1)) \right\}. \tag{26}$$

It has been shown in Corollary 1 that the proposed test $\phi_\alpha \in \mathcal{I}(s,\alpha)$ for $s \lesssim \sqrt{n}/\log p$. To investigate the power, we consider the local alternative space $\mathcal{H}_1(s,\tau) = \{\theta \in \Theta(s) : \mathbf{x}_{\text{new}}^\intercal (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \tau\}$, for a given $\tau > 0$. The power of a test ϕ over the parameter space $\mathcal{H}_1(s,\tau)$ is

$$\boldsymbol{\omega}(s,\tau,\phi) = \inf_{\theta \in \mathcal{H}_1(s,\tau)} \mathbb{E}_\theta \phi. \tag{27}$$

With a larger value of τ, the alternative parameter space is further away from the null parameter space and hence it is easier to construct a test of higher power. The minimax optimality can be reduced to identifying the smallest τ such that the size is controlled over $\mathcal{H}_0(s)$ and the corresponding power over $\mathcal{H}_1(s,\tau)$ is large, that is,

$$\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) = \arg\min_\tau \left\{ \tau : \sup_{\phi \in \mathcal{I}(s,\alpha)} \boldsymbol{\omega}(s,\tau,\phi) \ge 1-\eta \right\}, \tag{28}$$

where $\eta \in [0, 1)$ is a small positive constant controlling the type II error probability. The quantity $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ depends on $\mathbf{x}_{\text{new}}$, the sparsity level s and the constants $\alpha, \eta \in (0, 1)$. Throughout the discussion, we omit α and η in the arguments of $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ for simplicity. This quantity $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ is referred to as the minimax detection boundary of the hypothesis testing problem (2). In other words, $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ is the minimum distance such that there exists a test controlling size and achieving a good power. If a test ϕ satisfies the following property,

$$\phi \in \mathcal{I}(s, \alpha) \quad \text{and} \quad \boldsymbol{\omega}(s, \phi, \tau) \geq 1 - \eta \quad \text{for } \tau \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \tag{29}$$

then ϕ is defined to be minimax optimal. The minimax detection boundary in (28) is defined for a given sparsity level s, which is assumed to be known a priori. However, the exact sparsity level is typically unknown in practice. Hence, it is also of great importance to consider the optimality from the following two perspectives on adaptivity,

Q1. Whether it is possible to construct a test achieving the minimax detection boundary defined in (28) if the true sparsity level s is unknown.

Q2. What is the optimal procedure in the absence of accurate sparsity information?

To facilitate the definition of adaptivity, we consider two sparsity levels, $s \leq s_u$. Here s denotes the true sparsity level, which is typically not available in practice while s_u denotes the prior knowledge of an upper bound for the sparsity level. If we do not have a good prior knowledge about the sparsity level s, then s_u can be much larger than s. To answer Q1 and Q2, we assume that only the upper bound s_u is known instead of the exact sparsity level s. As a consequence, we focus on the set of tests $\mathcal{I}(s_u, \alpha)$, which is defined in (26) by replacing s with s_u. $\mathcal{I}(s_u, \alpha)$ consists of tests whose size is uniformly controlled over the parameter space $\mathcal{H}_0(s_u)$. In contrast to $\mathcal{I}(s, \alpha)$ in (26), the control of size in $\mathcal{I}(s_u, \alpha)$ is with respect to $\mathcal{H}_0(s_u)$, a larger parameter space than $\mathcal{H}_0(s)$, due to the fact that the true sparsity level s is unknown in constructing the testing procedure. Similar to (28), we define the adaptive detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ as

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) = \arg\min_{\tau} \left\{ \tau : \sup_{\phi \in \mathcal{I}(s_u, \alpha)} \boldsymbol{\omega}(s, \tau, \phi) \geq 1 - \eta \right\}. \tag{30}$$

Comparing (30) with (28), the imprecise information about the sparsity level only affects the control of size, where the power functions in (30) and (28) are evaluated over the same parameter space, $\mathcal{H}_1(s, \tau)$. If a test ϕ satisfies the following property,

$$\phi \in \mathcal{I}(s_u, \alpha) \quad \text{and} \quad \boldsymbol{\omega}(s, \tau, \phi) \geq 1 - \eta \quad \text{for } \tau \asymp \tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \tag{31}$$

then ϕ is defined to be adaptive optimal.

The quantities $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ and $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ do not depend on the specific testing procedure but mainly reflect the difficulty of the testing problem (2), which depends on the parameter space and also the loading vector $\mathbf{x}_{\text{new}}$. The question Q1 can be addressed through comparing $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ and $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$; if $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$, then the hypothesis testing problem (2) is defined to be adaptive, that is, even if one does not know the exact sparsity level, it is possible to construct a test as if the sparsity level is known; in contrast, if $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \ll \tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$, the hypothesis testing problem (2) is defined to be not adaptive. The information on the sparsity level is crucial. In this case, the adaptive detection boundary itself is of great interest as it quantifies

the effect of the knowledge of sparsity level. The question Q2 can be addressed using the adaptive detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ and an adaptive optimal test satisfying (31) would be the best that we can aim for if there is lack of accurate information on sparsity.

As a concluding remark, the minimax detection boundary characterizes the difficulty of the testing problem for the case of known sparsity level while the adaptive detection boundary characterizes a more challenging problem due to the unknown sparsity. The adaptive optimal test satisfying (31) is more useful in practice than that of a minimax optimal test because the exact sparsity level is typically unknown in applications.

4.2. Detection Boundary for Testing Problem (2)

We now demonstrate the optimality of the proposed procedure by considering the following two types of loadings $\mathbf{x}_{\text{new}}$, Exact Loading and Decaying Loading.

(E) **Exact Loading.** $\mathbf{x}_{\text{new}}$ is defined as an exact loading if it satisfies,

$$c \leq \max_{\{i:\mathbf{x}_{\text{new},i}\neq 0\}} |\mathbf{x}_{\text{new},i}| \,/\, \min_{\{i:\mathbf{x}_{\text{new},i}\neq 0\}} |\mathbf{x}_{\text{new},i}| \leq C, \tag{32}$$

for some positive constants $C \geq c > 0$. The condition (32) assumes that all non-zero coefficients of the loading vector $\mathbf{x}_{\text{new}}$ are of the same order of magnitude and hence the complexity of an exact loading $\mathbf{x}_{\text{new}}$ is captured by its sparsity level. We calibrate the sparsity levels of the regression vectors and the exact loading $\mathbf{x}_{\text{new}}$ as

$$s = p^{\gamma}, \quad s_u = p^{\gamma_u}, \quad \|\mathbf{x}_{\text{new}}\|_0 = p^{\gamma_{\text{new}}} \text{ for } \quad 0 \leq \gamma < \gamma_u \leq 1,\ 0 \leq \gamma_{\text{new}} \leq 1. \tag{33}$$

Based on the sparsity level of $\mathbf{x}_{\text{new}}$, we define the following types of loadings,

(E1) $\mathbf{x}_{\text{new}}$ is called an *exact sparse loading* if it satisfies (32) with $\gamma_{\text{new}} \leq 2\gamma$;

(E2) $\mathbf{x}_{\text{new}}$ is called an *exact dense loading* if it satisfies (32) with $\gamma_{\text{new}} > 2\gamma$.

Exact loadings are commonly seen in the genetic studies, where the loading $\mathbf{x}_{\text{new}}$ represents a specific observation's SNP and only takes the value from $\{0, 1, 2\}$.

(D) **Decaying Loading.** Let $|x_{\text{new},(1)}| \geq \left|x_{\text{new},(2)}\right| \geq \cdots \geq \left|x_{\text{new},(p)}\right|$ be the sorted coordinates of $|\mathbf{x}_{\text{new}}|$. We say that $\mathbf{x}_{\text{new}}$ is decaying at the rate δ if

$$\left|x_{\text{new},(i)}\right| \asymp i^{-\delta} \quad \text{for some constant } \delta \geq 0. \tag{34}$$

Depending on the decaying rate δ, we define the following two types of loadings,

(D1) $\mathbf{x}_{\text{new}}$ is called a *fast decaying loading* if it satisfies (34) with $\delta \geq \frac{1}{2}$;

(D2) $\mathbf{x}_{\text{new}}$ is called a *slow decaying loading* if it satisfies (34) with $0 \leq \delta < \frac{1}{2}$.

We focus on the exact loading and give a summary of the results for the decaying loading in Table 1. The detailed results about decaying loadings are deferred to Section A in the supplement (Cai et al., 2020). The following theorem establishes the lower bounds for the adaptive detection boundary for exact loadings.

THEOREM 3. *Suppose that* $s \leq s_u \lesssim n/\log p$. *We calibrate* s, s_u *and* $\|\mathbf{x}_{\text{new}}\|_0$ *by* γ, γ_u *and* γ_{new}, *respectively, as defined in* (33).

(E1) *If* $\mathbf{x}_{\text{new}}$ *is an exact sparse loading, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \frac{\sqrt{\|\mathbf{x}_{\text{new}}\|_0}\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}}; \tag{35}$$

(E2) *If* $\mathbf{x}_{\text{new}}$ *is an exact dense loading, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \begin{cases} \|x_{\text{new}}\|_\infty s_u \sqrt{\frac{\log p}{n}} & \text{if} \quad \gamma_{\text{new}} > 2\gamma_u; \\ \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} & \text{if} \quad \gamma_{\text{new}} \le 2\gamma_u. \end{cases} \tag{36}$$

We shall point out here that establishing the adaptive detection boundaries in Theorem 3 requires technical novelty. A closely related problem, adaptivity of confidence sets, has been carefully studied in the context of high-dimensional linear regression (Nickl and van de Geer, 2013; Cai and Guo, 2017, 2018a). However, it requires new technical tools to establish the adaptive detection boundaries, due to the different geometries demonstrated in Figure 1. The main idea of constructing the lower bounds in Nickl and van de Geer (2013); Cai and Guo (2017, 2018a) is illustrated in Figure 2(a), where one interior point is first chosen in the smaller parameter space $\Theta(s)$ and the corresponding least favorable set is constructed in the larger parameter space $\Theta(s_u)$ such that they are not distinguishable.

In comparison to Figure 2(a), the lower bound construction for the testing problem related to Figure 2(b) is more challenging due to the fact that the alternative parameter space $\mathcal{H}_1(s_u, \tau)$ does not contain the indifference region $0 < \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) < \tau$. A new technique, *transferring technique*, is developed for establishing the sharp lower bounds for the adaptive detection boundary. Define the index of $\mathbf{x}_{\text{new}}$ with the largest absolute value as $i_{\max} = \arg\max |\mathbf{x}_{\text{new},i}|$. In constructing the least favorable set in $\mathcal{H}_0(s_u)$, we first perturb the regression coefficients at other locations except for $i_{\max}$ and then choose the regression coefficient at $i_{\max}$ such that $\mathbf{x}_{\text{new},i_{\max}}(\boldsymbol{\beta}_{1,i_{\max}} - \boldsymbol{\beta}_{2,i_{\max}}) > 0$ and $\mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \le 0$; in construction of the corresponding least favorable set in $\mathcal{H}_1(s, \tau)$, we simply set the regression coefficient with index $i_{\max}$ to be the same as the corresponding coefficient at $i_{\max}$ in $\mathcal{H}_0(s_u)$ and set all other coefficients to be zero. The above construction is transferring the parameter space complexity from $\mathcal{H}_0(s_u)$ to $\mathcal{H}_1(s, \tau)$ by matching the regression coefficient at $i_{\max}$. Such a transferring technique can be of independent interest in establishing the adaptive detection boundaries for other inference problems.

The following corollary presents the matched upper bounds for the detection boundaries established in Theorem 3 over certain regimes.

COROLLARY 3. *Suppose that* $s \le s_u \lesssim \sqrt{n}/\log p$.

(E1) *If the loading* $\mathbf{x}_{\text{new}}$ *is an exact sparse loading, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \tag{37}$$

(E2) *If the loading* $\mathbf{x}_{\text{new}}$ *is an exact dense loading, then the results are divided into the following two cases,*

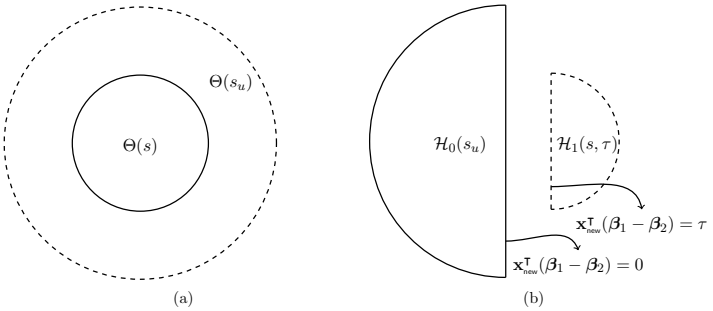


Fig. 1: (a) Null and alternative parameter spaces for the confidence set construction; (b) Null and alternative parameter spaces for the hypothesis testing problem.

(E2-a) *If* $\gamma < \gamma_u < \frac{1}{2}\gamma_{\text{new}}$, *then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_\infty s_u \sqrt{\frac{\log p}{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_\infty s \sqrt{\frac{\log p}{n}}. \tag{38}$$

(E2-b) *If* $\gamma < \frac{1}{2}\gamma_{\text{new}} \le \gamma_u$, *then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_\infty s \sqrt{\frac{\log p}{n}}. \tag{39}$$

The question Q1 about the possibility of adaptivity of the testing problem (2) can be addressed by the above corollary, where the testing problem is adaptive for the exact sparse loading case (E1) but not adaptive for the exact dense loading case (E2). The specific cut-off for the "dense" and "sparse" cases occurs at $\gamma_{\text{new}} = 2\gamma$. For the case (E2), depending on the value of γ_u, the adaptive detection boundaries can be quite different. The case (E2-a) corresponds to the case that the exact sparsity level is unknown but the upper bound s_u is relatively precise (both γ and γ_u are below $1/2{\cdot}\gamma_{\text{new}}$), then we can utilize the proposed procedure ϕ_α with the sparsity information s_u to construct a testing procedure matching the adaptive detection boundary; see the detailed construction in Section B in the supplement (Cai et al., 2020). In contrast, the case (E2-b) corresponds to the setting where the prior knowledge of the upper bound s_u is quite rough. For such a case, the proposed testing procedure ϕ_α defined in (15) achieves the adaptive detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$, but not the minimax detection boundary $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$.

Beyond answering Q1, we can also address the question Q2 with the following corollary, which considers the practical setting that there is limited information on sparsity and presents a unified optimality result for the case of exact loadings.

COROLLARY 4. *Suppose that* $s \le s_u \lesssim \sqrt{n}/\log p$ *and* $\gamma_u \ge \gamma_{\text{new}}/2$. *Then the testing procedure* ϕ_α *in* (15) *achieves the adaptive detection boundary* $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$ *for any* $\mathbf{x}_{\text{new}}$ *satisfying* (32).

The above corollary states that, in absence of accurate sparsity information, the proposed procedure ϕ_α is an adaptive optimal test for all exact loadings $\mathbf{x}_{\text{new}}$.

4.3. Comparison with Existing Optimality Results on CI

It is helpful to compare the established optimality results to the related work Cai and Guo (2017) on the minimaxity and adaptivity of confidence intervals for the linear contrast in the one-sample high-dimensional regression. Beyond the technical difference highlighted in Figure 1, we also observed the following three distinct features between the present paper and Cai and Guo (2017).

(a) The current paper closes the gap between the sparse loading regime and the dense loading regime in Cai and Guo (2017), where the lower bounds for the exact sparse loading only covered the case $\gamma_{\text{new}} \leq \gamma$ instead of the complete regime $\gamma_{\text{new}} \leq 2\gamma$ defined in this paper.

(b) In comparison to Cai and Guo (2017), the current paper considers the additional setting (E2-b), which corresponds to the case where the knowledge on the sparsity level is rough. This additional result is not only of technical interest, but has broad implications to practical applications. It addresses the important question, "what is the optimal testing procedure in a practical setting where no accurate sparsity information is available? " As shown in Corollary 4, the proposed procedure ϕ_α is an adaptive optimal test for all exact loadings $\mathbf{x}_{\text{new}}$.

(c) In addition, Theorem 4 develops the technical tools for a general loading $\mathbf{x}_{\text{new}}$, which includes the loadings not considered in Cai and Guo (2017). Specifically, we summarize in Table 1 the optimality results for the decaying loading defined in (34). As shown in Table 1, the fast decaying loading (D1) is similar to the exact sparse loading (E1) while the slow decaying loading (D2) is similar to the exact dense loading (E2). In contrast, the decaying loading has the distinct setting (D2-c) from the exact loading case where the hypotheses (2) can be tested adaptively if both γ and γ_u are above $1/2$; see the detailed discussion in Section A of the supplement (Cai et al., 2020).

δ	Setting	γ, γ_u	$\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$	Rel	$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$	Adpt
$[1/2, \infty)$	(D1)	$\gamma < \gamma_u$	$\|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$	$\asymp$	$\|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$	Yes
$[0, 1/2)$	(D2-a)	$\gamma < \gamma_u \leq \frac{1}{2}$	$s^{1-2\delta}\,(\log p)^{\frac{1}{2}-\delta}/\sqrt{n}$	$\ll$	$s_u^{1-2\delta}\,(\log p)^{\frac{1}{2}-\delta}/\sqrt{n}$	No
	(D2-b)	$\gamma < \frac{1}{2} \leq \gamma_u$	$s^{1-2\delta}\,(\log p)^{\frac{1}{2}-\delta}/\sqrt{n}$	$\ll$	$\|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$	No
	(D2-c)	$\frac{1}{2} \leq \gamma < \gamma_u$	$\|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$	$\asymp$	$\|\mathbf{x}_{\text{new}}\|_2/\sqrt{n}$	Yes

Table 1: Optimality for the decaying loading $\mathbf{x}_{\text{new}}$ defined in (34) over the regime $s \lesssim s_u \lesssim \sqrt{n}/\log p$. The column indexed with "Rel" compares $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ and $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ and the column indexed with "Adpt" reports whether the testing problem is adaptive in the corresponding setting.

