

Individualized Treatment Selection: An Optimal Hypothesis Testing Approach 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. The prediction of ITEs enables the comparison of the effectiveness of two treatment procedures for a specific individual. We propose a hypothesis testing approach to choosing between two available treatments for a given individual in the framework of high-dimensional linear models. The methodological novelty is the development of a testing procedure with the type-I error uniformly controlled for any future high-dimensional observation, while the existing methods can only handle certain specific forms of covariates observation. The procedure is based on a debiased estimator of the ITEs and its asymptotic normality. The asymptotic power of the proposed test is established and the finite sample performance is demonstrated in simulation studies. We introduce the optimality framework of hypothesis testing in high dimensions from both minimaxity and adaptivity perspectives and establish the optimality of the proposed procedure. The proposed method can be extended to conduct statistical inference for general linear contrasts, including both average treatment effect and the prediction problem. The procedure is 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:

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

where $\mathbf{Y}_k = (y_{k,1}, \dots, y_{k,n_k})^\top$ and $\mathbb{X}_k = (\mathbf{X}_{k,1}, \dots, \mathbf{X}_{k,n_k})^\top$ are the response and covariates observed independently for the n_k subjects in the treatment group k respectively, $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \dots, \epsilon_{k,n_k})^\top$ is the error vector independent of \mathbb{X}_k 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}_{new} , we consider testing the hypothesis

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

A novel procedure, High-dimensional Individualized Treatment Selection (HITS), is developed for testing the hypothesis (2). In addition, we construct point and interval estimators for the ITE $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$. The HITS method, while motivated by inference for ITE, can be also be used to make inference in other related problems including the average treatment effect (ATE) with \mathbf{x}_{new} chosen as the average of the covariates and prediction such as $\mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}_1$. See Sections 2.4 and 5 for more detailed discussion on these related problems.

In the following, we first describe our motivating example for Individualized Treatment Selection problem in Section 1.1 and then detail the statistical contributions of the current paper and comparison with literature results in Section 1.2 and 1.3, respectively.

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 propose to address this need by deriving an ITE model for RA using the biomarker linked EHR data at PHS. Our proposed high dimensional ITE inference 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 regression procedures can be used to obtain point estimators for $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ [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 these regularized estimators due to their inherent biases. These regularization methods tend to shrink small coefficients to zero and produce biased estimates for large coefficients. These biases can accumulate when projecting along the direction of \mathbf{x}_{new} and result in a significant bias in Δ_{new} . Statistically, the bias-variance tradeoff for estimating the ITE Δ_{new} is fundamentally different from that for estimating the regression vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

In this paper, we develop the novel HITS method that aims to choose between two treatments for a given individual based on the observed high-dimensional covariates. Statistically, we construct CIs and carry out hypothesis test for Δ_{new} under the challenging setting where \mathbf{x}_{new} is of high-dimension and of no special structure. Rigorous justifications for the inference procedures are given. It is shown that the test controls the type-I error. The power of the test as well as the coverage and length properties of the CIs are analyzed.

An optimality framework for hypothesis testing in the high-dimensional sparse linear model is introduced and the optimality of the proposed HITS method is established. The optimality framework is from two perspectives, minimaxity and adaptivity, where minimaxity captures the difficulty of the testing problem with true sparsity level known a priori and adaptivity captures a more challenging problem with respect to unknown sparsity. It is shown that the proposed HITS method is an optimal testing procedure

for problem (2) over various types of loadings \mathbf{x}_{new} in practical settings where the exact sparsity level is unknown. In a more general framework, the developed methods can be used to tackle the statistical inference for the linear contrasts $\mathbf{x}_{\text{new}}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and $\mathbf{x}_{\text{new}}^\top\boldsymbol{\beta}_k$ for $k = 1, 2$. As recognized in the literature [Cai and Guo, 2017, Athey et al., 2018, Zhu and Bradic, 2018], constructing a unified inference procedure for linear contrasts in high-dimensional regression with no restriction on the loading vector \mathbf{x}_{new} is extremely challenging. A more detailed discussion on the challenges of inference for general linear contrasts is given in Sections 4.3 and 8.

The main contribution of the current paper is two-fold,

1. We propose a novel calibration method for the plug-in estimators such that the proposed HITS estimator has an asymptotic normal distribution for any given \mathbf{x}_{new} . This is achieved by imposing an additional constraint in the construction of the projection direction for bias correction to guarantee that the variance of the HITS estimator dominates its bias for any \mathbf{x}_{new} . This additional constraint restricts the feasible set in one more direction and automatically enables a unified inference procedure for an arbitrary \mathbf{x}_{new} . See Figure 3 for illustration of the effect of this additional constraint. To the best of our knowledge, the proposed method is the first unified inference procedure with theoretical guarantee for general linear contrasts $\mathbf{x}_{\text{new}}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ and $\mathbf{x}_{\text{new}}^\top\boldsymbol{\beta}_k$ for $k = 1, 2$, where the high-dimensional loading \mathbf{x}_{new} is of no special structure.
2. An optimality framework is introduced from both the minimaxity and adaptivity perspectives. In particular, adaptivity results for hypothesis testing in high dimensions are obtained. It is shown that the proposed HITS method is optimal over different types of loadings, including both exact loadings and decaying loadings defined in Section 4.1. A novel technical tool, *transferring technique*, is developed to establish rate-optimal lower bound result. This tool can be of independent interest.

1.3. Comparison 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 the bias correction method 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}_{new} . It is still unknown prior to the current work how to make inference for all loadings \mathbf{x}_{new} .

More specifically, in the context of constructing CIs, Cai and Guo [2017] showed a significance difference between sparse and dense \mathbf{x}_{new} . The methods developed for a single regression coefficient can be extended to a sparse \mathbf{x}_{new} but the construction of a dense \mathbf{x}_{new} relies on a conservative upper bound and requires the information on sparsity level. Athey et al. [2018] constructed CI for the linear contrasts for \mathbf{x}_{new} only if the loading \mathbf{x}_{new} has a bounded weighted ℓ_2 norm. Zhu and Bradic [2018] constructed a CI for the linear contrast under the condition that the conditional expectation $\mathbb{E}[\mathbf{x}_{\text{new}}^\top \mathbf{X}_{1,i} \mid \mathbf{v}_1^\top \mathbf{X}_{1,i}, \dots, \mathbf{v}_{p-1}^\top \mathbf{X}_{1,i}]$ is a sparse linear combination of $\mathbf{v}_1^\top \mathbf{X}_{1,i}, \dots, \mathbf{v}_{p-1}^\top \mathbf{X}_{1,i}$, where $\{\mathbf{v}_j\}_{1 \leq j \leq p-1}$ span the space orthogonal to \mathbf{x}_{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}_{new} .

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. In Section 2, we discuss the proposed hypothesis testing and CI procedures for Δ_{new} . Theoretical properties of the proposed method are present in Section 3; Optimality of the hypothesis testing problem is discussed in Section 4; The proposed method is extended in Section 5 to quantify uncertainty for the prediction problem in high dimensional linear regression; The numerical performance of the proposed method 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 present in the supplementary materials.

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 $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$.

2. Methodology

In this section, we propose statistical inference procedures for the ITE $\Delta_{\text{new}} = \mathbf{x}_{\text{new}}^\top (\beta_1 - \beta_2)$. We first discuss the bias correction method for estimating the high dimensional regression parameter β_k in Section 2.1 and introduce a novel construction of projection direction $\mathbf{x}_{\text{new}}^\top \beta_k$ which adapt to any given loading \mathbf{x}_{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. Idea of Bias Correction

Without loss of generality, we discuss the bias correction idea for estimating $\mathbf{x}_{\text{new}}^\top \beta_1$ and the same approach can be extended to $k = 2$. With the observations $\mathbf{Y}_1 \in \mathbb{R}^{n_1}$ and $\mathbb{X}_1 \in \mathbb{R}^{n_1 \times p}$, we may estimate $\mathbf{B}_1 = (\beta_1^\top, \sigma_1)^\top$ as the scaled LASSO estimator

$$\hat{\mathbf{B}}_1 = (\hat{\beta}_1^\top, \hat{\sigma}_1)^\top = \arg \min_{\beta_1 \in \mathbb{R}^p, \sigma_1 \in \mathbb{R}^+} \frac{\|\mathbf{Y}_1 - \mathbb{X}_1 \beta_1\|_2^2}{2n_1 \sigma_1} + \frac{\sigma_1}{2} + \sqrt{\frac{A \log p}{n_1}} \sum_{j=1}^p W_{1,j} |\beta_{1,j}|, \quad (3)$$

where $A > 2$ is the pre-specified tuning parameter and $W_{1,j} = \sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} X_{k,i,j}^2}$ is the penalization weight of the j^{th} variable. A natural and simple way to estimate $\mathbf{x}_{\text{new}}^\top \beta_1$ is to plug in the scaled LASSO estimator $\hat{\beta}_1$ in (3). However, this plug-in estimator $\mathbf{x}_{\text{new}}^\top \hat{\beta}_1$ is known to suffer from the bias induced by the (scaled) LASSO method. 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 bias-corrected 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 $\hat{\beta}_{1,j}$ and then construct CIs centered at the corresponding bias-corrected estimators. This bias-correction idea can be generalized to general linear contrasts $\mathbf{x}_{\text{new}}^\top \beta$ for certain class of \mathbf{x}_{new} , where a key step is to estimate the bias $\mathbf{x}_{\text{new}}^\top (\hat{\beta}_1 - \beta_1)$. To this end, we aim to identify an effective projection direction $\hat{\mathbf{u}}_1 \in \mathbb{R}^p$ to construct a debiased estimator for $\mathbf{x}_{\text{new}}^\top \beta_1$ as

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

The error decomposition of the above bias-corrected estimator is

$$(\mathbf{x}_{\text{new}}^\top \hat{\beta}_1 + \hat{\mathbf{u}}_1^\top \hat{\mathbf{E}}_1) - \mathbf{x}_{\text{new}}^\top \beta_1 = -\mathbf{u}^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i} \epsilon_{1,i} + \left(\hat{\Sigma}_1 \mathbf{u} - \mathbf{x}_{\text{new}} \right)^\top (\hat{\beta}_1 - \beta_1) \quad (5)$$

where $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{X}_{k,i} \mathbf{X}_{k,i}^\top$. As highlighted in Zhang and Zhang [2014], Javanmard and Montanari [2014], Cai and Guo [2017], the key idea of constructing $\hat{\mathbf{u}}_1$ is to find a projection direction such that the variance of $\mathbf{u}^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i} \epsilon_{1,i}$ is minimized/controlled while the bias component $\hat{\Sigma}_1 \mathbf{u} - \mathbf{x}_{\text{new}}$ is constrained. Specifically, this idea is implemented in Cai and Guo [2017] via the following algorithm

$$\tilde{\mathbf{u}}_1 = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \left\{ \mathbf{u}^\top \hat{\Sigma}_1 \mathbf{u} : \left\| \hat{\Sigma}_1 \mathbf{u} - \mathbf{x}_{\text{new}} \right\|_\infty \leq \|\mathbf{x}_{\text{new}}\|_2 \lambda_1 \right\} \quad (6)$$

where $\lambda_1 \asymp \sqrt{\log p / n_1}$. Note that the estimator in Athey et al. [2018] is of a similar form as (6) but with some slight technical difference. We shall emphasize that the bias-corrected estimator (4) with the projection direction $\tilde{\mathbf{u}}_1$ defined in (6) is only effective for a small class of \mathbf{x}_{new} . Specifically, Cai and Guo [2017] showed that $\tilde{\mathbf{u}}_1$ is effective for

bias-correction if \mathbf{x}_{new} is of certain sparse structure and Athey et al. [2018] showed that the bias correction method is only effective if \mathbf{x}_{new} is of a bounded weighted ℓ_2 norm. On the other hand, as established in Proposition 2 in Section 8, if the loading \mathbf{x}_{new} is of certain dense structure, then $\tilde{\mathbf{u}}_1$ fails to correct the bias of the plug-in estimator. In next section, we discuss a novel method of constructing projection directions that adapt to arbitrary loading \mathbf{x}_{new} . This construction of projection directions enables us to correct the bias for an arbitrary loading \mathbf{x}_{new} , regardless of the structure of \mathbf{x}_{new} .

2.2. Construction of the Projection Direction

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

$$\hat{\mathbf{u}}_k = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \hat{\Sigma}_k \mathbf{u} \quad \text{subject to} \quad \left\| \hat{\Sigma}_k \mathbf{u} - \mathbf{x}_{\text{new}} \right\|_\infty \leq \|\mathbf{x}_{\text{new}}\|_2 \lambda_k \quad (7)$$

$$\left| \mathbf{x}_{\text{new}}^\top \hat{\Sigma}_k \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2 \right| \leq \|\mathbf{x}_{\text{new}}\|_2^2 \lambda_k, \quad (8)$$

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

$$\widehat{\mathbf{x}_{\text{new}}^\top \beta_k} = \mathbf{x}_{\text{new}}^\top \hat{\beta}_k + \hat{\mathbf{u}}_k^\top \hat{\mathbf{E}}_k. \quad (9)$$

The main difference between $\hat{\mathbf{u}}_1$ and $\tilde{\mathbf{u}}_1$ in (6) is the additional constraint (8). A general strategy developed in the bias correction literature [Zhang and Zhang, 2014, Javanmard and Montanari, 2014, Cai and Guo, 2017] is to minimize the variance component while constraining the bias. However, to develop a unified inference method for all \mathbf{x}_{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 (8) is imposed to constrain the variance such that it is dominating the bias component, which is essential for CI construction. Such a general optimization strategy can be of independent interest and applied to other inference problems. We call the construction defined in (7) and (8) for identifying the projection direction “Variance-enhancement Projection Direction”.