5. Uncertainty Quantification related to High-dimensional Prediction

As mentioned in the introduction, the hypothesis testing method developed in the current paper can also be used for the prediction problem in a single high-dimensional regression. Consider the regression model with i.i.d observations $\{(X_{i\cdot}, y_i)\}_{1\leq i\leq n}$ satisfying

$$y_i = X_{i\cdot}^{\intercal}\boldsymbol{\beta} + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \text{ for } 1 \leq i \leq n, \tag{40}$$

and $\{\epsilon_i\}_{1\le i\le n}$ is independent of the design matrix X. The problem of interest is inference for the conditional expectation $\mathbb{E}(y_i \mid X_{i\cdot} = \mathbf{x}_{\text{new}}) = \mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}$. Uncertainty quantification for $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}$ is a one-sample version of the testing problem (15). Due to its importance and for clarity, we present a separate result on this prediction problem. We use $\widehat{\boldsymbol{\beta}}$ to denote the Lasso estimator in (3) based on the observations $\{(X_{i\cdot}, y_i)\}_{1\le i\le n}$ and construct the following bias-corrected point estimator, $\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}} = \mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}^\intercal\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i\cdot}(y_i - \mathbf{X}_{i\cdot}^\intercal\widehat{\boldsymbol{\beta}})$ with the projection direction defined as

$$\widehat{\mathbf{u}} = \underset{\mathbf{u}\in\mathbb{R}^p}{\arg\min}\, \mathbf{u}^\intercal\widehat{\boldsymbol{\Sigma}}\mathbf{u} \quad \text{subject to } \left\|\widehat{\boldsymbol{\Sigma}}\mathbf{u} - \mathbf{x}_{\text{new}}\right\|_\infty \le \|\mathbf{x}_{\text{new}}\|_2\lambda$$
$$\left|\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\Sigma}}\mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2\right| \le \|\mathbf{x}_{\text{new}}\|_2^2\lambda,$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal$ and $\lambda \asymp \sqrt{\log p/n}$. The key difference between this construction and the projection construction for the single regression coefficient in Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014); Athey et al. (2018) is the additional constraint $\left|\mathbf{x}_{\text{new}}^\intercal\widehat{\boldsymbol{\Sigma}}\mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2\right| \le \|\mathbf{x}_{\text{new}}\|_2^2\lambda$, which guarantees the asymptotic limiting distribution for any $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$. We consider the capped-ℓ_1 sparsity as in (16), $\sum_{j=1}^p \min\{|\boldsymbol{\beta}_j|/\sigma\lambda_0, 1\} \le s$, and introduce the following general condition for the initial estimator $\widehat{\boldsymbol{\beta}}$ and then establish the limiting distribution for the point estimator $\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}}$ in Corollary 5.

(P) With probability at least $1 - g(n)$ where $g(n) \to 0$ as $n \to \infty$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \lesssim s\sqrt{\log p/n}$.

COROLLARY 5. *Suppose that the regression model* (40) *holds where* $s \le c\sqrt{n}/\log p$ *and the rows* $X_{i\cdot}$ *are i.i.d. p-dimensional sub-gaussian random vectors with* $\boldsymbol{\Sigma} = \mathbb{E}(X_{i\cdot}X_{i\cdot}^\intercal)$ *satisfying* $c_0 \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le C_0$ *for positive constants* $C_0, c_0 > 0$. *For any initial estimator* $\widehat{\boldsymbol{\beta}}$ *satisfying the condition* (P), *then* $\frac{1}{\sqrt{\sigma^2\widehat{\mathbf{u}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{u}}/n}}\left(\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}} - \mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}\right) \xrightarrow{d} N(0,1)$.

Based on this corollary, we construct $\widehat{\sigma}^2 = \|y - X\widehat{\boldsymbol{\beta}}\|_2^2/n$ and use $\widehat{\mathrm{V}} = \widehat{\sigma}^2\widehat{\mathbf{u}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{u}}/n$ to estimate the variance of $\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}}$ and construct the CI, $\left(\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{\mathbf{u}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{u}}/n}, \quad \widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{\mathbf{u}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{u}}/n}\right)$. If $\widehat{\sigma}^2$ is a consistent estimator of σ^2, then this constructed CI is guaranteed to have coverage for $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}$ for any $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$. The optimality theory established in Section 4 can be easily extended to the one-sample case.

6. Simulation Studies

In this section, we present numerical studies comparing our proposed HITS method and the existing state-of-the-art methods. In Section 6.2, we consider the setting with exact sparse regression vectors and loadings $\mathbf{x}_{\text{new}}$ generated from Gaussian distributions. Results for the setting with decaying loadings are given in Section D.1 in the supplement (Cai et al., 2020). In Section 6.3, we consider settings of approximately sparse regression. Throughout, we let $p = 501$ including intercept and $n_1 = n_2 = n$ with various choices of n. For simplicity, we generate the covariates $(\mathbf{X}_{k,i})_{-1}$ from the same multivariate normal distribution with zero mean and covariance $\boldsymbol{\Sigma} = [0.5^{1+|j-l|}]_{(p-1)\times(p-1)}$.

6.1. Algorithm Implementation

We specify the tuning parameter selection in our proposed HITS algorithm. For $k = 1, 2$, the initial estimator $\widehat{\boldsymbol{\beta}}_k$ in (3) is implemented by the Lasso algorithm in `glmnet` (Friedman et al., 2010) with the tuning parameter chosen by cross validation or the tuning-free scaled Lasso algorithm in the R package `FLARE` (Li et al., 2015).

Regarding the bias correction step, we first introduce the equivalent dual form to find the projection direction defined in (8) and (9).

PROPOSITION 2. *The constrained optimizer* $\widehat{\mathbf{u}}_k \in \mathbb{R}^p$ *for* $k = 1, 2$ *defined in* (8) *and* (9) *can be computed in the form of* $\widehat{\mathbf{u}}_k = -\frac{1}{2}[\widehat{\mathbf{v}}^k_{-1} + \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\widehat{\mathbf{v}}^k_1]$, *where* $\widehat{\mathbf{v}}^k \in \mathbb{R}^{p+1}$ *is defined as*

$$\widehat{\mathbf{v}}^k = \underset{\mathbf{v}\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \frac{1}{4}\mathbf{v}^\intercal \mathbb{H}^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbb{H}\mathbf{v} + \mathbf{x}_{\text{new}}^\intercal \mathbb{H}\mathbf{v} + \lambda_k \|\mathbf{x}_{\text{new}}\|_2 \cdot \|\mathbf{v}\|_1 \right\} \tag{41}$$

with $\mathbb{H} = \left[\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}, \mathbb{I}_{p\times p}\right] \in \mathbb{R}^{p\times(p+1)}$.

Proposition 2 shows that the constrained minimization problem can be transformed to the unconstrained minimization problem in (41). In the high-dimensional setting $p > n$, the objective in the dual problem is unbounded from below if the value of $\lambda_k \geq 0$ is too small. We shall select the smallest $\lambda_k > 0$ such that the objective in the dual problem (41) is bounded from below. The code for implementing our proposed method is available on the website `https://github.com/zijguo/ITE`.

6.2. Exact Sparse Regression with General Loading

We consider the exact sparse regression in the following. To simulate $\mathbf{Y}_1$ and $\mathbf{Y}_2$, we generate $\epsilon_{k,i}$ from the standard normal and set $\beta_{1,1} = -0.1, \beta_{1,j} = -\mathbf{1}(2 \leq j \leq 11)0.4(j-1), \beta_{2,1} = -0.5$, and $\beta_{2,j} = 0.2(j-1)\mathbf{1}(2 \leq j \leq 6)$. We consider the case with the loading $\mathbf{x}_{\text{new}}$ being a dense vector, generated via two steps. In the first step, we generate $\mathbf{x}_{\text{basis}} = [1, \mathbf{x}^\intercal_{\text{basis},-1}]^\intercal \in \mathbb{R}^p$ with $\mathbf{x}_{\text{basis},-1} \sim N(0, \boldsymbol{\Sigma})$. In the second step, we generate $\mathbf{x}_{\text{new}}$ based on $\mathbf{x}_{\text{basis}}$ in two specific settings,

Setting 1: generate $\mathbf{x}_{\text{new}}$ as a shrunk version of x_{basis} with

$$x_{\text{new},j} = \mathcal{S} \cdot \mathbf{1}(j \geq 12) \cdot x_{\text{basis},j}, \quad \text{for } j = 1, ..., p \tag{42}$$

and $\mathcal{S} \in \{1, 0.5, 0.2, 0.1\}$.

Setting 2: let $x_{\text{new},j} = \mathbf{1}(j = 1) - \frac{2}{3}\mathbf{1}(j = 2) + \mathcal{S} \cdot \mathbf{1}(j \geq 12) \cdot x_{\text{basis},j}$ for $j = 1, ..., p$ and $\mathcal{S} \in \{1, 0.5, 0.2, 0.1\}$.

Under the above configurations, the scale parameter $\mathcal{S}$ controls the magnitude of the noise variables in $\mathbf{x}_{\text{new}}$. As $\mathcal{S}$ increases, $\|\mathbf{x}_{\text{new}}\|_2$ increases but Δ_{new} remains the same for all choices of $\mathcal{S}$. Setting 1 corresponds to an alternative setting with $\Delta_{\text{new}} = 1.082$ and Setting 2 corresponds to the null setting with $\Delta_{\text{new}} = 0$.

We report the simulation results based on $1,000$ replications for each setting in Tables 2 and 3. Under Setting 1, as the sample size n increases and as the magnitude of the noise variables decreases, the statistical inference problem becomes "easier" in the sense that the CI length and root mean square error (RMSE) get smaller, the empirical rejection rate (ERR) gets closer to 100%, where ERR denotes the proportion of null hypotheses being rejected out of the total $1,000$

replications and is an empirical measure of power under the alternative. This is consistent with the established theoretical results. The most challenging setting for HITS is the case with $\mathcal{S} = 1$, where the noise variables are of high magnitude. As a result, the HITS procedure has a lower power in detecting the treatment effect even when $n = 1,000$. When $\mathcal{S}$ drops to 0.2, the power of the HITS is about 72% when $n = 200$ and 95% when $n = 400$. Across all sample sizes considered including when $n = 100$, the empirical coverages of the CIs are close to the nominal level.

In Table 2, HITS is compared with the plug-in Lasso estimator (shorthanded as Lasso) and plug-in debiased estimator (shorthanded as Deb) in terms of RMSE. For Lasso, the regression vectors are estimated by the scaled Lasso in the R package FLARE (Li et al., 2015); For Deb, the regression vectors are estimated by the debiased estimators Javanmard and Montanari (2014) using the code at `https://web.stanford.edu/~montanar/sslasso/code.html`. We remark that the debiased estimators are mainly developed for inference for a single regression coefficient but not for a general linear contrast. Across all settings, HITS always outperforms Deb; in comparison to Lasso, HITS has substantially smaller bias but at the expense of larger variance, reflecting the bias-variance trade-off. Specifically, when $\mathcal{S}$ is small (taking values in $\{0.2, 0, 1\}$), HITS generally has a smaller RMSE than Lasso. When $\mathcal{S} = 1, 0.5$ in which case $\mathbf{x}_{\text{new}}$ is dense, HITS has a much larger variability compared to Lasso. This suggests that under the challenging dense scenario, a high price is paid to ensure the validity of the interval estimation.

We further compare HITS with the two plug-in estimators in the context of hypothesis testing. The Lasso estimator is not useful in hypothesis testing or CI construction due to the fact that the bias component is as large as the variance. In contrast, the variance component of both HITS and Deb dominates the corresponding bias component. Due to this reason, we only report the empirical comparison between the HITS method and the Deb method. As illustrated in the coverage property, the empirical coverage of CIs based on the Deb estimator is about 10% below the nominal level while our proposed CI achieves the nominal level. In Table 3, we report the results in Setting 2 where the null hypothesis holds. We observe a similar pattern as in Setting 1 for estimation accuracy and relative performance compared to the Lasso and Deb estimators. All the ERRs in this case, which correspond to the empirical size, are close to the nominal level 5% for the HITS method while the corresponding ERRs cannot be controlled for the Deb estimators.

6.3. Approximate Sparse Regression

For approximately sparse regression vectors, we generate the first few coefficients of β_1 and β_2 as in Section 6.2, $\beta_{1,1} = -0.1, \beta_{1,j} = -0.4(j-1)$ for $2 \le j \le 11$ and $\beta_{2,1} = -0.5$, $\beta_{2,j} = 0.2(j-1)$ for $2 \le j \le 6$ and then generate the remaining coefficients under the following two settings.

(1) **Approximate sparse with decaying coefficients:** $\beta_{1,j} = (j-1)^{-\delta_1}$ for $12 \le j \le 501$ and $\beta_{2,j} = 0.5 \cdot (j-1)^{-\delta_1}$ for $7 \le j \le 501$ and we vary δ_1 across $\{0.5, 1, 2, 3\}$.

(2) **Capped-ℓ_1 sparse:** $\beta_{1,j} = \delta_2 \cdot \lambda_0$ and $\beta_{2,j} = \beta_{1,j}/2$ for $11 \le j \le 50$ with $\lambda_0 = \sqrt{2\log p/n}$ and $\beta_{1,j} = \beta_{2,j} = 0$ for $51 \le j \le 501$. We vary δ_2 across $\{0.5, 0.2, 0.1, 0.05\}$.

For the decaying coefficients setting, the decay rate δ_1 controls the sparsity. For the capped-ℓ_1 sparse setting, the sparsity is measured by capped-ℓ_1 sparsity defined in (16), where s_k denotes the capped-ℓ_1 sparsity of β_k, for $k = 1, 2$. We vary δ_2 over $\{0.5, 0.2, 0.1, 0.05\}$ and the upper bound for s_1 ranges over $11 + \{20, 8, 4, 2\}$ and the upper bound for s_2 ranges over $6 + \{10, 4, 2, 1\}$. In both cases, we consider the dense loading $\mathbf{x}_{\text{new}}$ generated in (42) with $\mathcal{S} = 0.2$.

		ERR		Coverage		Len	HITS			Lasso			Deb		
$\mathcal{S}$	n	HITS	Deb	HITS	Deb	HITS	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
	100	0.10	0.25	0.97	0.81	9.33	2.10	0.05	2.10	0.90	0.61	0.66	2.56	0.07	2.56
	200	0.11	0.26	0.97	0.86	7.65	1.79	0.01	1.79	0.61	0.41	0.45	1.97	0.02	1.97
1	400	0.20	0.30	0.97	0.85	5.38	1.30	0.03	1.30	0.43	0.31	0.30	1.62	0.02	1.62
	600	0.23	0.36	0.97	0.87	4.47	1.02	0.07	1.02	0.34	0.24	0.24	1.34	0.05	1.34
	1000	0.34	0.32	0.97	0.85	3.49	0.83	0.02	0.83	0.26	0.19	0.19	1.49	0.05	1.49
	100	0.16	0.36	0.96	0.83	4.80	1.10	0.13	1.09	0.81	0.64	0.49	1.27	0.03	1.27
	200	0.30	0.46	0.95	0.83	3.90	0.96	0.08	0.96	0.51	0.40	0.33	1.09	0.08	1.09
0.5	400	0.49	0.57	0.94	0.86	2.74	0.67	0.09	0.67	0.34	0.25	0.22	0.82	0.05	0.81
	600	0.59	0.62	0.96	0.84	2.30	0.57	0.02	0.57	0.29	0.23	0.17	0.73	0.02	0.73
	1000	0.76	0.52	0.96	0.87	1.80	0.45	0.02	0.45	0.23	0.18	0.14	0.76	0.12	0.75
	100	0.59	0.77	0.94	0.80	2.27	0.60	0.04	0.60	0.72	0.58	0.42	0.65	0.07	0.65
	200	0.72	0.82	0.95	0.83	1.81	0.45	0.03	0.45	0.50	0.41	0.28	0.51	0.00	0.51
0.2	400	0.95	0.95	0.95	0.85	1.28	0.32	0.05	0.32	0.32	0.26	0.20	0.39	0.04	0.38
	600	0.98	0.98	0.94	0.84	1.10	0.28	0.02	0.28	0.27	0.22	0.16	0.32	0.00	0.32
	1000	1.00	0.98	0.95	0.88	0.85	0.21	0.02	0.21	0.21	0.17	0.12	0.33	0.01	0.33
	100	0.75	0.91	0.91	0.80	1.67	0.48	0.06	0.48	0.70	0.56	0.42	0.51	0.07	0.50
	200	0.94	0.97	0.93	0.80	1.29	0.35	0.01	0.35	0.49	0.40	0.29	0.38	0.05	0.38
0.1	400	1.00	1.00	0.94	0.83	0.91	0.24	0.01	0.24	0.33	0.28	0.19	0.28	0.02	0.28
	600	1.00	1.00	0.96	0.87	0.80	0.19	0.02	0.19	0.28	0.24	0.16	0.22	0.02	0.22
	1000	1.00	1.00	0.94	0.84	0.62	0.16	0.02	0.16	0.20	0.16	0.12	0.24	0.01	0.24

Table 2: Performance of HITS, in comparison with the Deb Estimator, with respect to ERR as well as the empirical coverage (Coverage) and length (Len) of the CIs under dense Setting 1 where $\Delta_{\text{new}} = 1.082$. Reported also are the RMSE, bias and the standard error (SE) of the HITS estimator compared to the Lasso and Deb estimators.