An equivalent way of constructing the projection direction defined in (7) and (8) is,

$$\hat{\mathbf{u}}_k = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \hat{\Sigma}_k \mathbf{u} \quad \text{subject to} \quad \sup_{\mathbf{w} \in \mathcal{C}} \left| \langle \mathbf{w}, \hat{\Sigma}_k \mathbf{u} - \mathbf{x}_{\text{new}} \rangle \right| \leq \|\mathbf{x}_{\text{new}}\|_2 \lambda_k \quad (10)$$

where $\mathcal{C} = \left\{ \mathbf{e}_1, \dots, \mathbf{e}_p, \frac{1}{\|\mathbf{x}_{\text{new}}\|_2} \mathbf{x}_{\text{new}} \right\}$ with \mathbf{e}_i denoting the i th standard Euclidean basis vector. The feasible set in (10) constrains that the projected values $\langle \mathbf{w}, \hat{\Sigma}_k \mathbf{u} - \mathbf{x}_{\text{new}} \rangle$ of $\hat{\Sigma}_k \mathbf{u} - \mathbf{x}_{\text{new}}$ to any of the $p+1$ vectors in \mathcal{C} are small. In contrast, the existing debiased procedures only constrain that the projected values to all the standard Euclidean basis vectors, $\max_{1 \leq j \leq p} \left| \langle \mathbf{e}_j, \hat{\Sigma}_k \mathbf{u} - \mathbf{x}_{\text{new}} \rangle \right|$, are small, which is motivated from the decomposition (5) and the Hölder’s inequality $|(\hat{\Sigma}_1 \mathbf{u} - \mathbf{x}_{\text{new}})^\top (\hat{\beta}_1 - \beta_1)| \leq \|\hat{\Sigma}_1 \mathbf{u} - \mathbf{x}_{\text{new}}\|_\infty \|\hat{\beta}_1 - \beta_1\|_1$. In the case where \mathbf{x}_{new} is of complicated structures, the idea motivated from (5) is not effective. However, this newly proposed projection direction defined in (10) resolves the

problem by introducing the additional direction $\frac{1}{\|\mathbf{x}_{\text{new}}\|_2} \mathbf{x}_{\text{new}}$ contained in \mathcal{C} . See more discussion in Section 8.

2.3. Statistical Inference for Individualized Treatment Effect

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

$$\widehat{\Delta}_{\text{new}} = \widehat{\mathbf{x}}_{\text{new}}^\top \widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{x}}_{\text{new}}^\top \widehat{\boldsymbol{\beta}}_2. \quad (11)$$

Compared to the plug-in estimator $\mathbf{x}_{\text{new}}^\top (\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)$, the main advantage of $\widehat{\Delta}_{\text{new}}$ is that the variance of $\widehat{\Delta}_{\text{new}}$ is typically dominating the bias of $\widehat{\Delta}_{\text{new}}$ while the bias of the plug-in estimator $\mathbf{x}_{\text{new}}^\top (\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)$ is usually larger than its variance. (See Table 2 for the numerical illustration.) Such a rebalance of bias and variance is extremely useful for uncertainty quantification 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

$$\mathbf{V} = \frac{\sigma_1^2}{n_1} \widehat{\mathbf{u}}_1^\top \widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 + \frac{\sigma_2^2}{n_2} \widehat{\mathbf{u}}_2^\top \widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2, \quad (12)$$

which can be estimated by

$$\widehat{\mathbf{V}} = \frac{\widehat{\sigma}_1^2}{n_1} \widehat{\mathbf{u}}_1^\top \widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 + \frac{\widehat{\sigma}_2^2}{n_2} \widehat{\mathbf{u}}_2^\top \widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2. \quad (13)$$

With $\widehat{\Delta}_{\text{new}}$ and the variance estimator $\widehat{\mathbf{V}}$, we may construct a CI for the ITE Δ_{new} as

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

and conduct HITS based on

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

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

It is worth noting that the proposed HITS procedure utilizes the \mathbf{x}_{new} information in the construction of the projection direction $\widehat{\mathbf{u}}_k$, where both the constraints in (7) and (8) depend on the observation \mathbf{x}_{new} . For observations with different \mathbf{x}_{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.

REMARK 1. To illustrate the main idea, we use the scaled LASSO method defined in (3) to construct the initial estimators. However, the proposed method, specifically the way of constructing the projection direction and also the following bias-correction, works for a large class of initial estimators. This will be specified in Section 3.

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 $\bar{\mathbf{X}}_2^\top(\beta_2 - \beta_1)$ where $\bar{\mathbf{X}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i}$. The statistical inference for the ATE $\bar{\mathbf{X}}_2^\top(\beta_2 - \beta_1)$ can be viewed as a special case of $x_{\text{new}}^\top(\beta_2 - \beta_1)$ with $x_{\text{new}} = \bar{\mathbf{X}}_2$. Both the point estimator (11) 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], where the main distinction is the additional constraint (8) proposed in the current paper. Without this additional condition, Athey et al. [2018] requires either of the following additional conditions to guarantee the asymptotic limiting distribution, where the *Bounded Loading* condition is imposed in Lemma 1 and the *Overlap* condition is stated as Assumption 5 in Athey et al. [2018]. We state both conditions in the following,

- 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, the overlap condition does not hold if $e(x)$ is the commonly used logit or probit probability function over an unbounded support.

3. Theoretical Properties

We establish in this section the theoretical guarantees of the proposed CI and hypothesis testing procedures. Before stating the theoretical properties, we first introduce some general conditions for the initial estimators $\hat{\mathbf{B}}_1$ and $\hat{\mathbf{B}}_2$.

(B1) With probability at least $1 - g(n_1, n_2)$ where $g(n_1, n_2) \rightarrow 0$, for $k = 1, 2$,

$$\max \left\{ \frac{1}{\sqrt{n_k}} \|\mathbb{X}_k(\hat{\beta}_k - \beta_k)\|_2, \|\hat{\beta}_k - \beta_k\|_2 \right\} \lesssim \sqrt{\frac{\|\beta_k\| \log p}{n_k}}, \quad \|\hat{\beta}_k - \beta_k\|_1 \lesssim \|\beta_k\|_0 \sqrt{\frac{\log p}{n_k}}.$$

(B2) $\hat{\sigma}_k^2$ is a consistent estimator of σ_k^2 for $k = 1, 2$, that is,

$$\max_{k=1,2} \left| \frac{\hat{\sigma}_k^2}{\sigma_k^2} - 1 \right| \xrightarrow{p} 0.$$

Most of the estimators proposed in the high-dimension literature are shown to satisfy the above conditions of estimating the regression vector and the variance of the regression error under various conditions. See Sun and Zhang [2012], Belloni et al. [2011], Bickel et al. [2009] and the reference within for more details. In addition, we assume the following conditions on the design and the errors, $\{\mathbf{X}_{1,i}, \epsilon_{1,i}\}_{1 \leq i \leq n_1}$ and $\{\mathbf{X}_{2,i}, \epsilon_{2,i}\}_{1 \leq i \leq n_2}$.

- (A1) For $k = 1, 2$, the rows $\mathbf{X}_{k,i}$ are i.i.d. p -dimensional sub-gaussian random vectors with mean $\boldsymbol{\mu}_k = \mathbb{E}\mathbf{X}_{k,i}$ and the second order moment $\boldsymbol{\Sigma}_k = \mathbb{E}(\mathbf{X}_{k,i}\mathbf{X}_{k,i}^\top)$ where $1/M_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_k) \leq \lambda_{\max}(\boldsymbol{\Sigma}_k) \leq M_1$ for $M_1 \geq 1$.
- (A2) The error vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \dots, \epsilon_{k,n_k})^\top$ is independent of \mathbb{X}_k with zero mean and constant variance σ_k^2 .
- (A3) The errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ follow Gaussian distributions.

The assumptions (A1) and (A2) are standard assumptions for the design and the regression error in the high-dimension literature. The condition (A3) is strong but is only imposed here to establish the distributional results. Although this condition can be extended to the Sub-gaussian error, we use the Gaussianity condition (A3) to simplify the algorithm and also the technical analysis.

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

$$\widehat{\mathbf{x}}_{\text{new}}^\top \boldsymbol{\beta}_k - \mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}_k = \widehat{\mathbf{u}}_k^\top \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{X}_{k,i} \boldsymbol{\epsilon}_{k,i} + \left(\widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k). \quad (16)$$

This decomposition (16) reflects the bias and variance decomposition of the proposed bias-correction estimator $\widehat{\mathbf{x}}_{\text{new}}^\top \boldsymbol{\beta}_k$, where the first error term $\widehat{\mathbf{u}}_k^\top \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{X}_{k,i} \boldsymbol{\epsilon}_{k,i}$ is the variance term while the second error term $(\widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}})^\top (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ is the remaining bias term. A similar bias and variance decomposition for the estimator $\widehat{\Delta}_{\text{new}}$ defined in (11) can be established by applying the error decomposition in (16) with $k = 1, 2$. The following theorem establishes the rate of convergence for $\widehat{\Delta}_{\text{new}}$.

THEOREM 1. *Suppose that the conditions (A1), (A2) and (B1) hold and $\|\boldsymbol{\beta}_k\|_0 \leq cn_k/\log p$ for $k = 1, 2$, then with probability larger than $1 - p^{-c} - g(n_1, n_2) - \frac{1}{t}$,*

$$\left| \widehat{\Delta}_{\text{new}} - \Delta_{\text{new}} \right| \lesssim \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{\|\boldsymbol{\beta}_1\|_0 \log p}{n_1} + \frac{\|\boldsymbol{\beta}_2\|_0 \log p}{n_2} \right) + t \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right). \quad (17)$$

In the above theorem, part of the convergence rate, $\|\mathbf{x}_{\text{new}}\|_2 \left(\frac{\|\boldsymbol{\beta}_1\|_0 \log p}{n_1} + \frac{\|\boldsymbol{\beta}_2\|_0 \log p}{n_2} \right)$, is an upper bound for the remaining bias of the proposed bias-corrected estimator while $\|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right)$ is an upper bound for the corresponding variance term. The following theorem shows that under the additional normal error distribution condition (A3) and the stronger sparsity condition, the proposed estimator has an asymptotic normal distribution.

THEOREM 2. *Suppose that the conditions (A1), (A2), (A3) and (B1) hold and $\|\boldsymbol{\beta}_k\|_0 \leq c\sqrt{n_k}/\log p$ for $k = 1, 2$, then*

$$\frac{1}{\sqrt{V}} \left(\widehat{\Delta}_{\text{new}} - \Delta_{\text{new}} \right) \xrightarrow{d} N(0, 1) \quad (18)$$

where V is defined in (12).

A key component of establishing the limiting distribution for $\widehat{\Delta}_{\text{new}}$ is to show that the variance term in (12) dominates the upper bound for the bias term in (17). We present this important component in the following Lemma, which characterizes the magnitude of the variance level V in (12).

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

$$c_0 \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right) \leq \sqrt{V} \leq C_0 \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right), \quad (19)$$

for some positive constants $c_0, C_0 > 0$.

We shall highlight that such a characterization of the variance level, mainly the lower bound of (19), is only achieved through incorporating the novel additional constraint (8) to identify the projection direction. Without this additional constraint, as shown in Proposition 2, the variance level will be vanishing and hence the lower bound in (19) does not hold.

In the following, 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 the ITE Δ_{new} . Regarding the hypothesis testing problem, we consider the following parameter space

$$\Theta(s) = \left\{ \boldsymbol{\theta} = \begin{pmatrix} \mathbf{B}_1, \boldsymbol{\Sigma}_1 \\ \mathbf{B}_2, \boldsymbol{\Sigma}_2 \end{pmatrix} : \|\boldsymbol{\beta}_k\|_0 \leq s, 0 < \sigma_k \leq M_0, \lambda_{\min}(\boldsymbol{\Sigma}_k) \geq c_0, \text{ for } k = 1, 2 \right\}, \quad (20)$$

for some positive constants $M_0 > 0$ and $c_0 > 0$. Then we define the null hypothesis parameter space as

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

and the local alternative parameter space as

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

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), (A3) and (B1), (B2) hold and $\|\boldsymbol{\beta}_k\|_0 \leq c\sqrt{n_k}/\log p$ for $k = 1, 2$, 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,*

$$\lim_{\min\{n_1, n_2\} \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \mathcal{H}_0} \mathbb{P}_{\boldsymbol{\theta}}(\phi_\alpha = 1) \leq \alpha. \quad (23)$$

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

$$\lim_{\min\{n_1, n_2\} \rightarrow \infty} \mathbb{P}_{\boldsymbol{\theta}}(\phi_\alpha = 1) = 1 - \Phi^{-1} \left(z_\alpha - \frac{\delta_0}{\sqrt{V}} \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right) \right). \quad (24)$$

Together with Lemma 1, we observe that the proposed test is powerful as long as $\delta_0 \rightarrow \infty$, where δ_0 controls the local alternative defined in (22). The main message for a real application 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 the related numerical illustration.

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), (A3) and (B1), (B2) hold and $\|\beta_k\|_0 \leq c\sqrt{n_k}/\log p$ for $k = 1, 2$. Then the coverage property of the proposed CI defined in (14) is*

$$\lim_{\min\{n_1, n_2\} \rightarrow \infty} \mathbb{P}_{\theta}(\Delta_{\text{new}} \in \text{CI}) \geq 1 - \alpha \text{ for any } \mathbf{x}_{\text{new}} \in \mathbb{R}^p. \quad (25)$$

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 $\|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right)$, 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}_{new} , the lengths of the corresponding CIs for ITE Δ_{new} can be quite different, where the length is determined by $\|\mathbf{x}_{\text{new}}\|_2$ and the numerical illustration is present in Section 6.

4. Optimality for Hypothesis Testing

In this section, we investigate the optimality of the proposed procedure in the hypothesis testing framework from two perspectives, minimaxity and adaptivity. We introduce the optimality framework in Section 4.1, present the detailed results in Section 4.2, and compare with the optimality of confidence intervals in Section 4.3.

4.1. Optimality Framework for Hypothesis Testing: Minimaxity and Adaptivity

The performance of a testing procedure can be evaluated in terms of its size (type I error probability) and its power (or type II error probability). For a given test ϕ and a null parameter space $\mathcal{H}_0(s)$, its size is

$$\alpha(s, \phi) = \sup_{\theta \in \mathcal{H}_0(s)} \mathbb{E}_{\theta} \phi. \quad (26)$$

It has been shown in Corollary 1 that the proposed test ϕ_{α} satisfies $\alpha(s, \phi_{\alpha}) \leq \alpha$. To investigate the power, we consider the following local alternative parameter space

$$\mathcal{H}_1(s, \tau) = \{\theta \in \Theta(s) : \mathbf{x}_{\text{new}}^T (\beta_1 - \beta_2) = \tau\}, \quad (27)$$

for a given $\tau > 0$. The power of a test ϕ over the parameter space $\mathcal{H}_1(s, \tau)$ is

$$\omega(s, \tau, \phi) = \inf_{\theta \in \mathcal{H}_1(s, \tau)} \mathbb{E}_\theta \phi. \quad (28)$$

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,

$$\tau_{\min}(s, \mathbf{x}_{\text{new}}) = \arg \min_{\tau} \left\{ \tau : \sup_{\phi: \alpha(s, \phi) \leq \alpha} \omega(s, \tau, \phi) \geq 1 - \eta \right\}, \quad (29)$$

where $\eta \in [0, 1)$ is a small positive constant controlling the type II error probability. The quantity $\tau_{\min}(s, \mathbf{x}_{\text{new}})$ depends on \mathbf{x}_{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 $\tau_{\min}(s, \mathbf{x}_{\text{new}})$ for simplicity. This quantity $\tau_{\min}(s, \mathbf{x}_{\text{new}})$ is referred to as the minimax detection boundary of the hypothesis testing problem (2). In other words, $\tau_{\min}(s, \mathbf{x}_{\text{new}})$ 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,

$$\alpha(s, \phi) \leq \alpha \quad \text{and} \quad \omega(s, \phi, \tau) \geq 1 - \eta \quad \text{for} \quad \tau \asymp \tau_{\min}(s, \mathbf{x}_{\text{new}}) \quad (30)$$

then ϕ is defined to be minimax optimal.