		ERR		Coverage		Len	HITS			Lasso			Deb		
$\mathcal{S}$	n	HITS	Deb	HITS	Deb	HITS	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
	100	0.02	0.10	0.98	0.83	9.18	1.97	0.20	1.96	0.56	0.16	0.53	2.37	0.21	2.36
	200	0.03	0.10	0.97	0.84	7.61	1.75	0.07	1.75	0.38	0.15	0.35	1.98	0.10	1.97
1	400	0.03	0.09	0.96	0.87	5.35	1.31	0.10	1.31	0.26	0.10	0.24	1.58	0.08	1.58
	600	0.03	0.11	0.97	0.87	4.45	1.03	0.04	1.03	0.21	0.07	0.20	1.32	0.02	1.32
	1000	0.03	0.10	0.97	0.83	3.49	0.82	0.05	0.81	0.16	0.06	0.15	1.53	0.08	1.53
	100	0.02	0.13	0.97	0.82	4.68	1.00	0.04	1.00	0.38	0.21	0.31	1.24	0.02	1.24
	200	0.04	0.13	0.97	0.84	3.82	0.91	0.04	0.91	0.26	0.15	0.21	1.02	0.02	1.02
0.5	400	0.03	0.07	0.96	0.87	2.70	0.62	0.07	0.62	0.17	0.09	0.14	0.76	0.08	0.75
	600	0.03	0.09	0.97	0.86	2.24	0.52	0.02	0.52	0.15	0.09	0.12	0.68	0.04	0.68
	1000	0.04	0.13	0.95	0.83	1.75	0.45	0.00	0.45	0.11	0.06	0.09	0.78	0.02	0.78
	100	0.06	0.18	0.97	0.80	1.96	0.46	0.11	0.44	0.33	0.22	0.24	0.53	0.09	0.52
	200	0.05	0.13	0.96	0.85	1.62	0.38	0.02	0.38	0.23	0.17	0.16	0.44	0.03	0.44
0.2	400	0.03	0.12	0.96	0.85	1.13	0.27	0.00	0.27	0.16	0.11	0.11	0.34	0.01	0.34
	600	0.03	0.08	0.96	0.88	0.94	0.22	0.01	0.22	0.13	0.09	0.09	0.27	0.01	0.27
	1000	0.03	0.09	0.97	0.88	0.74	0.18	0.01	0.18	0.10	0.07	0.07	0.31	0.01	0.31
	100	0.07	0.20	0.93	0.77	1.12	0.29	0.05	0.29	0.31	0.21	0.22	0.33	0.04	0.32
	200	0.04	0.12	0.96	0.84	0.94	0.23	0.01	0.23	0.21	0.15	0.15	0.25	0.01	0.25
0.1	400	0.04	0.11	0.96	0.87	0.66	0.16	0.00	0.16	0.16	0.12	0.10	0.19	0.01	0.19
	600	0.03	0.10	0.95	0.87	0.55	0.13	0.01	0.13	0.13	0.09	0.09	0.16	0.00	0.16
	1000	0.03	0.08	0.96	0.87	0.43	0.10	0.00	0.10	0.10	0.07	0.06	0.17	0.00	0.17

Table 3: Performance of HITS, in comparison with the Deb Estimator, with respect to ERR as well as the empirical coverage (Coverage) and length (Len) of the CIs under dense setting 2 where $\Delta_{\text{new}} = 0$. Reported also are the RMSE, bias and the standard error (SE) of the HITS estimator compared to the Lasso and Deb stimators.

The simulation results are reported in in Table 4. The results for the approximate sparse settings are largely consistent with those for the exact sparse setting. We observe that the proposed CIs achieve the 95% coverage while CIs constructed based on the Deb estimators do not have desired coverage level and the Lasso estimators have a dominant bias component. Additionally, we note that, 1) for the case that the coefficients decay slowly (the upper part of Table 4 with $\delta_1 = 0.5$), the HITS CI over-covers since the variance of the point estimator is over-estimated due to the relatively dense regression vectors; once δ_1 becomes larger, say $\delta_1 \geq 1$, the coverage is achieved at the desired level; 2) the HITS CIs are longer than those based on the Deb estimator; however, the Deb CIs do not have the correct coverage.

In addition, HITS is computationally more efficient. The average of the ratio of the computational time of Deb over that for HITS is reported under the column indexed with "TRatio ". The computational time of the Deb estimator can be as large as fifteen times that of our proposed HITS estimator. The numerical results demonstrate that, for both the exact sparse and approximate sparse settings, HITS not only has the desired coverage property for arbitrary loading $\mathbf{x}_{\text{new}}$, but also is computationally efficient.

7. Real Data Analysis

Tumor Necrosis Factor (TNF) is an inflammatory cytokine important for immunity and inflammation. TNF blockade therapy has found its success in treating RA (Taylor and Feldmann, 2009). However, the effect of anti-TNF varies greatly among patients and multiple genetic markers have been identified as predictors of anti-TNF response (Padyukov et al., 2003; Liu et al., 2008; Chatzikyriakidou et al., 2007). We seek to estimate ITE of anti-TNF in reducing inflammation for treating RA using EHR data from PHS as described in Section 1. Here, the inflammation is measured by the inflammation marker, C-reactive Protein (CRP). Since a higher value of CRP is more indicative of a worse treatment response, we define $Y = -\log \text{CRP}$.

The analyses include $n = 183$ RA patients who are free of coronary artery disease, out of which $n_1 = 92$ were on the combination therapy of anti-TNF and methotraxate and $n_2 = 91$ on methotraxate alone. To sufficiently control for potential confounders, we extracted a wide range of predictors from the EHR and included both potential confounders and predictors of CRP in $\mathbf{X}$, resulting a total of $p = 171$ predictors. Examples of predictors include diagnostic codes of RA and comorbidities such as systemic lupus erythematosus (SLE) and diabetes, past history of lab results including CRP, rheumatoid factor (RF), and anticyclic citrullinated peptide (CCP), prescriptions of other RA medications including Gold and Plaquenil, as well as counts of NLP mentions for a range of clinical terms including disease conditions and medications. Since counts of diagnosis or medication codes, referred to as codified (COD) mentions, are highly correlated with the corresponding NLP mentions in the narrative notes, we combine the counts of COD and NLP mentions of the same clinical concept to represent its frequency. The predictors also include a number of single-nucleotide polymorphism (SNP) markers and genetic risk scores identified as associated with RA risk or progression. All count variables were transformed via $x \mapsto \log(1 + x)$ and lab results were transformed by $x \mapsto \log(x)$ since their distributions are highly skewed. Missing indicator variables were created for past history of lab measurements since the availability of lab results can be indicative of disease severity. We assume that conditional on $\mathbf{X}$, the counterfactual outcomes are independent of the treatment actually received.

We applied the proposed HITS procedures to infer about the benefit of anti-TNF for individual

Approximate sparse with decaying coefficients

		ERR		Coverage		Len		HITS			Lasso			Deb			
δ_1	n	HITS	Deb	HITS	Deb	HITS	Deb	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	TRatio
0.5	100	0.01	0.50	1.00	0.76	9.99	2.77	1.06	0.19	1.04	1.28	0.95	0.87	1.16	0.06	1.16	13.58
	200	0.05	0.67	1.00	0.81	5.84	2.15	0.67	0.03	0.67	0.82	0.60	0.57	0.79	0.04	0.78	10.08
	400	0.47	0.90	1.00	0.90	3.21	1.56	0.44	0.05	0.43	0.53	0.35	0.40	0.50	0.02	0.50	4.93
	600	0.82	0.96	0.99	0.89	2.35	1.31	0.36	0.05	0.36	0.40	0.25	0.32	0.40	0.01	0.40	4.24
1	100	0.53	0.77	0.94	0.79	2.44	1.72	0.62	0.07	0.62	0.74	0.58	0.46	0.69	0.05	0.69	14.97
	200	0.77	0.88	0.97	0.85	1.91	1.44	0.46	0.04	0.46	0.47	0.38	0.28	0.50	0.07	0.50	11.98
	400	0.96	0.96	0.96	0.85	1.33	1.13	0.34	0.02	0.33	0.34	0.27	0.21	0.40	0.01	0.40	5.70
	600	0.99	0.98	0.96	0.84	1.14	0.97	0.28	0.03	0.28	0.26	0.20	0.17	0.33	0.02	0.33	4.76
2	100	0.49	0.71	0.93	0.80	2.31	1.70	0.61	0.12	0.60	0.77	0.63	0.45	0.67	0.00	0.67	14.97
	200	0.74	0.85	0.95	0.86	1.83	1.42	0.47	0.01	0.47	0.49	0.40	0.28	0.51	0.04	0.51	12.18
	400	0.96	0.96	0.94	0.87	1.29	1.12	0.32	0.03	0.32	0.33	0.27	0.19	0.39	0.04	0.39	5.78
	600	0.99	0.98	0.95	0.87	1.10	0.96	0.27	0.03	0.27	0.27	0.22	0.16	0.32	0.02	0.32	4.80
3	100	0.50	0.72	0.91	0.78	2.29	1.68	0.62	0.11	0.62	0.77	0.63	0.44	0.67	0.01	0.67	14.97
	200	0.77	0.87	0.95	0.82	1.82	1.42	0.48	0.01	0.48	0.49	0.40	0.29	0.53	0.06	0.53	12.11
	400	0.96	0.94	0.97	0.85	1.30	1.11	0.32	0.01	0.32	0.33	0.27	0.19	0.38	0.01	0.38	5.73
	600	0.99	0.99	0.95	0.86	1.10	0.96	0.27	0.03	0.27	0.27	0.22	0.16	0.33	0.02	0.33	4.83

Capped-ℓ_1 sparse

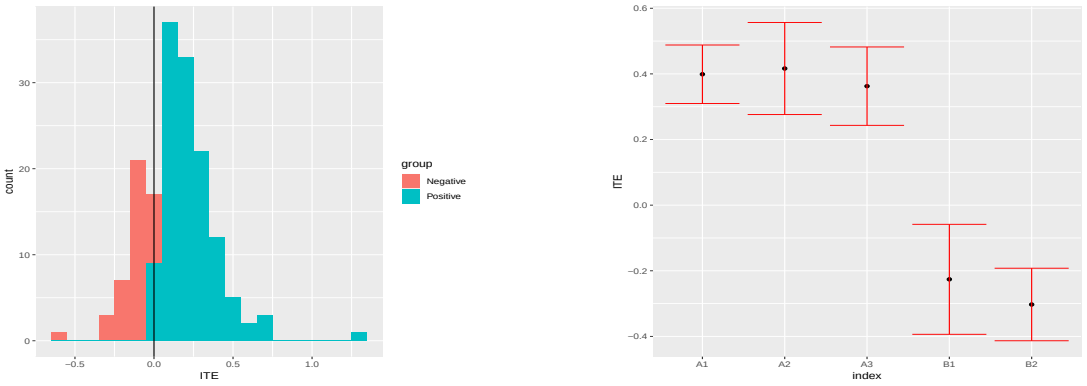
		ERR		Coverage		Len		HITS			Lasso			Deb			
δ_2	n	HITS	Deb	HITS	Deb	HITS	Deb	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	TRatio
0.5	100	0.28	0.66	0.98	0.76	3.84	1.97	0.76	0.14	0.75	0.92	0.70	0.60	0.79	0.02	0.79	13.78
	200	0.70	0.87	0.98	0.87	2.18	1.51	0.47	0.01	0.47	0.55	0.43	0.34	0.53	0.04	0.53	10.80
	400	0.97	0.96	0.97	0.87	1.39	1.16	0.33	0.05	0.32	0.33	0.26	0.21	0.40	0.04	0.40	5.27
	600	0.99	0.99	0.97	0.85	1.15	0.98	0.27	0.02	0.27	0.27	0.21	0.17	0.32	0.01	0.32	4.43
0.2	100	0.46	0.73	0.96	0.79	2.60	1.75	0.62	0.11	0.61	0.78	0.62	0.47	0.69	0.00	0.69	14.40
	200	0.76	0.87	0.96	0.85	1.91	1.44	0.47	0.03	0.47	0.51	0.41	0.31	0.52	0.06	0.52	11.49
	400	0.95	0.94	0.94	0.83	1.32	1.12	0.34	0.05	0.34	0.34	0.27	0.21	0.41	0.05	0.41	5.43
	600	0.99	0.99	0.95	0.87	1.11	0.96	0.28	0.02	0.28	0.27	0.21	0.17	0.33	0.01	0.33	4.54
0.1	100	0.50	0.71	0.95	0.79	2.41	1.70	0.61	0.12	0.59	0.78	0.64	0.44	0.65	0.02	0.65	14.50
	200	0.81	0.88	0.95	0.83	1.85	1.42	0.47	0.04	0.47	0.49	0.39	0.30	0.53	0.09	0.52	11.62
	400	0.97	0.96	0.95	0.83	1.29	1.12	0.33	0.04	0.32	0.33	0.27	0.20	0.38	0.03	0.38	5.50
	600	0.98	0.98	0.97	0.88	1.10	0.96	0.27	0.00	0.27	0.28	0.23	0.16	0.32	0.00	0.32	4.59
0.05	100	0.53	0.70	0.94	0.79	2.33	1.70	0.60	0.10	0.60	0.75	0.61	0.44	0.66	0.03	0.66	14.45
	200	0.78	0.87	0.96	0.86	1.82	1.42	0.45	0.03	0.45	0.49	0.39	0.29	0.50	0.06	0.50	11.59
	400	0.94	0.94	0.93	0.84	1.29	1.12	0.33	0.00	0.33	0.34	0.28	0.20	0.39	0.00	0.39	5.47
	600	0.99	0.98	0.96	0.87	1.10	0.96	0.27	0.01	0.27	0.28	0.22	0.16	0.33	0.01	0.33	4.63

Table 4: Performance of HITS, in comparison with the Deb Estimator, with respect to ERR as well as the empirical coverage (Coverage) and length (Len) of the CIs under approximate sparse regression settings. Reported also are the RMSE, bias and the standard error (SE) of the HITS estimator compared to the Lasso and Deb estimators; the ratio of computational time of Deb to HITS ("TRatio").

patients. Out of the $p = 171$ predictors, 8 of which were assigned with non-zero coefficients in either treatment groups. The leading predictors, as measured by the magnitude of difference between two Lasso estimators $\{\widehat{\boldsymbol{\beta}}_{1,j} - \widehat{\boldsymbol{\beta}}_{2,j}, j = 1, ..., p\}$, include counts of SLE COD or NLP mentions, indicator of no past history of CRP measurements, and SNPs including rs12506688, rs8043085 and rs2843401. Confidence intervals for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ based on debiased estimators are also reported. These predictors are generally consistent with results previously reported in clinical studies. The anti-TNF has been shown as effective among patients with presentations of both RA and SLE (Danion et al., 2017). The rs8043085 SNP located in the RASGRP1 gene is associated with an increased risk of sero-positive RA (Eyre et al., 2012) and the combination therapy has been previously reported as being more beneficial for sero-postive RA than for sero-negative RA (Seegobin et al., 2014). The rs2843401 SNP in the MMLE1 gene has been reported as protective of RA risk (Eyre et al., 2012), which appears to be associated with lower benefit of anti-TNF. The rs12506688 is in the RB-J gene which is a key upstream negative regulator of TNF-induced osteoclastogenesis.

	$\widehat{\boldsymbol{\beta}}_1$	$\text{CI}_{\boldsymbol{\beta}_1}$	$\widehat{\boldsymbol{\beta}}_2$	$\text{CI}_{\boldsymbol{\beta}_2}$	$\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2$	$\text{CI}_{\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2}$
Echo	0.02	[0.02, 0.18]	-0.03	[-0.19, -0.03]	0.04	[0.08, 0.33]
rs2843401	-0.03	[-0.07, -0.02]	0	[-0.02, 0.03]	-0.03	[-0.10, -0.01]
rs12506688	-0.08	[-0.18, -0.07]	0	[-0.03, 0.08]	-0.08	[-0.23, -0.07]
rs8043085	0	[-0.03, 0.15]	-0.05	[-0.21, -0.03]	0.05	[0.06, 0.29]
race black	0	[-0.30, 0.62]	-0.02	[-0.70, 0.22]	0.02	[-0.10, 0.90]
prior CRP missing	-0.17	[-0.7, -0.16]	0	[-0.23, 0.31]	-0.17	[-1.24, 0.30]
Gold	-0.01	[-0.13, -0.01]	0	[-0.11, 0.02]	-0.01	[-0.12, 0.06]
SLE	0	[-0.07, 0.08]	-0.16	[-0.34, -0.18]	0.16	[0.14, 0.40]

Table 5: Lasso Estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ for the predictors of CRP along with their 95% CIs. All predictors not included the table received zero Lasso estimates for both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.



(a) (b)

Fig. 2: (a) Histogram of the estimated ITE for the observed set of $\mathbf{x}_{\text{new}}$ where the vertical line represents the median value; (b) point estimate and 95% CIs for 5 choices of $\mathbf{x}_{\text{new}}$ where the x-axis indexes $\mathbf{x}_{\text{new}}^{\mathsf{T}}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)$.

We obtained estimates of Δ_{new} for the observed set of $\mathbf{x}_{\text{new}}$. As shown in Figure 2(a), the predicted

ITE ranges from -1.3 to 0.6 with median -0.14. About 72% of the patients in this population appear to benefit from combination therapy. We also obtained CIs for a few examples of $\mathbf{x}_{\text{new}}$, including (A) those with rs12506688 = 0, rs2843401 = 0, prior CRP not missing, rs8043085 > 0, and ≥ 1 SLE mention; and (B) those with rs12506688 > 0, rs2843401 > 0, prior CRP missing, rs8043085 = 0, and no SLE mention. There are three such patients in (A) (indexed by A1, A2, A3) and two in (B), (indexed by B1, B2). The point estimates and their corresponding 95% CIs are shown in Figure 2(b). The estimated ITEs were around -0.4 for A1-A3 with 95% CIs all below -0.2, suggesting that adding anti-TNF is beneficial for these patients. On the contrary, anti-TNF may even be detrimental for B1 and B2 whose estimated ITEs are 0.23 (95% CI: [0.058, 0.39]) and 0.30 (95% CI: [0.19, 0.41]), respectively. These results support prior findings that the benefit of combination therapy is heterogeneous across patients.

8. Discussions

We introduced the HITS procedure for inference on ITE with high-dimensional covariates. Both the theoretical and numerical properties of HITS are established. Unlike the debiasing methods proposed in the literature, HITS has the major advantage of not requiring the covariate vector $\mathbf{x}_{\text{new}}$ to be sparse or of other specific structures. A key innovation lies in the novel construction of the projection direction with an additional constraint (9). We elaborate the importance of this step to further illustrate the challenges of statistical inference for dense loading. The following result shows that the algorithm without (9) fails to correct the bias of $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$ for a certain class of $\mathbf{x}_{\text{new}}$.

PROPOSITION 3. *The minimizer $\widetilde{\mathbf{u}}_1$ in (6) is zero if either of the following conditions on $\mathbf{x}_{\text{new}}$ is satisfied:* (F1) $\|\mathbf{x}_{\text{new}}\|_2/\|\mathbf{x}_{\text{new}}\|_\infty \geq 1/\lambda_1$; (F2) *The non-zero coordinates of $\mathbf{x}_{\text{new}}$ are of the same order of magnitude and $\|\mathbf{x}_{\text{new}}\|_0 \geq C\sqrt{n_1/\log p}$ for some positive constant $C > 0$.*

Since $\|\mathbf{x}_{\text{new}}\|_2/\|\mathbf{x}_{\text{new}}\|_\infty$ can be viewed as a measure of sparsity of $\mathbf{x}_{\text{new}}$, both Conditions (F1) and (F2) state that the optimization algorithm (6) fails to produce a non-zero projection direction if the loading $\mathbf{x}_{\text{new}}$ is dense to certain degree. That is, without the additional constraint (9), the projection direction $\widetilde{\mathbf{u}}_1$ does not correct the bias of estimating $\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1$.

We shall provide some geometric insights about Proposition 3. The feasible set for constructing $\widetilde{\mathbf{u}}$ in (6) depends on both $\mathbf{x}_{\text{new}}$ and $\|\mathbf{x}_{\text{new}}\|_2$. If p is large and $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$ is dense, this feasible set is significantly enlarged in comparison to the feasible set corresponding to the simpler case $\mathbf{x}_{\text{new}} = \mathbf{e}_i$. As illustrated in Figure 3, the larger and smaller dashed squares represent the feasible sets for a dense $\mathbf{x}_{\text{new}}$ and $\mathbf{x}_{\text{new}} = \mathbf{e}_i$, respectively. Since zero vector is contained in the feasible set for a dense $\mathbf{x}_{\text{new}}$, the optimizer in (6) is zero and the bias correction is not effective. With the additional constraint (9), even in the presence of dense $\mathbf{x}_{\text{new}}$, the feasible set is largely shrunk to be the solid parallelogram, as the intersection of the larger dashed square and the parallel lines introduced by the constraint (9). Interestingly, the additional constraint (9) simply restricts the feasible set from one additional direction determined by $\mathbf{x}_{\text{new}}$ and automatically enables a unified inference procedure for an arbitrary $\mathbf{x}_{\text{new}}$.

The high dimensional outcome modeling adopted in HITS allows us to extensively adjust for confounders to overcome treatment by indication bias frequently encountered in observational studies. The present paper focused on the supervised setting. For EHR applications, in addition to the labeled data with outcome variables observed, there are often also a large amount of unlabeled data available where only the covariates are observed. It is known that for certain inference problems,

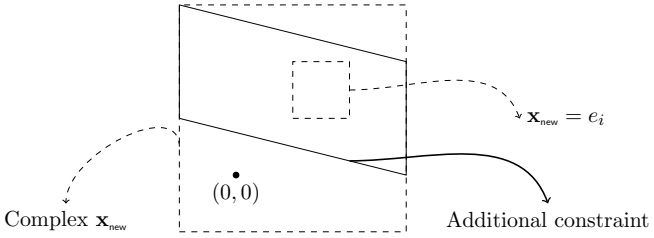


Fig. 3: Geometric illustration of Proposition 3: the solid parallelogram corresponds to the feasible set of (8) and (9) for a dense $\mathbf{x}_{\text{new}}$ while the large dashed triangle corresponds to that of (6); the small dashed square corresponds to the feasible set of (6) for $\mathbf{x}_{\text{new}} = e_i$.

the unlabeled data can be used to significantly improve the inference accuracy (Cai and Guo, 2020; Chakrabortty and Cai, 2018). Inference for ITE in the semi-supervised setting warrants future research.

9. Proofs

We present in this section the proof for the optimality results, Theorem 3 and Corollary 3 and also the proof of Lemma 1 for non-vanishing variance. For reasons of space, the proofs of all other results are deferred to Section C in the supplement (Cai et al., 2020).

9.1. Proof of Theorem 3 and Corollary 3

In the following theorem, we first introduce a general machinery for establishing the detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ for the hypothesis testing problem (2).

THEOREM 4 (DETECTION BOUNDARY LOWER BOUND). *Suppose* $s \le s_u \lesssim \min\{p, \frac{n}{\log p}\}$. *Reorder* $\mathbf{x}_{\text{new}}$ *such that* $|x_{\text{new},1}| \ge |x_{\text{new},2}| \ge \cdots \ge |x_{\text{new},p}|$. *For any* (q, L) *satisfying* $q \le s_u$ *and* $L \le \|x_{\text{new}}\|_0$, *the adaptation detection boundary* $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ *in* (30) *satisfies* $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \ge \tau^*$ *with*

$$\tau^* = C\frac{1}{\sqrt{n}} \cdot \max\left\{ \sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2}, \sum_{j=\max\{L-q+2,1\}}^{L} |x_{\text{new},j}| \sqrt{\max\{\log(cL/q^2), 0\}} \right\}. \tag{43}$$

To establish the lower bounds in Theorem 3 and Corollary 3, we simply apply the general lower bound in (43) to the case of exact loadings. Specifically, τ^* is reduced to the following expression by taking $L = \|\mathbf{x}_{\text{new}}\|_0$ and $q \asymp \min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\}$,

$$\tau^* = \frac{\|x_{\text{new}}\|_\infty}{\sqrt{n}} \cdot \max\left\{ \min\{\sqrt{s}, \sqrt{\|x_{\text{new}}\|_0}\}, \min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\} \sqrt{\max\left\{ \log\left(\frac{c\|\mathbf{x}_{\text{new}}\|_0}{\min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\}^2} \right), 0 \right\}} \right\}. \tag{44}$$

For the case (E1), we have $\|\mathbf{x}_{\text{new}}\|_0 \le s^2 \le s_u^2$ and $\tau^* \asymp \|\mathbf{x}_{\text{new}}\|_\infty \sqrt{\|\mathbf{x}_{\text{new}}\|_0/n}$; hence the lower bound (35) follows. For the case (E2), if $\gamma_{\text{new}} > 2\gamma_u$, we have $\tau^* \asymp \frac{\|x_{\text{new}}\|_\infty}{\sqrt{n}} s_u \sqrt{\log p}$; if $\gamma_{\text{new}} \le 2\gamma_u$, we have $\tau^* \asymp \frac{\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} \sqrt{\|\mathbf{x}_{\text{new}}\|_0}$. Hence, the lower bounds in (36) follow.

By applying Corollary 1, we establish that the detection boundaries $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ in (37) and (39) are achieved by the hypothesis testing procedure ϕ_α defined in (15). All the other detection boundaries will be achieved by the procedure $\phi(q, s_u)$ defined in (69) in the supplement (Cai et al., 2020).

9.2. Proof of Lemma 1

Part of (20), $\sqrt{\mathrm{V}} \le C_1\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1}+1/\sqrt{n_2}\right)$, is a consequence of the high probability concentration, $\min_{k\in\{1,2\}}\mathbf{P}\left(\|\widehat{\mathbf{\Sigma}}_k\widehat{\mathbf{u}}_k-\mathbf{x}_{\text{new}}\|_\infty \le C\|\mathbf{x}_{\text{new}}\|_2\sqrt{\log p/n_k}\right)\ge 1-p^{-c}$ which is the second high probability inequality of Lemma 4 established in Cai and Guo (2017). Hence, $\mathbf{\Sigma}_1^{-1}\mathbf{x}_{\text{new}}$ satisfies the constraints (8) and (9) and $\mathrm{V}\le \frac{\sigma_1^2}{n_1}\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_1^{-1}\widehat{\mathbf{\Sigma}}_1\mathbf{\Sigma}_1^{-1}\mathbf{x}_{\text{new}}+\frac{\sigma_2^2}{n_2}\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_2^{-1}\widehat{\mathbf{\Sigma}}_2\mathbf{\Sigma}_2^{-1}\mathbf{x}_{\text{new}}$. By Lemma 10 (specifically, the last high probability inequality) of Cai and Guo (2020), with probability larger than $1-p^{-c}$, we have

$$\left|\frac{\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_1^{-1}\widehat{\mathbf{\Sigma}}_1\mathbf{\Sigma}_1^{-1}\mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_1^{-1}\mathbf{x}_{\text{new}}}-1\right|\lesssim\sqrt{\frac{\log p}{n_1}}\ \text{ and }\ \left|\frac{\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_2^{-1}\widehat{\mathbf{\Sigma}}_2\mathbf{\Sigma}_2^{-1}\mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_2^{-1}\mathbf{x}_{\text{new}}}-1\right|\lesssim\sqrt{\frac{\log p}{n_2}} \tag{45}$$

Then we establish $\mathbb{P}\left(\sqrt{\mathrm{V}}\le C_1\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1}+1/\sqrt{n_2}\right)\right)\ge 1-p^{-c}$.

The proof of the lower bound $\sqrt{\mathrm{V}}\ge c_1\|\mathbf{x}_{\text{new}}\|_2$ is similar to that of Lemma 3.1 of Javanmard and Montanari (2014) through constructing another optimization algorithm. The main difference is that the proof in Javanmard and Montanari (2014) is for an individual regression coefficient and the following proof for a general linear contrast mainly relies on the additional constraint (9), instead of (8). To be specific, we define a proof-facilitating optimization problem,

$$\bar{\mathbf{u}}_1=\min_{u\in\mathbb{R}^p}\mathbf{u}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{u}\quad\text{subject to}\ \left|\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{u}-\|\mathbf{x}_{\text{new}}\|_2^2\right|\le\|\mathbf{x}_{\text{new}}\|_2^2\lambda_1. \tag{46}$$

Note that $\widehat{\mathbf{u}}_1$ satisfies the feasible set of (46) and hence

$$\widehat{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1\ge\bar{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\bar{\mathbf{u}}_1\ge\bar{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\bar{\mathbf{u}}_1+t((1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2-\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\bar{\mathbf{u}}_1)\ \text{for any}\ t\ge 0, \tag{47}$$

where the last inequality follows from the constraint of (46). For any given $t\ge 0$,

$$\bar{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\bar{\mathbf{u}}_1+t((1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2-\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\bar{\mathbf{u}}_1)\ge\min_{\mathbf{u}\in\mathbb{R}^p}\mathbf{u}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{u}+t((1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2-\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{u}). \tag{48}$$

By solving the minimization problem of the right hand side of (48), we have the minimizer u^* satisfies $\widehat{\mathbf{\Sigma}}_1u^*=\frac{t}{2}\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}$ and hence the minimum of the right hand side of (48) is $-\frac{t^2}{4}\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}+t(1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2$. Combined with (47) and (48), we have

$$\widehat{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1\ge\max_{t\ge0}\left[-\frac{t^2}{4}\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}+t(1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2\right]$$

and the minimum is achieved at $t^*=2\frac{(1-\lambda_1)\|\mathbf{x}_{\text{new}}\|_2^2}{\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}}>0$ and hence $\widehat{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1\ge\frac{(1-\lambda_1)^2\|\mathbf{x}_{\text{new}}\|_2^4}{\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}}$. By Lemma 10 of Cai and Guo (2020), with probability larger than $1-p^{-c}$, we have $\left|\frac{\mathbf{x}_{\text{new}}^\intercal\widehat{\mathbf{\Sigma}}_1\mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^\intercal\mathbf{\Sigma}_1\mathbf{x}_{\text{new}}}-1\right|\lesssim\sqrt{\log p/n_1}$ and hence $\widehat{\mathbf{u}}_1^\intercal\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1\ge c\|\mathbf{x}_{\text{new}}\|_2^2$. Similarly, we establish $\widehat{\mathbf{u}}_2^\intercal\widehat{\mathbf{\Sigma}}_2\widehat{\mathbf{u}}_2\ge c\|\mathbf{x}_{\text{new}}\|_2^2$ and hence $\mathbb{P}(\sqrt{\mathrm{V}}\ge c_1\|\mathbf{x}_{\text{new}}\|_2\left(1/\sqrt{n_1}+1/\sqrt{n_2}\right))\ge 1-p^{-c}$.

Acknowledgement

The research of Tianxi Cai was supported in part by NIH grants R21 CA242940 and R01 HL089778. The research of Tony Cai was supported in part by NSF grants DMS-1712735 and and DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056. The research of Zijian Guo was supported in part by NSF grants DMS-1811857, DMS-2015373 and NIH grant R01GM140463.

Supplementary Materials

Additional discussions, simulations and the remaining proofs are presented in the supplementary materials. We also provide in the supplementary materials a privacy preserving perturbed EHR data along with results from analyzing this dataset. The code for implementing our proposed method is available on the website `https://github.com/zijguo/ITE`.

References

Albain, K. S., W. E. Barlow, S. Shak, G. N. Hortobagyi, R. B. Livingston, I.-T. Yeh, P. Ravdin, R. Bugarini, F. L. Baehner, N. E. Davidson, et al. (2010). Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology 11*(1), 55–65.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc. B 80*(4), 597–623.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*(4), 791–806.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Bongartz, T., A. J. Sutton, M. J. Sweeting, I. Buchan, E. L. Matteson, and V. Montori (2006). Anti-tnf antibody therapy in rheumatoid arthritis and the risk of serious infections and malignancies: systematic review and meta-analysis of rare harmful effects in randomized controlled trials. *Journal of the American Medical Association 295*(19), 2275–2285.

Breedveld, F. C., M. H. Weisman, A. F. Kavanaugh, S. B. Cohen, K. Pavelka, R. v. Vollenhoven, J. Sharp, J. L. Perez, and G. T. Spencer-Green (2006). The premier study: a multicenter, randomized, double-blind clinical trial of combination therapy with adalimumab plus methotrexate versus methotrexate alone or adalimumab alone in patients with early, aggressive rheumatoid arthritis who had not had previous methotrexate treatment. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology 54*(1), 26–37.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T., T. T. Cai, and Z. Guo (2020). Supplement to "optimal statistical inference for individualized treatment effects in high-dimensional models".

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist. 45*(2), 615–646.

Cai, T. T. and Z. Guo (2018a). Accuracy assessment for high-dimensional linear regression. *Ann. Statist. 46*(4), 1807–1836.

Cai, T. T. and Z. Guo (2018b). Supplement to "accuracy assessment for high-dimensional linear regression".

Cai, T. T. and Z. Guo (2020). Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(2), 391–419.

Calabrese, L. H., C. Calabrese, and E. Kirchner (2016). The 2015 american college of rheumatology guideline for the treatment of rheumatoid arthritis should include new standards for hepatitis b screening: comment on the article by singh et al. *Arthritis Care & Research 68*(5), 723–724.

Candès, E. and T. Tao (2007). The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351.

Chakrabortty, A. and T. Cai (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics 46*(4), 1541–1572.

Chantrill, L. A., A. M. Nagrial, C. Watson, A. L. Johns, M. Martyn-Smith, S. Simpson, S. Mead, M. D. Jones, J. S. Samra, A. J. Gill, et al. (2015). Precision medicine for advanced pancreas cancer: the individualized molecular pancreatic cancer therapy (impact) trial. *Clinical Cancer Research 21*(9), 2029–2037.

Chatzikyriakidou, A., I. Georgiou, P. Voulgari, A. Venetsanopoulou, and A. Drosos (2007). Combined tumour necrosis factor-alpha and tumour necrosis factor receptor genotypes could predict rheumatoid arthritis patients' response to anti-tnf-alpha therapy and explain controversies of studies based on a single polymorphism. *Rheumatology (Oxford, England) 46*(6), 1034.

Chen, S. S., D. L. Donoho, and M. A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM review 43*(1), 129–159.

Danion, F., L. Sparsa, L. Arnaud, G. Alsaleh, F. Lefebvre, V. Gies, T. Martin, C. Lukas, J. Durckel, M. Ardizzone, et al. (2017). Long-term efficacy and safety of antitumour necrosis factor alpha treatment in rhupus: an open-label study of 15 patients. *RMD open 3*(2), e000555.