The minimax detection boundary in (29) 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 (29) if the true sparsity level s is unknown
- Q2. What is the optimal procedure in absence of accurate sparsity information?

To facilitate the definition of adaptivity, we consider the case of 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 the aforementioned questions on adaptivity, we assume that only the upper bound s_u is known instead of the exact sparsity level s . As a consequence, a testing procedure needs to be constructed with the size uniformly controlled over the parameter space $\mathcal{H}_0(s_u)$,

$$\alpha(s_u, \phi) = \sup_{\theta \in \mathcal{H}_0(s_u)} \mathbb{E}_\theta \phi \leq \alpha. \quad (31)$$

In contrast to (26), the control of size 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 is unknown. Similar to (29), we define the adaptive detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ as

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) = \arg \min_{\tau} \left\{ \tau : \sup_{\phi: \alpha(s_u, \phi) \leq \alpha} \omega(s, \tau, \phi) \geq 1 - \eta \right\}. \quad (32)$$

Comparing (32) with (29), the imprecise information about the sparsity level only affects the control of size, where the power functions in (32) and (29) are evaluated over the same parameter space, $\mathcal{H}_1(s, \tau)$. If a test ϕ satisfies the following property,

$$\alpha(s_u, \phi) \leq \alpha \quad \text{and} \quad \omega(s, \tau, \phi) \geq 1 - \eta \quad \text{for } \tau \asymp \tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \quad (33)$$

then ϕ is defined to be adaptive optimal. 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 would be the best that we can aim for if there is lack of accurate information on sparsity.

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}_{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 in the construction of the test, 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.

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 level. The construction of an adaptive optimal test satisfying (33) 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}_{new} , Exact Loading, indexed with (E), and Decaying Loading, indexed with (D).

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

$$c_0 \leq \max_{\{i: \mathbf{x}_{\text{new},i} \neq 0\}} |\mathbf{x}_{\text{new},i}| / \min_{\{i: \mathbf{x}_{\text{new},i} \neq 0\}} |\mathbf{x}_{\text{new},i}| \leq C_0, \quad (34)$$

for some positive constants $C_0 \geq c_0 > 0$. The condition (34) assumes that all non-zero coefficients of the loading vector \mathbf{x}_{new} are of the same order of magnitude and hence the complexity of an exact loading \mathbf{x}_{new} is captured by its sparsity level. We calibrate the sparsity levels of the regression vectors as

$$s = p^\gamma, \quad s_u = p^{\gamma_u} \quad \text{for } 0 \leq \gamma < \gamma_u \leq 1, \quad (35)$$

and calibrate the sparsity of the exact loading \mathbf{x}_{new} as

$$\|\mathbf{x}_{\text{new}}\|_0 = p^{\gamma_{\text{new}}} \quad \text{for } 0 \leq \gamma_{\text{new}} \leq 1. \quad (36)$$

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

(E1) \mathbf{x}_{new} is called an *exact sparse loading* if it satisfies (34) with $\gamma_{\text{new}} \leq 2\gamma$;

(E2) \mathbf{x}_{new} is called an *exact dense loading* if it satisfies (34) with $\gamma_{\text{new}} > 2\gamma$;

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

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

$$|x_{\text{new},(i)}| \asymp i^{-\delta} \quad \text{for some constant } \delta \geq 0. \quad (37)$$

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

(D1) \mathbf{x}_{new} is called a *fast decaying loading* if it satisfies (37) with $\delta \geq \frac{1}{2}$;

(D2) \mathbf{x}_{new} is called a *slow decaying loading* if it satisfies (37) with $0 \leq \delta < \frac{1}{2}$.

To simplify the presentation, we present the optimality results for the case $n_1 \asymp n_2$, denoted by n and the results can be extended to the case where n_1 and n_2 are of different orders. In the following, 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 will be deferred to Section A in the supplement. The following theorem establishes the lower bounds for the adaptive detection boundary for exact loadings.

THEOREM 3. Suppose that $s \leq s_u \lesssim \frac{n}{\log p}$. We calibrate s, s_u and $\|\mathbf{x}_{\text{new}}\|_0$ by γ, γ_u and γ_{new} , respectively, as defined in (35) and (36).

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \frac{\sqrt{\|\mathbf{x}_{\text{new}}\|_0} \|\mathbf{x}_{\text{new}}\|_{\infty}}{\sqrt{n}} \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}}; \quad (38)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \begin{cases} \|\mathbf{x}_{\text{new}}\|_{\infty} s_u \sqrt{\frac{\log p}{n}} & \text{if } \gamma_{\text{new}} > 2\gamma_u; \\ \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} & \text{if } \gamma_{\text{new}} \leq 2\gamma_u. \end{cases} \quad (39)$$

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 $0 < \mathbf{x}_{\text{new}}^\top(\beta_1 - \beta_2) < \tau$. A new technique, *transferring technique*, is developed for establishing the sharp lower bounds for the adaptive detection boundary. Define the index of \mathbf{x}_{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 $\mathbf{x}_{\text{new},i_{\max}}(\beta_{1,i_{\max}} - \beta_{2,i_{\max}}) > 0$ and $\mathbf{x}_{\text{new}}^\top(\beta_1 - \beta_2) \leq 0$; 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.

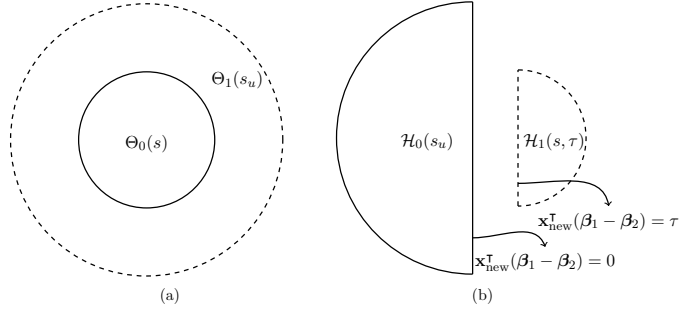


Fig. 1: (a) Null and alternative parameter spaces for the confidence set construction; (b) Null and alternative parameter spaces for the hypothesis testing problem.

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

COROLLARY 3. Suppose that $s \leq s_u \lesssim \frac{\sqrt{n}}{\log p}$.

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \quad (40)$$

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

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_\infty s_u \sqrt{\frac{\log p}{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_\infty s \sqrt{\frac{\log p}{n}}. \quad (41)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \|x_{\text{new}}\|_{\infty} s \sqrt{\frac{\log p}{n}}. \quad (42)$$

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 of the supplement. 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 $s \leq s_u \lesssim \frac{\sqrt{n}}{\log p}$ and $\gamma_u \geq \frac{1}{2}\gamma_{\text{new}}$. 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}_{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 exact loadings \mathbf{x}_{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].