Eberhard, D. A., B. E. Johnson, L. C. Amler, A. D. Goddard, S. L. Heldens, R. S. Herbst, W. L. Ince, P. A. Jänne, T. Januario, D. H. Johnson, et al. (2005). Mutations in the epidermal growth factor receptor and in kras are predictive and prognostic indicators in patients with non–small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *Journal of Clinical Oncology 23*(25), 5900–5909.

Emery, P., F. C. Breedveld, S. Hall, P. Durez, D. J. Chang, D. Robertson, A. Singh, R. D. Pedersen, A. S. Koenig, and B. Freundlich (2008). Comparison of methotrexate monotherapy with a combination of methotrexate and etanercept in active, early, moderate to severe rheumatoid arthritis (comet): a randomised, double-blind, parallel treatment trial. *The Lancet 372*(9636), 375–382.

Evans, W. E. and M. V. Relling (2004). Moving towards individualized medicine with pharmacogenomics. *Nature 429*(6990), 464.

Eyre, S., J. Bowes, D. Diogo, A. Lee, A. Barton, P. Martin, A. Zhernakova, E. Stahl, S. Viatte, K. McAllister, et al. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics 44*(12), 1336.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass 96*(456), 1348–1360.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gainer, V. S., A. Cagan, V. M. Castro, S. Duey, B. Ghosh, A. P. Goodson, S. Goryachev, R. Metta, T. D. Wang, N. Wattanasin, et al. (2016). The biobank portal for partners personalized medicine: A query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *Journal of Personalized Medicine 6*(1), 11.

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Statist. 7*(1), 443–470.

Javanmard, A. and J. D. Lee (2017). A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971v3*.

Javanmard, A. and J. D. Lee (2020). A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(3), 685–718.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res. 15*(1), 2869–2909.

Kohane, I. S., S. E. Churchill, and S. N. Murphy (2012). A translational engine at the national scale: informatics for integrating biology and the bedside. *J. Am. Med. Inform. Assn 19*(2), 181–185.

La Thangue, N. B. and D. J. Kerr (2011). Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nature Reviews Clinical Oncology 8*(10), 587–596.

Li, X., T. Zhao, X. Yuan, and H. Liu (2015). The flare package for high dimensional linear regression and precision matrix estimation in r. *J. Mach. Learn. Res. 16*(1), 553–557.

Liao, K. P., T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, I. Kohane, et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research 62*(8), 1120–1127.

Liu, C., F. Batliwalla, W. Li, A. Lee, R. Roubenoff, E. Beckman, H. Khalili, A. Damle, M. Kern, R. M. Plenge, et al. (2008). Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-tnf treatment in rheumatoid arthritis. *Molecular medicine 14*(9-10), 575.

Moon, H., H. Ahn, R. L. Kodell, S. Baek, C.-J. Lin, and J. J. Chen (2007). Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine 41*(3), 197–207.

Nickl, R. and S. van de Geer (2013). Confidence sets in sparse regression. *Ann. Statist. 41*(6), 2852–2876.

Ong, F., K. Das, J. Wang, H. Vakil, J. Kuo, W. Blackwell, S. Lim, M. Goodarzi, K. Bernstein, J. Rotter, et al. (2012). Personalized medicine and pharmacogenetic biomarkers: progress in molecular oncology testing. *Expert Review of Molecular Diagnostics 12*(6), 593–602.

Padyukov, L., J. Lampa, M. Heimbürger, S. Ernestam, T. Cederholm, I. Lundkvist, P. Andersson, Y. Hermansson, A. Harju, L. Klareskog, et al. (2003). Genetic markers for the efficacy of tumour necrosis factor blocking therapy in rheumatoid arthritis. *Annals of the Rheumatic Diseases 62*(6), 526–529.

Qian, M. and S. A. Murphy (2011). Performance guarantees for individualized treatment rules. *Ann. Statist. 39*(2), 1180–1210.

Seegobin, S. D., M. H. Ma, C. Dahanayake, A. P. Cope, D. L. Scott, C. M. Lewis, and I. C. Scott (2014). Acpa-positive and acpa-negative rheumatoid arthritis differ in their requirements for combination dmards and corticosteroids: secondary analysis of a randomized controlled trial. *Arthritis Research & Therapy 16*(1), R13.

Simon, G., A. Sharma, X. Li, T. Hazelton, F. Walsh, C. Williams, A. Chiappori, E. Haura, T. Tanvetyanon, S. Antonia, et al. (2007). Feasibility and efficacy of molecular analysis-directed individualized therapy in advanced non–small-cell lung cancer. *Journal of Clinical Oncology 25*(19), 2741–2746.

Song, R., M. Kosorok, D. Zeng, Y. Zhao, E. Laber, and M. Yuan (2015, Feb). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat 4*(1), 59–68.

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 101*(2), 269–284.

Taylor, P. C. and M. Feldmann (2009). Anti-tnf biologic agents: still the therapy of choice for rheumatoid arthritis. *Nature Reviews Rheumatology 5*(10), 578.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B 58*(1), 267–288.

Tripuraneni, N. and L. Mackey (2019). Debiasing linear prediction. *arXiv preprint arXiv:1908.02341*.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202.

van der Heijde, D., L. Klareskog, V. Rodriguez-Valverde, C. Codreanu, H. Bolosiu, J. Melo-Gomes, J. Tornero-Molina, J. Wajdula, R. Pedersen, S. Fatenejad, et al. (2006). Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the tempo study, a double-blind, randomized trial. *Arthritis & Rheumatism 54*(4), 1063–1074.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B 76*(1), 217–242.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Am. Statist. Ass 107*(499), 1106–1118.

Zhou, X., N. Mayer-Hamblett, U. Khan, and M. R. Kosorok (2017). Residual weighted learning for estimating individualized treatment rules. *J. Am. Statist. Ass 112*(517), 169–187.

Zhu, Y. and J. Bradic (2017). A projection pursuit framework for testing general high-dimensional hypothesis. *arXiv preprint arXiv:1705.01024*.

Zhu, Y. and J. Bradic (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Am. Statist. Ass 113*(524), 1583–1600.

A. Detection Boundary for Decaying Loading

In the following, we consider the optimality result about decaying loading. Specifically, we calibrate the i-th largest element $\mathbf{x}_{\text{new},(i)}$ by the decaying rate parameter δ as defined in (34). A larger value of δ means that the loading decays faster; for the case $\delta = 0$, the loading is not decaying at all.

THEOREM 5. *Suppose that $s \le s_u \lesssim \frac{n}{\log p}$. We calibrate s, s_u and the decaying of $\mathbf{x}_{\text{new}}$ by γ, γ_u and δ, respectively, as defined in* (33) *and* (34) *in the main paper.*

(D1) If $\mathbf{x}_{\text{new}}$ is a fast decaying loading with $\delta \ge \frac{1}{2}$, then

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \frac{1}{\sqrt{n}} \cdot (1 + \sqrt{\log s} \cdot \mathbf{1}(\delta = \frac{1}{2})) \tag{49}$$

(D2) If $\mathbf{x}_{\text{new}}$ is a slow decaying loading with $0 \le \delta < \frac{1}{2}$, then

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \begin{cases} c_p \frac{s_u^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta} & \text{if} \quad \gamma_u < \frac{1}{2} \\ \sqrt{\frac{p^{1-2\delta}}{n}} & \text{if} \quad \gamma_u \ge \frac{1}{2} \end{cases} \tag{50}$$

where $c_p = \sqrt{\frac{\log(\log p)}{\log p}}$.

Similar to the exact sparse loading in Theorem 3, the detection lower bounds in the above theorem can be attained under regularity conditions. The following corollary presents the matched upper bound for the detection boundaries established in Theorem 5 over certain regimes.

COROLLARY 6. *Suppose that $s \le s_u \lesssim \frac{\sqrt{n}}{\log p}$.*

(D1) *If $\mathbf{x}_{\text{new}}$ is a fast decaying loading with $\delta \ge \frac{1}{2}$, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \tag{51}$$

In particular, for $\delta = \frac{1}{2}$, the detection boundary holds if $\gamma > 0$; otherwise the detection boundary holds up to a $\sqrt{\frac{\log p}{\log s}}$ factor.

(D2) *If $\mathbf{x}_{\text{new}}$ is a slow decaying loading with $0 \le \delta < \frac{1}{2}$, then the minimaxity detection boundary and adaptive detection boundary hold up to a $\sqrt{\log p}$ order*

(D2-a) *If the true sparsity s and the knowledge of s_u satisfies satisfies $\gamma < \gamma_u \le \frac{1}{2}$, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \frac{s_u^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}. \tag{52}$$

(D2-b) *If the true sparsity s and the knowledge of s_u satisfies $\gamma < \frac{1}{2} \le \gamma_u$, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \sqrt{\frac{p^{1-2\delta}}{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}. \tag{53}$$

(D2-c) *If the true sparsity s and the knowledge of s_u satisfies $\frac{1}{2} \le \gamma < \gamma_u$, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \sqrt{\frac{p^{1-2\delta}}{n}}. \tag{54}$$

We will provide some remarks for the above corollary. As an analogy to the exact sparse loading, (D1), (D2-a) and (D2-b) correspond to (E1), (E2-a) and (E2-b), respectively.

(D1) This corresponds to a large class of fast decaying loadings. in this case, even without the exact information about the sparsity level, we can conduct the hypothesis testing procedure as well as we know the exact sparsity level.

(D2) For the case of slowly decaying loading $\mathbf{x}_{\text{new}}$, we first discuss the following two cases,

(D2-a) This is similar to (E2-a), where the prior knowledge of sparsity s_u is relatively precise. We can use the sparsity level s_u to construct a testing procedure matching the adaptive detection boundary. See the proof of Corollary 6 for details.

(D2-b) This is similar to (E2-b), where the prior knowledge of sparsity s_u is rough. For such a case, the proposed testing procedure ϕ_α defined in (15) achieves the adaptive detection boundary.

Although the decaying loading shares some similarity with the exact sparse loading, there still exist significant distinctions in terms of the exact detection boundary. Interestingly, there exists an additional case (D2-c), which correspond to the case that the true sparsity itself is relatively dense. In this case, although the true sparsity level is high and the knowledge of sparsity is not available, the hypothesis testing problem itself is adaptive, which means, without any knowledge on the true sparsity level, we can conduct the hypothesis testing problem as well as the case of known sparsity.

We conclude this section by establishing a uniform optimality result of the proposed testing procedure ϕ_α in (15) over the decaying loading $\mathbf{x}_{\text{new}}$, which parallels Corollary 4 in the main paper for the case of exact loading,

COROLLARY 7. *Suppose that $s \le s_u \lesssim \frac{\sqrt{n}}{\log p}$ and $\gamma_u \ge \frac{1}{2}$. Then the testing procedure ϕ_α in* (15) *achieves the adaptive detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}}$ for any $\mathbf{x}_{\text{new}}$ satisfying* (34).

The above Corollary states that, in absence of accurate sparsity information, the proposed procedure ϕ_α is an adaptive optimal test for all decaying loadings $\mathbf{x}_{\text{new}}$.

B. Sparsity-assisted Hypothesis Testing Procedure

In this section, we consider the setting that there is additional information on the sparsity and present the method of constructing confidence interval for Δ_{new} and conducting hypothesis testing for (2) with incorporating the given sparsity information. Without loss of generality, we can assume the loading $\mathbf{x}_{\text{new}}$ is ordered as follows,

$$|\mathbf{x}_{\text{new},1}| \ge |\mathbf{x}_{\text{new},2}| \ge \cdots \ge |\mathbf{x}_{\text{new},p}|. \tag{55}$$

For $k = 1, 2$, we define $\widetilde{\boldsymbol{\beta}}_{k,j}$ to be the de-biased estimator introduced by Javanmard and Montanari (2014); Zhang and Zhang (2014); van de Geer et al. (2014) with the corresponding covariance

matrix of $\widetilde{\boldsymbol{\beta}}_{k,\cdot} \in \mathbb{R}^p$ as $\widehat{\mathrm{Var}}^k$. Define the vector ξ as a sparsified version of $\mathbf{x}_{\text{new}}$,

$$\xi_j = x_{\text{new},j} \quad \text{for } 1 \le j \le q \quad \text{and} \quad \xi_j = 0 \quad \text{for } q+1 \le j \le p, \tag{56}$$

where q is an integer to be specified later. Define $\mathbf{d} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ and the index sets G_1 and G_2 as

$$G_1 = \left\{ j : q+1 \le j \le p,\ \max\left\{ \left| \widetilde{\boldsymbol{\beta}}_{1,j} / \sqrt{\widehat{\mathrm{Var}}^1_{j,j}} \right|, \left| \widetilde{\boldsymbol{\beta}}_{2,j} / \sqrt{\widehat{\mathrm{Var}}^2_{j,j}} \right| \right\} > \sqrt{2.01 \log(2p)} \right\}$$

and

$$G_2 = \{ j : q+1 \le j \le p, j \notin G_1 \}$$

We define the estimator $\widehat{\xi^\intercal \mathbf{d}}$ as in (12) with $\mathbf{x}_{\text{new}}$ replaced with the sparsified loading ξ. In particular, the projection directions $\check{\mathbf{u}}_k$ for $k = 1, 2$ are constructed as

$$\check{\mathbf{u}}_k = \underset{\mathbf{u} \in \mathbb{R}^p}{\arg\min}\, \mathbf{u}^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} \quad \text{subject to } \left\| \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \xi \right\|_\infty \le \|\xi\|_2 \lambda_k$$
$$\left| \xi^\intercal \widehat{\boldsymbol{\Sigma}}_k \mathbf{u} - \|\xi\|_2^2 \right| \le \|\xi\|_2^2 \lambda_k,$$

where $\lambda_k \asymp \sqrt{\log p / n_k}$. Similar to (10), we estimate $\xi^\intercal(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ by

$$\widehat{\xi^\intercal \mathbf{d}} = (\xi^\intercal \widehat{\boldsymbol{\beta}}_1 + \check{\mathbf{u}}_1^\intercal \widehat{\mathbf{E}}_1) - (\xi^\intercal \widehat{\boldsymbol{\beta}}_2 + \check{\mathbf{u}}_2^\intercal \widehat{\mathbf{E}}_2)$$

and propose the following estimator of Δ_{new},

$$\check{\Delta}_{\text{new}} = \widehat{\xi^\intercal \mathbf{d}} + \sum_{j \in G_1} x_{\text{new},j} \left(\widetilde{\boldsymbol{\beta}}_{1,j} - \widetilde{\boldsymbol{\beta}}_{2,j} \right). \tag{57}$$

Note that, with probability larger than $1 - p^{-c}$,

$$\max_{j \in G_1} \left| \left(\widetilde{\boldsymbol{\beta}}_{1,j} - \widetilde{\boldsymbol{\beta}}_{2,j} \right) - \mathbf{d}_j \right| \le \sqrt{\widehat{\mathrm{Var}}^1_{j,j} + \widehat{\mathrm{Var}}^2_{j,j}} \sqrt{2.01 \log p}. \tag{58}$$

and

$$\max_{j \in G_2} |\boldsymbol{\beta}_{1,j}| \le 2\sqrt{\widehat{\mathrm{Var}}^1_{j,j}} \sqrt{2.01 \log(2p)} \quad \text{and} \quad \max_{j \in G_2} |\boldsymbol{\beta}_{2,j}| \le 2\sqrt{\widehat{\mathrm{Var}}^2_{j,j}} \sqrt{2.01 \log(2p)}. \tag{59}$$

As a consequence, we apply (59) and obtain

$$\begin{aligned}
\sum_{j \in G_2} |\boldsymbol{\beta}_{1,j}| &\le \sqrt{2\frac{\log p}{n_1}} \sigma_1 \sum_{j \in G_2} \min \left\{ \left| \frac{\boldsymbol{\beta}_{1,j}}{\sqrt{2\frac{\log p}{n_1}} \sigma_1} \right|, \frac{2\sqrt{\widehat{\mathrm{Var}}^1_{j,j}} \sqrt{2.01 \log(2p)}}{\sqrt{2\frac{\log p}{n_1}} \sigma_1} \right\} \\
&\le \sqrt{2\frac{\log p}{n_1}} \sigma_1 \sum_{j \in G_2} \min \left\{ \left| \frac{\boldsymbol{\beta}_{1,j}}{\sqrt{2\frac{\log p}{n_1}} \sigma_1} \right|, 1 \right\} \cdot \max \left\{ 1, \max_{j \in G_2} \frac{2\sqrt{\widehat{\mathrm{Var}}^1_{j,j}} \sqrt{2.01 \log(2p)}}{\sqrt{2\frac{\log p}{n_1}} \sigma_1} \right\} \\
&\le \max \left\{ \sqrt{2\frac{\log p}{n_1}} \sigma_1, \max_{j \in G_2} 2\sqrt{\widehat{\mathrm{Var}}^1_{j,j}} \sqrt{2.01 \log(2p)} \right\} s_u
\end{aligned} \tag{60}$$

where the last inequality follows from the assumption of capped ℓ_1 sparse of $\boldsymbol{\beta}_1$. Similarly, we have

$$\sum_{j\in G_2}|\boldsymbol{\beta}_{2,j}| \le \max\left\{\sqrt{2\frac{\log p}{n_2}}\sigma_2, \max_{j\in G_2}\sqrt{\widehat{\mathrm{Var}}^2_{j,j}}2\sqrt{2.01\log(2p)}\right\}s_u. \tag{61}$$