- 1 The current paper closes the gap between the definitions of exact sparse and dense loading in Cai and Guo [2017], where the lower bounds for 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 the current paper.
- 2 The result in the current paper considers the additional setting (E2-b) than those in Cai and Guo [2017], which corresponds to the case where the knowledge on the sparsity level is rough. This additional result is not only of technical interest, but is of 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}_{new} .

3 In addition, the technical tools for general loading \mathbf{x}_{new} have been developed in Theorem 4, which can handle the loadings not considered in Cai and Guo [2017]. Specifically, we summarize in Table 1 the optimality results for the decaying loading defined in (37). 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 hypothesis testing problem (2) is adaptive if both γ and γ_u are above $1/2$. See the detailed discussion in Section A of the supplement.

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)	$\gamma < \gamma_u$	$\frac{\ \mathbf{x}_{\text{new}}\ _2}{\sqrt{n}}$	\asymp	$\frac{\ \mathbf{x}_{\text{new}}\ _2}{\sqrt{n}}$	Yes
$[0, 1/2)$	(D2-a)	$\gamma < \gamma_u \leq \frac{1}{2}$	$\frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}$	\ll	$\frac{s_u^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}$	No
	(D2-b)	$\gamma < \frac{1}{2} \leq \gamma_u$	$\frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}$	\ll	$\frac{\ \mathbf{x}_{\text{new}}\ _2}{\sqrt{n}}$	No
	(D2-c)	$\frac{1}{2} \leq \gamma < \gamma_u$	$\frac{\ \mathbf{x}_{\text{new}}\ _2}{\sqrt{n}}$	\asymp	$\frac{\ \mathbf{x}_{\text{new}}\ _2}{\sqrt{n}}$	Yes

Table 1: Optimality for the decaying loading \mathbf{x}_{new} defined in (37) over the regime $s \lesssim s_u \lesssim \frac{\sqrt{n}}{\log p}$. 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 or not for 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 $\{(X_i, y_i)\}_{1 \leq i \leq n}$ satisfying

$$y_i = X_i^\top \boldsymbol{\beta} + \epsilon_i \quad \text{where } \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \quad \text{for } 1 \leq i \leq n, \quad (43)$$

and $\{\epsilon_i\}_{1 \leq i \leq n}$ is independent of the design matrix X . The problem of interest is inference for the conditional expectation $\mathbb{E}(y_i | X_i = \mathbf{x}_{\text{new}}) = \mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}$. Uncertainty quantification for $\mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}$ is a simpler version of the testing problem (15). Due to its importance and for clarity, we present a separate result on this prediction problem. We use $\hat{\boldsymbol{\beta}}$ to denote the scaled Lasso estimator of $\boldsymbol{\beta}$ based on the observations $\{(X_i, y_i)\}_{1 \leq i \leq n}$ and construct the following bias-corrected point estimator,

$$\widehat{\mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}} = \mathbf{x}_{\text{new}}^\top \hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}^\top \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \cdot (y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}) \quad (44)$$

with the projection direction defined as

$$\begin{aligned} \hat{\mathbf{u}} = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \hat{\Sigma} \mathbf{u} \quad \text{subject to} \quad & \left\| \hat{\Sigma} \mathbf{u} - \mathbf{x}_{\text{new}} \right\|_\infty \leq \|\mathbf{x}_{\text{new}}\|_2 \lambda \\ & \left| \mathbf{x}_{\text{new}}^\top \hat{\Sigma} \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2 \right| \leq \|\mathbf{x}_{\text{new}}\|_2^2 \lambda, \end{aligned}$$

where $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$ 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 $\left| \mathbf{x}_{\text{new}}^\top \hat{\Sigma} \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2 \right| \leq \|\mathbf{x}_{\text{new}}\|_2^2 \lambda$, which guarantees the asymptotic limiting distribution for any $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$. We introduce the following general condition for the initial estimator $\hat{\beta}$ and then establish the limiting distribution for the point estimator $\widehat{\mathbf{x}_{\text{new}}^\top \beta}$ in Corollary 5.

(P) With probability at least $1 - g(n)$ where $g(n) \rightarrow 0$,

$$\max \left\{ \frac{1}{\sqrt{n}} \|X(\hat{\beta} - \beta)\|_2, \|\hat{\beta} - \beta\|_2 \right\} \lesssim \sqrt{\frac{\|\beta\| \log p}{n}}, \quad \|\hat{\beta} - \beta\|_1 \lesssim \|\beta\|_0 \sqrt{\frac{\log p}{n}}.$$

COROLLARY 5. *Suppose that the regression model (43) holds where $\|\beta\|_0 \leq c\sqrt{n}/\log p$ and the rows X_i are i.i.d. p -dimensional sub-gaussian random vectors with mean $\mu = \mathbb{E}X_i$ and the second order moment $\Sigma = \mathbb{E}(X_i X_i^\top)$ satisfying $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$ for positive constants $C_0, c_0 > 0$. For any initial estimator $\hat{\beta}$ satisfying the condition (P), then*

$$\frac{1}{\sqrt{\sigma^2 \hat{\mathbf{u}}^\top \hat{\Sigma} \hat{\mathbf{u}}/n}} \left(\widehat{\mathbf{x}_{\text{new}}^\top \beta} - \mathbf{x}_{\text{new}}^\top \beta \right) \xrightarrow{d} N(0, 1). \quad (45)$$

Based on this corollary, we use $\hat{V} = \hat{\sigma}^2 \hat{\mathbf{u}}^\top \hat{\Sigma} \hat{\mathbf{u}}/n$ to quantify the uncertainty of the estimator $\widehat{\mathbf{x}_{\text{new}}^\top \beta}$, which leads to the following CI for $\mathbf{x}_{\text{new}}^\top \beta$,

$$\text{CI}_{\text{Pred}} = \left(\widehat{\mathbf{x}_{\text{new}}^\top \beta} - z_{\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{n} \hat{\mathbf{u}}^\top \hat{\Sigma} \hat{\mathbf{u}}}, \quad \widehat{\mathbf{x}_{\text{new}}^\top \beta} + z_{\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{n} \hat{\mathbf{u}}^\top \hat{\Sigma} \hat{\mathbf{u}}} \right).$$

If $\hat{\sigma}^2$ is a consistent estimator of σ^2 , then this constructed CI will be guaranteed to have coverage for $\mathbf{x}_{\text{new}}^\top \beta$ 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

This section investigates the numerical performance of HITS and the CI construction through simulation studies and shows that the HITS method outperforms the existing methods in the literature.

Before presenting the numerical results, we first introduce the equivalent dual form to find the projection direction defined in (7) and (8).

PROPOSITION 1. *The constrained optimizer $\hat{\mathbf{u}}_k \in \mathbb{R}^p$ for $k = 1, 2$ defined in (7) and (8) can be computed in the form of*

$$\hat{\mathbf{u}}_k = \hat{\mathbf{v}}_{-1}^k + \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2} \hat{\mathbf{v}}_1^k. \quad (46)$$

where $\hat{\mathbf{v}}^k \in \mathbb{R}^{p+1}$ is defined as

$$\hat{\mathbf{v}}^k = \arg \min_{\mathbf{v} \in \mathbb{R}^{p+1}} \left\{ \frac{1}{4} \mathbf{v}^\top \mathbb{H}^\top \hat{\Sigma}_k \mathbb{H} \mathbf{v} + \mathbf{x}_{\text{new}}^\top \mathbb{H} \mathbf{v} + \lambda_k \|\mathbf{x}_{\text{new}}\|_2 \cdot \|\mathbf{v}\|_1 \right\} \quad (47)$$

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

By the above proposition, we transform the constrained minimization problem to the unconstrained minimization problem in (47), which can be solved using standard penalized least squares algorithms.