Hence, we have

$$\begin{aligned}\left|\sum_{j\in G_2}x_{\text{new},j}(\boldsymbol{\beta}_{1,j}-\boldsymbol{\beta}_{2,j})\right| &\le \max_{j\in G_2}|x_{\text{new},j}|\cdot\left(\sum_{j\in G_2}|\boldsymbol{\beta}_{1,j}|+\sum_{j\in G_2}|\boldsymbol{\beta}_{2,j}|\right)\\ &\le \max_{j\in G_2}|x_{\text{new},j}|\cdot s_u\sum_{k=1}^{2}\max\left\{\sqrt{2\frac{\log p}{n_k}}\sigma_2, \max_{j\in G_2}\sqrt{\widehat{\mathrm{Var}}^k_{j,j}}2\sqrt{2.01\log(2p)}\right\}\end{aligned} \tag{62}$$

Combining the bounds (58) and (62), we have

$$\left|\sum_{j\in G_1}x_{\text{new},j}\left(\widetilde{\boldsymbol{\beta}}_{1,j}-\widetilde{\boldsymbol{\beta}}_{2,j}\right)-\sum_{j=q+1}^{p}x_{\text{new},j}\mathbf{d}_j\right| \le S(s_u) \tag{63}$$

where

$$\begin{aligned}S(s_u) &= \sum_{j\in G_1}\sqrt{\widehat{\mathrm{Var}}^1_{j,j}+\widehat{\mathrm{Var}}^2_{j,j}}\sqrt{2.01\log p}\\ &+\max_{j\in G_2}|x_{\text{new},j}|\cdot s_u\sum_{k=1}^{2}\max\left\{\sqrt{2\frac{\log p}{n_k}}\sigma_2, \max_{j\in G_2}\sqrt{\widehat{\mathrm{Var}}^k_{j,j}}2\sqrt{2.01\log(2p)}\right\}\end{aligned} \tag{64}$$

It follows from Theorem 2 that

$$\frac{\widehat{\xi^\intercal\mathbf{d}}-\xi^\intercal\mathbf{d}}{\sqrt{\check{\mathrm{V}}}}\to N(0,1), \tag{65}$$

where

$$\check{\mathrm{V}} = \frac{\sigma_1^2}{n_1}\check{\mathbf{u}}_1^\intercal\widehat{\boldsymbol{\Sigma}}_1\check{\mathbf{u}}_1+\frac{\sigma_2^2}{n_2}\check{\mathbf{u}}_2^\intercal\widehat{\boldsymbol{\Sigma}}_2\check{\mathbf{u}}_2. \tag{66}$$

We construct the CI as

$$\mathrm{CI}(q,s_u) = \left(\check{\Delta}_{\text{new}}-z_{\alpha/2}\sqrt{\widetilde{\mathrm{V}}}-S(s_u), \check{\Delta}_{\text{new}}+z_{\alpha/2}\sqrt{\widetilde{\mathrm{V}}}+S(s_u)\right) \tag{67}$$

with $S(s_u)$ defined in (64) and

$$\widetilde{\mathrm{V}} = \frac{\widehat{\sigma}_1^2}{n_1}\check{\mathbf{u}}_1^\intercal\widehat{\boldsymbol{\Sigma}}_1\check{\mathbf{u}}_1+\frac{\widehat{\sigma}_2^2}{n_2}\check{\mathbf{u}}_2^\intercal\widehat{\boldsymbol{\Sigma}}_2\check{\mathbf{u}}_2. \tag{68}$$

We propose the following decision rule,

$$\phi(q,s_u) = \mathbf{1}\left(\check{\Delta}_{\text{new}}-z_{\alpha/2}\sqrt{\widetilde{\mathrm{V}}}-S(s_u)>0\right). \tag{69}$$

Combining (65) and (63), we establish the coverage property of the confidence interval $\mathrm{CI}(q,s_u)$ proposed in (67) and also control the type I error of the testing procedure $\phi(q,s_u)$ defined in (69).

It remains to control the length of $z_{\alpha/2}\sqrt{\widetilde{\mathrm{V}}} + S(s_u)$. We focus on the decaying loading $|x_{\mathsf{new},j}| \asymp j^{-\delta}$ for $0 \le \delta < \frac{1}{2}$. Following from (101), we have

$$\sqrt{\widetilde{\mathrm{V}}} \asymp \|\xi\|_2 \lesssim \sqrt{\sum_{j=1}^{q} |x_{\mathsf{new},j}|^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \lesssim q^{\frac{1}{2}-\delta} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{70}$$

and

$$S(s_u) \lesssim q^{-\delta} s_u \sqrt{\log p} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{71}$$

We take $q = \lfloor s_u^2 \log p \rfloor$ for both decaying loading and the exact loading and have

$$\left| z_{\alpha/2}\sqrt{\widetilde{\mathrm{V}}} + S(s_u) \right| \lesssim (s_u^2 \log p)^{\frac{1}{2}-\delta} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{72}$$

C. Additional Proofs

We present the proofs of Theorems 1 and 2 in Section C.1 and the proof of Proposition 1 in Section C.2; we present the proof of Proposition 2 in Section C.3; we present the proof of Corollaries 1 and 2 in Section C.4; we present the proof of Proposition 3 in Section C.5; we present the proof of Theorem 4 in Section C.6; we present the proof of Theorem 5 and Corollary 6 in Section C.7.

C.1. Proof of Theorems 1 and 2

By combining the error decompositions for $k = 1$ and $k = 2$ in (17), we have

$$\begin{aligned} \widehat{\mathbf{x}_{\mathsf{new}}^{\intercal}\boldsymbol{\beta}_1} - \widehat{\mathbf{x}_{\mathsf{new}}^{\intercal}\boldsymbol{\beta}_2} - \mathbf{x}_{\mathsf{new}}^{\intercal}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) &= \widehat{\mathbf{u}}_1^{\intercal} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^{\intercal} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{2,i}\epsilon_{2,i} \\ &+ \left(\widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 - \mathbf{x}_{\mathsf{new}}\right)^{\intercal} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) - \left(\widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2 - \mathbf{x}_{\mathsf{new}}\right)^{\intercal} \left(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) \end{aligned} \tag{73}$$

By Hölder's inequality, for $k = 1, 2$, we have

$$\left| (\widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\mathsf{new}})^{\intercal} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \right| \le \|\widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\mathsf{new}}\|_\infty \cdot \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \lesssim \|\mathbf{x}_{\mathsf{new}}\|_2 \sqrt{\frac{\log p}{n_k}} \cdot s_k \sqrt{\frac{\log p}{n_k}},$$

where the second inequality follows from the optimization constraint (8) and the condition (B1). Hence, we establish that, with probability larger than $1 - g(n_1, n_2)$,

$$\begin{aligned} &\left| \left(\widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 - \mathbf{x}_{\mathsf{new}}\right)^{\intercal} (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) - \left(\widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2 - \mathbf{x}_{\mathsf{new}}\right)^{\intercal} \left(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\right) \right| \\ \lesssim\ & \|\mathbf{x}_{\mathsf{new}}\|_2 \left(\frac{\|\boldsymbol{\beta}_1\|_0 \log p}{n_1} + \frac{\|\boldsymbol{\beta}_2\|_0 \log p}{n_2} \right). \end{aligned} \tag{74}$$

Proof of Theorem 1 Under the assumption (A1)

$$\mathbb{E}_{\cdot|\mathbb{X}} \left(\widehat{\mathbf{u}}_1^{\intercal} \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^{\intercal} \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i}\epsilon_{2,i} \right)^2 \lesssim \mathrm{V}, \tag{75}$$

where V is defined in (13). By (20), with probability larger than $1-p^{-c}-\frac{1}{t^2}$, then

$$\left|\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right| \lesssim t\|\mathbf{x}_{\text{new}}\|_2\,(1/\sqrt{n_1}+1/\sqrt{n_2}) \tag{76}$$

Combing (74) and (76), we establish Theorem 1.

Proof of Theorem 2 We establish Theorem 2 by assuming the Gaussian error assumption (A2). Conditioning on $\mathbb{X}$, we establish the following by applying Condition (A2), $\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i} \sim N(0,\mathrm{V})$ where V is defined in (13). After normalization, we have

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right) \mid \mathbb{X} \sim N(0,1) \tag{77}$$

and then after integrating with respect to $\mathbb{X}$, we have

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right) \sim N(0,1) \tag{78}$$

Combing (74) with (20), we show that with probability larger than $1-p^{-c}-g(n_1,n_2)$,

$$\frac{1}{\sqrt{\mathrm{V}}}\left|\left(\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1-\mathbf{x}_{\text{new}}\right)^\intercal(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1) - \left(\widehat{\mathbf{\Sigma}}_2\widehat{\mathbf{u}}_2-\mathbf{x}_{\text{new}}\right)^\intercal\left(\widehat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_2\right)\right| \le \frac{\|\boldsymbol{\beta}_1\|_0\log p}{\sqrt{n_1}} + \frac{\|\boldsymbol{\beta}_2\|_0\log p}{\sqrt{n_2}}.$$

Together with (78), we establish Theorem 2.

C.2. Proof of Proposition 1

This proposition is a modification of proofs of Theorem 1 and Theorem 2 presented in C.1. We shall focus on three parts that are different from the proof in C.1. Firstly, we modify the error decomposition of (73) as follows, by including the additional approximation error,

$$\begin{aligned}\widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_1} - \widehat{\mathbf{x}_{\text{new}}^\intercal\boldsymbol{\beta}_2} - \mathbf{x}_{\text{new}}^\intercal(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2) &= \widehat{\mathbf{u}}_1^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\,(\epsilon_{1,i}+\mathbf{r}_{1,i}) - \widehat{\mathbf{u}}_2^\intercal\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{1,i}\,(\epsilon_{2,i}+\mathbf{r}_{2,i}) \\ &+ \left(\widehat{\mathbf{\Sigma}}_1\widehat{\mathbf{u}}_1-\mathbf{x}_{\text{new}}\right)^\intercal(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1) - \left(\widehat{\mathbf{\Sigma}}_2\widehat{\mathbf{u}}_2-\mathbf{x}_{\text{new}}\right)^\intercal\left(\widehat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_2\right)\end{aligned} \tag{79}$$

We can control the additional error terms involving $\mathbf{r}_1$ as

$$|\widehat{\mathbf{u}}_1^\intercal\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\mathbf{r}_{1,i}| \le \sqrt{\frac{1}{n_1^2}\sum_{i=1}^{n_1}(\widehat{\mathbf{u}}_1^\intercal\mathbf{X}_{1,i})^2}\|\mathbf{r}_1\|_2 = o_p(\sqrt{\mathrm{V}}),$$

which follows from the assumption on $\mathbf{r}_1$. Similar argument is applied to the term involved with $\mathbf{r}_2$.

Secondly, we check the Lindeberg's condition and establish the asymptotic normality as a modification of (78) by allowing for non-Gaussian errors. We write

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right) = \sum_{k=1}^{2}\sum_{i=1}^{n_k}W_{k,i}$$

with

$$W_{1,i} = \frac{1}{n_1\sqrt{\mathrm{V}}}\widehat{\mathbf{u}}_1^\intercal \mathbf{X}_{1,i}\epsilon_{1,i} \text{ for } 1 \le i \le n_1 \quad \text{and} \quad W_{2,i} = -\frac{1}{n_2\sqrt{\mathrm{V}}}\widehat{\mathbf{u}}_2^\intercal \mathbf{X}_{2,i}\epsilon_{2,i} \text{ for } 1 \le i \le n_1 \tag{80}$$

Conditioning on $\mathbf{X}$, then $\{W_{1,i}\}_{1\le i\le n_1}$ and $\{W_{2,i}\}_{1\le i\le n_2}$ are independent random variables with $\mathbb{E}(W_{1,i} \mid \mathbf{X}) = 0$ for $1 \le l \le n_1$, $\mathbb{E}(W_{2,i} \mid \mathbf{X}) = 0$ for $1 \le k \le n_2$ and $\sum_{k=1}^{2}\sum_{i=1}^{n_k} \mathrm{Var}(W_{k,i} \mid \mathbf{X}) = 1$. To establish the asymptotic normality, it is sufficient to check the Lindeberg's condition, that is, for any constant $c > 0$,

$$\begin{aligned}
\sum_{k=1}^{2}\sum_{i=1}^{n} \mathbb{E}\left(W_{ki}^2 \mathbf{1}\{|W_{ki}| \ge c\} \mid \mathbf{X}\right) &\le \sum_{k=1}^{2}\sum_{i=1}^{n} \frac{\sigma_k^2}{n_k^2 \mathrm{V}} (\widehat{\mathbf{u}}_k^\intercal \mathbf{X}_{k,i})^2 \mathbb{E}\left(\frac{\epsilon_{k,i}^2}{\sigma_k^2}\mathbf{1}\left\{|\epsilon_{ki}| \ge \frac{cn_k\sqrt{\mathrm{V}}}{\|\mathbf{x}_{\mathsf{new}}\|_2 \tau_k}\right\} \mid \mathbf{X}\right) \\
&\le \max_{1\le k\le 2}\max_{1\le i\le n_k} \mathbb{E}\left(\frac{\epsilon_{k,i}^2}{\sigma_k^2}\mathbf{1}\left\{|\epsilon_{ki}| \ge \frac{cn_k\sqrt{\mathrm{V}}}{\|\mathbf{x}_{\mathsf{new}}\|_2 \tau_k}\right\} \mid \mathbf{X}\right) \\
&\lesssim \left(\frac{cn_k\sqrt{\mathrm{V}}}{\|\mathbf{x}_{\mathsf{new}}\|_2 \tau_k}\right)^{-\nu}
\end{aligned}$$

where the first inequality follows from the optimization constraint $\|\mathbf{X}_k\mathbf{u}\|_\infty \le \|\mathbf{x}_{\mathsf{new}}\|_2\tau_k$ in (25), the second inequality follows from the definition of V in (13) and the last inequality follows from the condition $\max_{k=1,2}\max_{1\le i\le n}\mathbb{E}(\epsilon_{k,i}^{2+\nu} \mid \mathbf{X}_{k,i}) \le M_0$. Hence, conditioning on $\mathbf{X}$, we establish the asymptotic normality of $\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right)$. By calculating its characteristic function, we can apply bounded convergence theorem to establish

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{\mathbf{u}}_1^\intercal \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1,i}\epsilon_{1,i} - \widehat{\mathbf{u}}_2^\intercal \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2,i}\epsilon_{2,i}\right) \xrightarrow{d} N(0,1).$$

Thirdly, we need to verify that $\widehat{\beta}$ satisfies condition (B1). In particular, we state the following lemma, which is essentially rewriting Theorem 6.3. of Bühlmann and van de Geer (2011) in the terminology of the current paper.

LEMMA 2. *For $k = 1, 2$, the Lasso estimator $\widehat{\boldsymbol{\beta}}_k$ defined in* (3) *with $A \ge 8e\|\boldsymbol{\epsilon}_{k,i}\|_{\psi_2}\max_{1\le j\le p}\frac{\|\mathbf{X}_{k,i}\|_{\psi_2}}{\boldsymbol{\Sigma}_{k,j,j}}$, satisfies*

$$\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \lesssim \frac{1}{\sqrt{n_k \log p}}\|\mathbf{r}_k\|_2^2 + \sqrt{\frac{\log p}{n_k}}\sum_{j=1}^{p}\min\{|\boldsymbol{\beta}_{k,j}|/\sigma_k\lambda_0, 1\} \lesssim s_k\sqrt{\frac{\log p}{n_k}}.$$

where $\|\boldsymbol{\epsilon}_{k,i}\|_{\psi_2}$ and $\|\mathbf{X}_{k,i}\|_{\psi_2}$ denote the sub-gaussian norms of $\boldsymbol{\epsilon}_{k,i} \in \mathbb{R}$ and $\mathbf{X}_{k,i} \in \mathbb{R}^p$, respectively, and $\boldsymbol{\Sigma}_{k,j,j}$ denotes the j-th diagonal entry of $\boldsymbol{\Sigma}_k$.

To apply Theorem 6.3. of Bühlmann and van de Geer (2011), it is sufficient to check

$$\max_{1\le j\le p}\frac{1}{W_{k,j}}\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{\epsilon}_{k,i}\mathbf{X}_{k,i,j}\right| \le A\sqrt{\frac{\log p}{n_k}},$$

where $\mathbf{X}_{k,i,j}$ denote the j-th variable of i-th observation in the k-th treatment group. Note that $\mathbb{E}\boldsymbol{\epsilon}_{k,i}\mathbf{X}_{k,i,j} = 0$ and $\boldsymbol{\epsilon}_{k,i}\mathbf{X}_{k,i,j}$ is random variable with sub-exponential norm smaller than $2\|\boldsymbol{\epsilon}_{k,i}\|_{\psi_2}\|\mathbf{X}_{k,i}\|_{\psi_2}$. By equation (73) of Javanmard and Montanari (2014) or Corollary 5.17 of Vershynin (2012), we establish that, with probability larger than $1 - p^{-c}$, for some positive constant $c > 0$,

$$\max_{1\le j\le p}\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{\epsilon}_{k,i}\mathbf{X}_{k,i,j}\right| \le 8e\sqrt{\frac{\log p}{n_k}}\|\boldsymbol{\epsilon}_{k,i}\|_{\psi_2}\|\mathbf{X}_{k,i}\|_{\psi_2}$$

By definition of G_3 and Lemma 4 in Cai and Guo (2017), we have $P\left(\left|\frac{W_{k,j}}{\boldsymbol{\Sigma}_{k,j,j}} - 1\right| \gtrsim \sqrt{\log p/n_k}\right) \lesssim p^{-c}$. Hence, with probability larger than $1 - p^{-c}$ for some positive $c > 0$, we have

$$\max_{1\le j\le p}\frac{1}{W_{k,j}}\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{\epsilon}_{k,i}\mathbf{X}_{k,i,j}\right| \le 8e\sqrt{\frac{\log p}{n_k}}\|\boldsymbol{\epsilon}_{k,i}\|_{\psi_2}\max_{1\le j\le p}\frac{\|\mathbf{X}_{k,i}\|_{\psi_2}}{\boldsymbol{\Sigma}_{k,j,j}}.$$

C.3. Proof of Proposition 2

In the following proof, we omit the index k to simply the notation, that is, $\mathbf{u} = \mathbf{u}_k$, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_k$ and $\lambda = \lambda_k$. We introduce the corresponding Lagrange function,

$$\begin{aligned} L(\mathbf{u},\tau,\eta,\tau_0,\eta_0) &= \mathbf{u}^\intercal\widehat{\boldsymbol{\Sigma}}\mathbf{u} + \tau^\intercal\left(\widehat{\boldsymbol{\Sigma}}\mathbf{u} - \mathbf{x}_{\text{new}} - \|\mathbf{x}_{\text{new}}\|_2\lambda\mathbf{1}\right) + \eta^\intercal\left(\mathbf{x}_{\text{new}} - \widehat{\boldsymbol{\Sigma}}\mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2\lambda\mathbf{1}\right) \\ &+ \tau_0\left(\frac{\mathbf{x}_{\text{new}}^\intercal}{\|\mathbf{x}_{\text{new}}\|_2}\widehat{\boldsymbol{\Sigma}}\mathbf{u} - (1+\lambda)\|\mathbf{x}_{\text{new}}\|_2\right) + \eta_0\left((1-\lambda)\|\mathbf{x}_{\text{new}}\|_2 - \frac{\mathbf{x}_{\text{new}}^\intercal}{\|\mathbf{x}_{\text{new}}\|_2}\widehat{\boldsymbol{\Sigma}}u\right) \end{aligned} \tag{81}$$

where $\tau \in \mathbb{R}^p$, $\eta \in \mathbb{R}^p$ and $\{\tau_i, \eta_i\}_{0\le i\le p}$ are positive constants. Then we derive the dual function $g(\tau,\eta,\tau_0,\eta_0) = \arg\min_{\mathbf{u}} L(u,\tau,\eta,\tau_0,\eta_0)$. By taking the first order-derivative of $L(u,\tau,\eta,\tau_0,\eta_0)$, we establish that the minimizer u^* of $L(u,\tau,\eta,\tau_0,\eta_0)$ satisfies

$$2\widehat{\boldsymbol{\Sigma}}\mathbf{u}^* + \widehat{\boldsymbol{\Sigma}}\left[(\tau-\eta) + (\tau_0-\eta_0)\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right] = 0. \tag{82}$$

By introducing $\gamma = \tau - \eta$ and $\gamma_0 = \tau_0 - \eta_0$, we have the expression of $L(u^*,\tau,\eta,\tau_0,\eta_0)$ as

$$\begin{aligned} g(\gamma,\eta,\gamma_0,\eta_0) = -\frac{1}{4}\left[\gamma + \gamma_0\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right]^\intercal\widehat{\boldsymbol{\Sigma}}\left[\gamma + \gamma_0\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right] - \mathbf{x}_{\text{new}}^\intercal\gamma - \|\mathbf{x}_{\text{new}}\|_2\lambda\mathbf{1}^\intercal(\gamma+2\eta) \\ - \|\mathbf{x}_{\text{new}}\|_2\gamma_0 - \lambda\|\mathbf{x}_{\text{new}}\|_2(\gamma_0+2\eta_0), \quad \text{where } \eta_i \ge -\gamma_i \text{ and } \eta_i \ge 0 \text{ for } 0 \le i \le p \end{aligned}$$

The computation of the maximum over η_0 and $\{\eta_i\}_{1\le i\le p}$ is based on the following observation, if $\gamma_i \ge 0$, then $\max_{\eta_i\ge\max\{0,-\gamma_i\}}\gamma_i + 2\eta_i = \gamma_i$; if $\gamma_i < 0$, then $\max_{\eta_i\ge\max\{0,-\gamma_i\}}\gamma_i + 2\eta_i = -\gamma_i$; Hence,

$$\max_{\eta_i\ge\max\{0,-\gamma_i\}}\gamma_i + 2\eta_i = |\gamma_i|. \tag{83}$$

By applying (83), we establish

$$\begin{aligned} \max_{\eta,\eta_0} g(\gamma,\eta,\gamma_0,\eta_0) = -\frac{1}{4}\left[\gamma + \gamma_0\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right]^\intercal\widehat{\boldsymbol{\Sigma}}\left[\gamma + \gamma_0\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right] \\ - \mathbf{x}_{\text{new}}^\intercal\left(\gamma + \gamma_0\frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\right) - \lambda\|\mathbf{x}_{\text{new}}\|_2\left(|\gamma_0| + \|\gamma\|_1\right) \end{aligned} \tag{84}$$

Then it is equivalent to solve the dual problem defined in (41). By (82), we establish

$$\widehat{\mathbf{u}}_k = \widehat{\mathbf{v}}_{-1}^k + \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2}\widehat{\mathbf{v}}_1^k.$$

C.4. Proof of Corollaries 1 and 2

Note that $\left|\frac{\widehat{\mathrm{V}}}{\mathrm{V}} - 1\right| \leq \sum_{j=1}^2 \left|\frac{\widehat{\sigma}_j^2}{\sigma_j^2} - 1\right|$ and hence $\frac{\widehat{\mathrm{V}}}{\mathrm{V}} \xrightarrow{p} 1$ follows from the condition (B2). Together with Theorem 2, we establish these two corollaries.

C.5. Proof of Proposition 3

Under the condition (F1), the projection $\mathbf{u} = \mathbf{0}$ belongs to the feasible set in (6) and hence the minimizer $\widetilde{\mathbf{u}}_1$ of (6) is zero since $\widehat{\mathbf{\Sigma}}_1$ is semi-positive-definite matrix.

If the non-zero coordinates of the loading $\mathbf{x}_{\text{new}}$ are of the same order of magnitude, we have $\|\mathbf{x}_{\text{new}}\|_2 \asymp \|\mathbf{x}_{\text{new}}\|_0 \|\mathbf{x}_{\text{new}}\|_\infty$. Then the condition $\|\mathbf{x}_{\text{new}}\|_0 \geq C\sqrt{n_1/\log p}$ will imply the condition (F1).

C.6. Proof of Theorem 4

Suppose that we observe a random variable Z which has a distribution $\mathbf{P}_\theta$ where the parameter θ belongs to the parameter space $\mathcal{H}$. Let π_i denote the prior distribution supported on the parameter space $\mathcal{H}_i$ for $i = 0, 1$. Let $f_{\pi_i}(z)$ denote the density function of the marginal distribution of Z with the prior π_i on $\mathcal{H}_i$ for $i = 0, 1$. More specifically, $f_{\pi_i}(z) = \int f_\theta(z)\,\pi_i(\theta)\,d\theta$, for $i = 0, 1$. Denote by $\mathbb{P}_{\pi_i}$ the marginal distribution of Z with the prior π_i on $\mathcal{H}_i$ for $i = 0, 1$. For any function g, we write $\mathbb{E}_{\pi_{\mathcal{H}_0}}(g(Z))$ for the expectation of $g(Z)$ with respect to the marginal distribution of Z with the prior $\pi_{\mathcal{H}_0}$ on $\mathcal{H}_0$. We define the χ^2 distance between two density functions f_1 and f_0 by

$$\chi^2(f_1, f_0) = \int \frac{(f_1(z) - f_0(z))^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1 \tag{85}$$

and the total variation distance by $L_1(f_1, f_0) = \int |f_1(z) - f_0(z)|\, dz$. It is well known that

$$L_1(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}. \tag{86}$$

LEMMA 3. *Suppose that π_i is a prior on the parameter space $\mathcal{F}_i$ for $i = 0, 1$, then we have*

$$\inf_{\theta \in \mathcal{F}_1} \mathbb{E}_\theta \phi \leq L_1(f_{\pi_1}, f_{\pi_0}) + \sup_{\theta \in \mathcal{F}_0} \mathbb{E}_\theta \phi \tag{87}$$

In addition, suppose that $L_1(f_{\pi_1}, f_{\pi_0}) < 1 - \alpha(1 + o(1)) - \eta$ for $0 < \alpha < \frac{1}{2}$, $\mathcal{F}_0 \subset \mathcal{H}_0(s_u)$ and $\mathcal{F}_1 \subset \mathcal{H}_1(s, \tau)$, then

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)|. \tag{88}$$

Proof of Lemma 3 It follows from the definition of $L_1(f_1, f_0)$ that

$$\mathbb{E}_{\pi_1}\phi - \mathbb{E}_{\pi_0}\phi \leq L_1(f_{\pi_1}, f_{\pi_0}). \tag{89}$$

Then (87) follows from

$$\inf_{\theta\in\mathcal{F}_1}\mathbb{E}_\theta\phi \le \mathbb{E}_{\pi_1}\phi \le L_1\left(f_{\pi_1}, f_{\pi_0}\right) + \mathbb{E}_{\pi_0}\phi \le L_1\left(f_{\pi_1}, f_{\pi_0}\right) + \sup_{\theta\in\mathcal{F}_0}\mathbb{E}_\theta\phi,$$

where the first and last inequalities follows from the definition of inf and sup and the second inequality follows from (89). The lower bound in (88) follows from the definition of $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ and the fact that

$$\boldsymbol{\omega}(s,\tau,\phi) \le \inf_{\theta\in\mathcal{F}_1}\mathbb{E}_\theta\phi \le L_1\left(f_{\pi_1}, f_{\pi_0}\right) + \sup_{\theta\in\mathcal{F}_0}\mathbb{E}_\theta\phi \le 1-\eta.$$

To establish the lower bound results, we divide the whole proof into two parts, where the first proof depends on the location permutation and the second proof does not depend on this.

Permuted Location Lower Bound We first establish the following lower bound through permuting the locations of non-zero coefficients,

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \ge \frac{1}{\sqrt{n}} \sum_{j=\max\{L-q+2,1\}}^{L} |x_{\text{new},j}| \sqrt{\max\{\log\left(cL/q^2\right), 0\}}. \tag{90}$$

For this case, we assume that $q \le \sqrt{cL}$; otherwise the lower bound in (90) is trivial. To simplify the notation of the proof, we fix $\boldsymbol{\beta}_2 = 0$ and denote $\boldsymbol{\beta}_1 = \boldsymbol{\eta}$. In addition, we set $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathrm{I}$ and $\sigma_1 = \sigma_2 = 1$. Without loss of generality, we set $x_{\text{new},i} \ge 0$ for $1 \le i \le p$ and $x_{\text{new},i} \ge x_{\text{new},i+1}$ for $1 \le i \le p-1$. By applying Lemma 3, we need to construct two parameters spaces $\mathcal{F}_0$ and $\mathcal{F}_1$ with considering the following three perspectives,

(a) $\mathcal{F}_0 \subset \mathcal{H}_0(s_u)$ and $\mathcal{F}_1 \subset \mathcal{H}_1(s,\tau)$.

(b) to constrain the distribution distance $L_1\left(f_{\pi_1}, f_{\pi_0}\right)$

(c) to maximize the functional distance $\min_{\theta_0\in\mathcal{F}_0,\theta_1\in\mathcal{F}_1}|\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)|$

To establish the lower bound (90), we construct the following parameter spaces,

$$\begin{aligned}
\mathcal{F}_0 &= \left\{\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\eta}, 1, \mathrm{I}\\ \mathbf{0}, 1, \mathrm{I}\end{pmatrix} : \boldsymbol{\eta}_1 = \rho\cdot\frac{\sum_{j=L-q+2}^{L} x_{\text{new},j}}{x_{\text{new},1}}, \|\boldsymbol{\eta}_{-1}\|_0 = q-1, \boldsymbol{\eta}_j \in \{0, -\rho\} \text{ for } 2\le j\le L\right\}\\
\mathcal{F}_1 &= \left\{\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\eta}, 1, \mathrm{I}\\ \mathbf{0}, 1, \mathrm{I}\end{pmatrix} : \boldsymbol{\eta}_1 = \rho\cdot\frac{\sum_{j=L-q+2}^{L} x_{\text{new},j}}{x_{\text{new},1}}, \boldsymbol{\eta}_{-1} = \mathbf{0}\right\}
\end{aligned} \tag{91}$$

For $\theta \in \mathcal{F}_0$, we have $\Delta_{\text{new}} = \rho\cdot\left(\sum_{j=L-q+2}^{L} x_{\text{new},j} - \sum_{j\in\text{supp}(\boldsymbol{\eta}_{-1})} x_{\text{new},j}\right) \le 0$, which is due to the construction of $\boldsymbol{\eta}_{-1}$ and the decreasing magnitude of the elements in $\mathbf{x}_{\text{new}}$; For $\theta \in \mathcal{F}_1$, we have $\Delta_{\text{new}} = \rho\cdot\sum_{j=L-q+2}^{L} x_{\text{new},j} \ge 0$. Hence, we have shown that

$$\mathcal{F}_0 \subset \mathcal{H}_0(s_u) \quad \text{and} \quad \mathcal{F}_1 \subset \mathcal{H}_1(s,\tau) \quad \text{for} \quad \tau = \rho\cdot\sum_{j=L-q+2}^{L} x_{\text{new},j} \tag{92}$$

To establish the distributional difference, we introduce π_0 to be the uniform prior on the parameter space $\mathcal{F}_0$ and π_1 to denote the mass point prior on the parameter space $\mathcal{F}_1$. Since L_1 distance is symmetric, we have

$$L_1(f_{\pi_1}, f_{\pi_0}) \leq \sqrt{\chi^2(f_{\pi_0}, f_{\pi_1})}. \tag{93}$$

As a remarkable difference from the typical lower bound construction, the null parameter space $\mathcal{F}_0$ is composite but the alternate parameter space $\mathcal{F}_1$ is simple. We use the symmetric property of the L_1 distance to control the distributional difference between this composite null and simple alternative in (93). We take $\rho = \frac{1}{2}\sqrt{\frac{2\log[(L-1)/(q-1)^2]}{n}}$. By Lemma 3 and Lemma 4 in Cai and Guo (2018b), we rephrase (3.33) in Cai and Guo (2018b) as

$$\chi^2(f_{\pi_0}, f_{\pi_1}) + 1 \leq \exp\left(\frac{(q-1)^2}{L-q}\right)\left(1 + \frac{1}{\sqrt{L-1}}\right)^{q-1} \leq \exp\left(\frac{(q-1)^2}{L-q} + \frac{q-1}{\sqrt{L-1}}\right)$$

The above inequality is further upper bounded by $\exp(\frac{1}{2}w^2 + w)$ for $w = \frac{q}{\sqrt{L}}$. Under the condition $\frac{q}{\sqrt{L}} \leq c$, we have $L_1(f_{\pi_1}, f_{\pi_0}) \leq \sqrt{\exp(\frac{1}{2}c^2 + c) - 1}$. By taking $c = \sqrt{1 + 2\log[c_0^2 + 1]} - 1$, we have $L_1(f_{\pi_1}, f_{\pi_0}) < c_0$. Then it suffices to control the functional difference $\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)|$, where $\mathrm{T} = \Delta_{\text{new}}$ and hence we have

$$\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)| \gtrsim \sqrt{\frac{2\log[L/q^2]}{n}} \cdot \sum_{j=L-q+2}^{L} x_{\text{new},j} \tag{94}$$

Fixed Location Lower Bound We will establish the following lower bound,

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \frac{1}{\sqrt{n}} \cdot \sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2}. \tag{95}$$

In this case, we do not perturb the location of non-zeros in constructing the null and alternative space but only perturb the coefficients corresponding to s-largest coefficients. To simplify the notation of the proof, we fix $\boldsymbol{\beta}_2 = 0$ and denote $\boldsymbol{\beta}_1 = \boldsymbol{\eta}$. In addition, we set $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathrm{I}$ and $\sigma_1 = \sigma_2 = 1$. Without loss of generality, we set $x_{\text{new},i} \geq 0$ for $1 \leq i \leq p$ and $x_{\text{new},i} \geq x_{\text{new},i+1}$ for $1 \leq i \leq p-1$. To establish the lower bound (90), we construct the following parameter space,

$$\begin{aligned}
\mathcal{F}_0 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} 0, 1, \mathrm{I} \\ \mathbf{0}, 1, \mathrm{I} \end{pmatrix} \right\} \\
\mathcal{F}_1 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta}, 1, \mathrm{I} \\ \mathbf{0}, 1, \mathrm{I} \end{pmatrix} : \boldsymbol{\eta}_j = \rho \cdot \frac{x_{\text{new},j}}{\sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2}} \quad \text{for } 1 \leq j \leq s \right\}
\end{aligned} \tag{96}$$

For $\theta \in \mathcal{F}_0$, we have $\Delta_{\text{new}} = 0$; For $\theta \in \mathcal{F}_1$, we have $\Delta_{\text{new}} = \rho \cdot \sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2} \geq 0$. Hence, we have shown that

$$\mathcal{F}_0 \subset \mathcal{H}_0(s_u) \quad \text{and} \quad \mathcal{F}_1 \subset \mathcal{H}_1(s, \tau) \quad \text{for} \quad \tau = \rho \cdot \sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2} \tag{97}$$

Let π_0 and π_1 denote the point mass prior over the parameter space $\mathcal{F}_0$ and $\mathcal{F}_1$, respectively. It follows from (7.25) in Cai and Guo (2017) that

$$\chi^2(f_{\pi_1}, f_{\pi_0}) \leq \exp\left(2n\rho^2\right) - 1 \tag{98}$$

By taking $\rho = \sqrt{\frac{\log(1+c_0^2)}{2n}}$, we have $L_1(f_{\pi_1}, f_{\pi_0}) \leq c_0$. Then it suffices to control the functional difference $\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)|$, where $\mathrm{T} = \Delta_{\text{new}}$ and hence we have

$$\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathrm{T}(\theta_0) - \mathrm{T}(\theta_1)| \gtrsim \frac{1}{\sqrt{n}} \cdot \sqrt{\sum_{j=1}^{s} x_{\text{new},j}^2}. \tag{99}$$

C.7. Proof of Theorem 5 and Corollary 6

The lower bound is an application of the general detection boundary (43), which is translated to the following lower bound,

$$\tau^* = \frac{1}{\sqrt{n}} \cdot \max\left\{ \sqrt{\sum_{j=1}^{s} j^{-2\delta}}, \sum_{j=\max\{L-q+2,1\}}^{L} j^{-\delta} \sqrt{\max\{\log\left(cL/q^2\right), 0\}} \right\}. \tag{100}$$

We also need the following fact, for integers $l_1 > 2$ and $l_2 > l_1$

$$\int_{l_1}^{l_2+1} x^{-\delta} dx \leq \sum_{j=l_1}^{l_2} j^{-\delta} \leq \int_{l_1-1}^{l_2} x^{-\delta} dx$$

Hence, we further have

$$\sum_{j=l_1}^{l_2} j^{-\delta} \in \begin{cases} \frac{1}{\delta-1}[(l_1-1)^{1-\delta} - l_2^{1-\delta}, l_1^{1-\delta} - (l_2+1)^{1-\delta}] & \text{for } \delta > 1 \\ [\log \frac{l_2+1}{l_1}, \log \frac{l_2}{l_1-1}] & \text{for } \delta = 1 \\ \frac{1}{1-\delta}[(l_2+1)^{1-\delta} - l_1^{1-\delta}, l_2^{1-\delta} - (l_1-1)^{1-\delta}] & \text{for } \delta < 1 \end{cases} \tag{101}$$

For the case (D1), we first consider the case $2\delta > 1$ and hence $\sum_{j=1}^{s} j^{-2\delta} = 1 + \sum_{j=2}^{s} j^{-2\delta} \asymp 1$; for the case $2\delta = 1$, we have $\sum_{j=1}^{s} j^{-2\delta} = 1 + \sum_{j=2}^{s} j^{-2\delta} \asymp \log s$; Hence, the lower bound (49) follows.