We next present simulation results from two settings with different structures for \mathbf{x}_{new} . We discuss the setting with \mathbf{x}_{new} generated from Gaussian distributions in Section 6.1 and the setting with decaying loadings in Section 6.2. Throughout, we consider $p = 501$ including intercept and equal sample sizes for the two treatment groups with $n_1 = n_2 = n$ and 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 $\Sigma = [0.5^{1+|j-l|}]_{(p-1) \times (p-1)}$. To generate \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 $\beta_{2,j} = 0.2(j-1)\mathbf{1}(2 \leq j \leq 6)$.

6.1. General Loading

We first consider the case with the loading \mathbf{x}_{new} being a dense vector, generated via two steps. In the first step, we generate $\mathbf{x}_{\text{basis}} = [1, \mathbf{x}_{\text{basis},-1}^\top]^\top \in \mathbb{R}^p$ with $\mathbf{x}_{\text{basis},-1} \sim N(0, \Sigma)$. In the second step, we generate \mathbf{x}_{new} based on x_{basis} in two specific ways,

- (a) In Setting 1, we generate \mathbf{x}_{new} as a shrunk version of x_{basis} with

$$x_{\text{new},j} = \mathcal{S} \cdot \mathbf{1}(j \geq 12) \cdot x_{\text{basis},j}, \quad \text{for } j = 1, \dots, p \quad (48)$$

and $\mathcal{S} \in \{1, 0.5, 0.2, 0.1, 0.05\}$.

- (b) For Setting 2, we let

$$x_{\text{new},j} = \mathbf{1}(j = 1) - \frac{2}{3}\mathbf{1}(j = 2) + \mathcal{S} \cdot \mathbf{1}(j \geq 12) \cdot x_{\text{basis},j} \quad \text{for } j = 1, \dots, p \quad (49)$$

and $\mathcal{S} \in \{1, 0.5, 0.2, 0.1, 0.05\}$.

Under the above configurations, the scale parameter \mathcal{S} controls the magnitude of the noise variables in \mathbf{x}_{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 1000 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 (RM) get smaller, the empirical rejection rate (ERR) (i.e. power) gets closer to 100%. 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 $n = 1000$. 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 coverage of the CIs are close to the nominal level.

Besides reporting the performance of the proposed estimator, we also compare the HITS method to two other estimators, 1) the plug-in Lasso estimator; and 2) the plug-in debiased estimator. For the plug-in Lasso estimator, we estimating the regression vectors by the scaled Lasso estimator implemented in the R package FLARE [Li et al., 2015]. For the plug-in debiased estimators, we introduce an alternative debiased estimator, $\widetilde{\Delta}_{\text{new}} = \mathbf{x}_{\text{new}}^{\top}(\widetilde{\beta}_1 - \widetilde{\beta}_2)$, where $\widetilde{\beta}_{k,j}$ is obtained based on the debiased plug-in estimator proposed in Javanmard and Montanari [2014] using the code posted in <https://web.stanford.edu/~montanar/sslasso/code.html>.

In Table 2, we first compare the proposed estimator (HITS) to these two plug-in estimators (plug-in Lasso estimator is shorthand as “Lasso” and plug-in debiased estimator is shorthand as “Deb”) in terms of RM (Root Mean Squared Error). Across all settings, in comparison to the plug-in debiased estimator, the proposed HITS estimator always have a smaller RM; in comparison to the plug-in Lasso estimator, the proposed estimator attained substantially smaller bias but at the expense of larger variability, reflecting the bias-variance trade-off. Specifically, when \mathcal{S} is small (taking values in $\{0.2, 0, 1, 0.05\}$), the proposed estimator generally has a smaller RM than the plug-in Lasso estimator. When $\mathcal{S} = 1, 0.5$ in which case \mathbf{x}_{new} is very dense, the proposed HITS estimator has a much larger variability compared to plug-in Lasso estimator. This suggests that under the very challenging dense scenario, a high price is paid to ensure the validity of the interval estimation.

We further compare the proposed estimator with these two plug-in estimators in terms of hypothesis testing. We first comment that the plug-in 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 the proposed HITS estimator and also the plug-in debiased estimator dominates the corresponding bias component. Due to this reason, we only report the empirical comparison between the HITS method and the plug-in debiased method. As illustrated in the coverage property, the empirical coverage of CIs based on the plug-in debiased is about 10% below the nominal level while our proposed CI achieves the nominal level.

In Table 3, we report results under setting 2 where the null hypothesis holds. We observe a similar pattern as in setting 1 for estimation precision and relative performance compared to the plug-in Lasso and debiased estimators. All the ERRs in this case, which correspond to the empirical size, are close to the nominal level 5% for the proposed HITS method while the corresponding ERRs cannot be controlled for the plug-in debiased

estimators.

6.2. Decaying Loading

In this session, we generate the decaying loading \mathbf{x}_{new} as $x_{\text{new},j} = 0.5 \cdot j^{-\delta}$ where $\delta \in \{0, 0.1, 0.25, 0.5\}$, and defer more detailed simulation results to Section D in the supplement. We report the performance of the proposed HITS method for the decaying loading in Table 4. 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 $\widehat{\Delta}_{\text{new}}$ is substantially more computationally intensive than $\widetilde{\Delta}_{\text{new}}$. The calculation of $\widehat{\Delta}_{\text{new}}$ requires four fittings of Lasso-type algorithms twice whereas $\widetilde{\Delta}_{\text{new}}$ requires $2p + 2$ fittings.

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 potentially predictive 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 and $n_2 = 91$ on methotrexate alone. To sufficiently control for potential confounders, we extracted a wide range of predictors from the EHR including past medical history and various indications of disease duration and severity on RA. We included both potential confounders and predictors of CRP in \mathbf{X} . This results in a total of $p = 171$ candidate predictors, including diagnostic codes of RA and comorbidities such as systemic lupus erythematosus (SLE) and diabetes, past history of lab results including C-reactive protein (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 med-

S	n	ERR		Coverage		Len	HITS			Lasso			Deb		
		HITS	Deb	HITS	Deb	HITS	RM	Bias	SE	RM	Bias	SE	RM	Bias	SE
1	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
	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
0.5	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
	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
0.2	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
	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
0.1	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
	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
0.05	100	0.78	0.92	0.85	0.72	1.47	0.49	0.07	0.48	0.72	0.58	0.44	0.48	0.07	0.47
	200	0.98	1.00	0.93	0.82	1.13	0.31	0.03	0.31	0.47	0.38	0.28	0.33	0.08	0.32
	400	1.00	1.00	0.95	0.86	0.80	0.21	0.02	0.21	0.33	0.27	0.19	0.23	0.03	0.23
	600	1.00	1.00	0.93	0.86	0.71	0.19	0.00	0.19	0.27	0.22	0.16	0.20	0.01	0.20
	1000	1.00	1.00	0.94	0.87	0.55	0.14	0.02	0.14	0.19	0.15	0.12	0.20	0.02	0.20

Table 2: 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 dense setting 1 where $\Delta_{\text{new}} = 1.08$. Reported also are the RM (Root Mean Squared Error), 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”).

ication 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 is indicative of disease severity. We assume that conditional on \mathbf{X} , the counterfactual outcomes are independent of the treatment assignment A .