For the case (D2), we first consider $\gamma_u \geq \frac{1}{2}$, we take $q = \sqrt{p}$ in (100) and have

$$\sum_{j=\max\{p-q+2,1\}}^{p} j^{-\delta} \geq \frac{1}{1-\delta}\left((p+1)^{1-\delta} - (p-\sqrt{p}-2)^{1-\delta}\right) \asymp (p - c\sqrt{p})^{-\delta}\sqrt{p} \asymp p^{\frac{1}{2}-\delta}. \tag{102}$$

For the case $\gamma_u < \frac{1}{2}$, we take $L = s_u^2 \log p$, then we have

$$\sum_{j=\max\{L-s_u+2,1\}}^{L} j^{-\delta} \geq \frac{1}{1-\delta}\left((L+1)^{1-\delta} - (L-s_u+2)^{1-\delta}\right) \asymp (L - cs_u)^{-\delta} s_u \asymp s_u^{1-2\delta} (\log p)^{-\delta}.$$

Hence we have

$$\sum_{j=\max\{L-s_u+2,1\}}^{L} j^{-\delta} \sqrt{\left(\log\left(cL/s_u^2\right)\right)_+} \geq s_u^{1-2\delta} (\log p)^{\frac{1}{2}-\delta} \sqrt{\frac{\log(\log p)}{\log p}}.$$

Combined with (102), we establish the lower bound (50). Since

$$\|\mathbf{x}_{\text{new}}\|_2 = \sqrt{\sum_{j=1}^{p} j^{-2\delta}} \asymp \begin{cases} 1 & \text{for } \delta > 1/2 \\ \sqrt{\log p} & \text{for } \delta = 1/2 \\ p^{\frac{1}{2}-\delta} & \text{for } \delta < 1/2 \end{cases}$$

we apply Corollary 1 to establish the upper bounds and show that the detection boundaries $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ in (51), (53) and (54) are achieved by the hypothesis testing procedure ϕ_α defined in (15). In addition, the detection boundaries $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ and $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ in (52) and $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ in (53) are achieved by the hypothesis testing procedure introduced in (69).

D. Additional Simulation Results

D.1. Additional Simulation Results for Decaying Loading

In this session, we present additional simulation results for the decaying loading case, where $\mathbf{x}_{\text{new}}$ is generated as, $x_{\text{new},j} = \text{Ratio} \cdot j^{-\delta}$, where $\delta \in \{0, 0.1, 0.25, 0.5\}$ and $\text{Ratio} \in \{0.25, 0.375, 0.5\}$. We report the performance of the proposed HITS method for the decaying loading in Table 6. Specifically, with an increasing sample size, the empirical power reaches 100%, the empirical coverage rate reaches 95% and the length of CIs gets shorter. A similar bias-and-variance tradeoff is observed, where across all settings, in comparison to the plug-in Lasso estimator, both the proposed HITS estimator and the plug-in debiased estimator attained substantially smaller bias but at the expense of larger variability. For the slow or no decay settings with $\delta = 0$ or 0.1, the proposed HITS estimator has uniformly higher power and shorter length of CIs than the debiased plug-in estimator $\widetilde{\Delta_{\text{new}}}$ while the coverage of the CIs constructed based on both estimators are close to 95%. In the relatively faster decay setting with $\delta = 0.5$, $\widetilde{\Delta_{\text{new}}}$ and our proposed $\widehat{\Delta_{\text{new}}}$ perform more similarly. This is not surprising since the case of fast decaying loading is similar to the sparse loading case and the plugging-in of the debiased estimators can be shown to work if the loading is sufficiently sparse (or decaying sufficiently fast). However, we shall emphasize that $\widetilde{\Delta_{\text{new}}}$ is substantially more computationally intensive than $\widehat{\Delta_{\text{new}}}$. The calculation of $\widehat{\Delta_{\text{new}}}$ requires four fittings of Lasso-type algorithms twice whereas $\widetilde{\Delta_{\text{new}}}$ requires $2p + 2$ fittings.

D.2. Additional Data Analysis

Due to privacy constraints, the EHR data cannot be made publicly available. We have provided a de-identified privacy preserving EHR dataset and along with the code for analyzing this dataset through the online supplement. The data was generated by sampling from the real EHR data, adding noise and scaling factors to the original data as well as removing variable names to fully de-identify. The generated data contains $p = 171$ variables and the observation numbers are $n_1 = n_2 = 100$. We report some numerical results on this generated data. For group 1, we have randomly sampled 30 subjects and use their corresponding covariate observations as $\mathbf{x}_{\text{new}}$. We report our proposed 95% CIs for ITE on the top of Figure A1 with respect to these 30 sampled observations. If the CI for ITE is above zero, then it suggests that the first treatment is better than the second one; If the CI is below zero, it indicates the superiority of the second treatment. When the CI covers zero, it means that there is no clear evidence that one treatment is better than the other. On the top of Figure A1, it is interesting to note that the constructed CIs for three observations (with

			ERR		Coverage		Len		HITS			Lasso			Deb		
δ	Ratio	n	HITS	Deb	HITS	Deb	HITS	Deb	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
		200	0.86	0.79	0.96	0.95	1.54	1.75	0.39	0.03	0.39	0.20	0.13	0.15	0.45	0.01	0.45
	0.25	300	0.99	0.95	0.96	0.97	1.11	1.39	0.26	0.03	0.26	0.16	0.11	0.12	0.33	0.01	0.33
		400	1.00	0.98	0.96	0.96	1.00	1.19	0.25	0.02	0.25	0.13	0.09	0.10	0.30	0.01	0.30
		200	0.93	0.90	0.96	0.96	2.29	2.60	0.55	0.04	0.55	0.27	0.21	0.18	0.62	0.02	0.62
0	0.375	300	0.99	0.96	0.95	0.94	1.63	2.06	0.42	0.04	0.42	0.22	0.17	0.15	0.53	0.01	0.53
		400	1.00	0.99	0.95	0.95	1.46	1.76	0.35	0.01	0.35	0.19	0.14	0.12	0.43	0.05	0.43
		200	0.98	0.93	0.97	0.97	3.00	3.45	0.69	0.05	0.69	0.35	0.28	0.21	0.83	0.05	0.83
	0.5	300	1.00	1.00	0.97	0.97	2.17	2.74	0.51	0.09	0.50	0.29	0.23	0.17	0.63	0.03	0.63
		400	1.00	1.00	0.95	0.94	1.94	2.33	0.50	0.05	0.50	0.25	0.20	0.15	0.61	0.00	0.61
		200	0.97	0.96	0.95	0.95	1.02	1.09	0.25	0.04	0.25	0.18	0.13	0.12	0.27	0.00	0.27
	0.25	300	1.00	0.99	0.95	0.96	0.70	0.87	0.18	0.03	0.18	0.15	0.10	0.10	0.21	0.00	0.21
		400	1.00	1.00	0.96	0.96	0.65	0.75	0.16	0.03	0.16	0.12	0.09	0.08	0.19	0.00	0.19
		200	0.99	0.99	0.95	0.97	1.43	1.59	0.36	0.06	0.35	0.25	0.20	0.16	0.39	0.00	0.39
0.1	0.375	300	1.00	1.00	0.96	0.98	1.01	1.26	0.24	0.05	0.23	0.20	0.17	0.12	0.29	0.00	0.29
		400	1.00	1.00	0.95	0.96	0.91	1.08	0.23	0.02	0.22	0.17	0.14	0.10	0.26	0.02	0.26
		200	1.00	1.00	0.97	0.97	1.87	2.09	0.44	0.05	0.44	0.30	0.26	0.17	0.50	0.02	0.50
	0.5	300	1.00	1.00	0.96	0.97	1.32	1.66	0.32	0.07	0.32	0.26	0.22	0.14	0.40	0.00	0.40
		400	1.00	1.00	0.96	0.94	1.20	1.42	0.30	0.05	0.29	0.22	0.19	0.12	0.35	0.00	0.35
		200	0.99	1.00	0.93	0.92	0.59	0.62	0.16	0.02	0.16	0.17	0.11	0.13	0.17	0.01	0.17
	0.25	300	1.00	1.00	0.95	0.95	0.44	0.50	0.12	0.03	0.11	0.14	0.10	0.09	0.13	0.00	0.13
		400	1.00	1.00	0.94	0.95	0.39	0.45	0.11	0.02	0.10	0.12	0.09	0.08	0.12	0.01	0.12
		200	1.00	1.00	0.96	0.96	0.81	0.84	0.21	0.05	0.20	0.22	0.18	0.12	0.21	0.00	0.21
0.25	0.375	300	1.00	1.00	0.95	0.96	0.56	0.68	0.14	0.04	0.14	0.18	0.15	0.10	0.16	0.00	0.16
		400	1.00	1.00	0.96	0.96	0.52	0.59	0.13	0.02	0.13	0.15	0.13	0.09	0.15	0.00	0.15
		200	1.00	1.00	0.96	0.96	1.01	1.07	0.25	0.05	0.25	0.27	0.24	0.13	0.27	0.02	0.27
	0.5	300	1.00	1.00	0.91	0.94	0.69	0.86	0.19	0.05	0.19	0.22	0.19	0.12	0.22	0.01	0.22
		400	1.00	1.00	0.93	0.95	0.64	0.74	0.17	0.03	0.17	0.19	0.16	0.10	0.19	0.00	0.19
		200	0.96	0.99	0.94	0.91	0.48	0.41	0.13	0.02	0.13	0.15	0.10	0.11	0.12	0.00	0.12
	0.25	300	1.00	1.00	0.89	0.91	0.34	0.35	0.10	0.02	0.10	0.13	0.09	0.09	0.10	0.00	0.10
		400	1.00	1.00	0.93	0.92	0.31	0.32	0.09	0.01	0.09	0.11	0.07	0.08	0.09	0.00	0.09
		200	1.00	1.00	0.91	0.92	0.49	0.46	0.14	0.03	0.13	0.20	0.16	0.12	0.13	0.00	0.13
0.5	0.375	300	1.00	1.00	0.94	0.95	0.38	0.39	0.11	0.01	0.10	0.15	0.12	0.10	0.11	0.01	0.11
		400	1.00	1.00	0.92	0.95	0.33	0.36	0.09	0.02	0.09	0.13	0.11	0.08	0.09	0.00	0.09
		200	1.00	1.00	0.92	0.94	0.53	0.54	0.15	0.04	0.14	0.24	0.21	0.13	0.14	0.00	0.14
	0.5	300	1.00	1.00	0.94	0.94	0.42	0.45	0.11	0.02	0.11	0.18	0.15	0.10	0.12	0.01	0.12
		400	1.00	1.00	0.92	0.93	0.36	0.40	0.10	0.02	0.10	0.17	0.14	0.09	0.11	0.00	0.11

Table 6: Performance of the HITS hypothesis testing, in comparison with the plug-in Debiased Estimator ("Deb"), with respect to empirical rejection rate (ERR) as well as the empirical coverage (Coverage) and length (Len) of the CIs under the decaying loading $x_{\text{new},j} = Ratio * j^{-\delta}$. Reported also are the RMSE, bias and the standard error (SE) of the HITS estimator compared to the plug-in Lasso estimator ("Lasso") and the plug-in Debiased Estimator ("Deb").

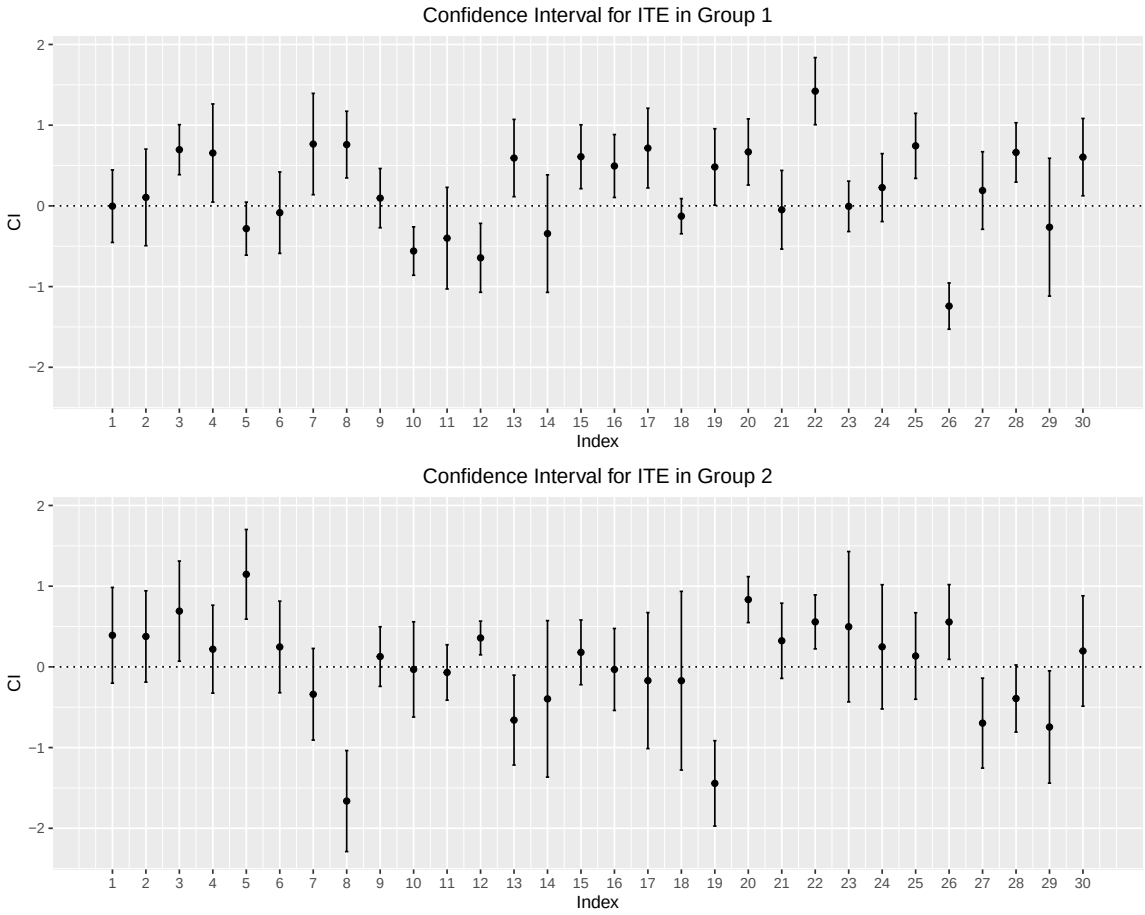


Fig. A1: Plotted confidence intervals for randomly sampled observations from groups 1 and 2. The top figure corresponds to 30 observations sampled from group 1 while the bottom corresponds to 30 observations from group 2.

indexes 10, 12, 26) are below zero. This indicates that they would benefit from being assigned to the second treatment assignment, instead of the current treatment assignment. Similar analysis has been implemented on 30 randomly sampled observations from group 2. The results are reported on the bottom of Figure A1. We observe the effect heterogeneity, that is, some observations benefit from the first treatment while some benefit from the second.