We applied our proposed procedures to estimate the ITE for individual 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 $\{\hat{\beta}_{1,j} - \hat{\beta}_{2,j}, j = 1, \dots, p\}$, include counts of SLE COD or NLP mentions, indicator of no past history of CRP measurements, and SNPs including rs12506688, rs8043085 and

\mathcal{S}	n	ERR		Coverage		Len	HITS			Lasso			Deb		
		HITS	Deb	HITS	Deb		RM	Bias	SE	RM	Bias	SE	RM	Bias	SE
1	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
	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
0.5	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
	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
0.2	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
	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
0.1	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
	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
0.05	100	0.15	0.25	0.88	0.70	0.81	0.26	0.06	0.25	0.31	0.22	0.22	0.27	0.06	0.26
	200	0.10	0.16	0.91	0.78	0.68	0.20	0.01	0.20	0.22	0.15	0.16	0.20	0.01	0.20
	400	0.07	0.10	0.95	0.88	0.48	0.12	0.01	0.12	0.16	0.12	0.10	0.13	0.01	0.13
	600	0.05	0.11	0.95	0.88	0.40	0.10	0.00	0.10	0.13	0.09	0.08	0.12	0.00	0.12
	1000	0.06	0.11	0.95	0.91	0.31	0.08	0.00	0.08	0.10	0.07	0.07	0.12	0.01	0.12

Table 3: 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 dense setting 2 where $\Delta_{\text{new}} = 0$. Reported also are the RM (Root Mean Squared Error), 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”).

δ	n	ERR		Coverage		Len		HITS			Lasso			Deb		
		HITS	Deb	HITS	Deb	HITS	Deb	RM	Bias	SE	RM	Bias	SE	RM	Bias	SE
0	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
	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
0.1	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
	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
0.25	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
	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
0.5	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
	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 4: 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 $x_{\text{new},j} = 0.5 * j^{-\delta}$. Reported also are the RM (Root Mean Squared Error), 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”).

rs2843401. 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-positive 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.

	β_1	CI_{β_1}	β_2	CI_{β_2}	$\beta_1 - \beta_2$	$CI_{\beta_1 - \beta_2}$
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: Estimates of β_1 , β_2 and $\beta_1 - \beta_2$ for the predictors of CRP along with their 95% CIs. All predictors not included the table received zero estimate for both β_1 and β_2 .

We obtained estimates of Δ_{new} for the observed set of \mathbf{x}_{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

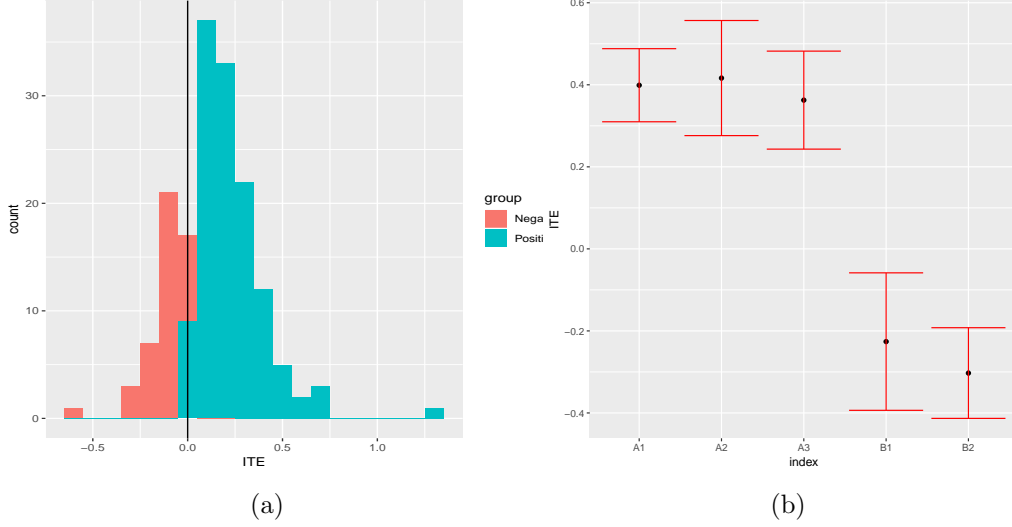


Fig. 2: (a) Histogram of the estimated ITE for the observed set of \mathbf{x}_{new} where the vertical line represents the median value; (b) point estimate and 95% CIs for 5 choices of \mathbf{x}_{new} where the x-axis indexes $\mathbf{x}_{\text{new}}^T(\hat{\beta}_1 - \hat{\beta}_2)$.

a few examples of \mathbf{x}_{new} , including (A) those with $\text{rs12506688} = 0$, $\text{rs2843401} = 0$, prior CRP not missing, $\text{rs8043085} > 0$, and ≥ 1 SLE mention; and (B) those with $\text{rs12506688} > 0$, $\text{rs2843401} > 0$, prior CRP missing, $\text{rs8043085} = 0$, 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 procedures for making inference about ITE with high-dimensional covariates. Both the theoretical and numerical properties of the HITS testing and inference procedures are established. Unlike the high dimensional debiased methods proposed in the literature, the proposed methods have the major advantage of not requiring the covariate vector \mathbf{x}_{new} to be sparse or of other specific forms. A key innovation of the proposed procedure lies in the novel construction of the projection direction for debiased where we imposed an additional constraint (8) when identifying the projection direction. We next discuss in more detail the importance of this step as well as an alternative approach to further illustrate the challenges of statistical inference for dense loading.

The following proposition shows that the algorithm without (8) fails to correct the bias of $\mathbf{x}_{\text{new}}^T \beta_1$ for a certain class of \mathbf{x}_{new} .

PROPOSITION 2. *The minimizer $\tilde{\mathbf{u}}_1$ in (6) is zero if either of the following conditions on \mathbf{x}_{new} is satisfied: (F1) $\frac{\|\mathbf{x}_{\text{new}}\|_2}{\|\mathbf{x}_{\text{new}}\|_\infty} \geq \frac{1}{\lambda_1}$; (F2) The non-zero coordinates of \mathbf{x}_{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 $\frac{\|\mathbf{x}_{\text{new}}\|_2}{\|\mathbf{x}_{\text{new}}\|_\infty}$ can be viewed as a measure of sparsity of \mathbf{x}_{new} , both Conditions (F1) and (F2) state that the optimization algorithm (6) fails to produce a non-vanishing projection direction if the loading \mathbf{x}_{new} is dense to certain degree. That is, without the additional constraint (8), the projection direction $\tilde{\mathbf{u}}_1$ does not correct the bias of estimating $\mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}_1$.

We shall provide some geometric insights about Proposition 2. The feasible set for constructing $\tilde{\mathbf{u}}$ in (6) depends on both \mathbf{x}_{new} and $\|\mathbf{x}_{\text{new}}\|_2$. If the dimension 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}_{new} and $\mathbf{x}_{\text{new}} = \mathbf{e}_i$, respectively. Since zero vector is contained in the feasible set for a dense \mathbf{x}_{new} , the optimizer in (6) is zero and the bias correction is not effective. With the additional constraint (8), even in the presence of dense \mathbf{x}_{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 (8). Interestingly, the additional constraint (8) simply restricts the feasible set from one additional direction determined by \mathbf{x}_{new} and automatically enables a unified inference procedure for an arbitrary \mathbf{x}_{new} .

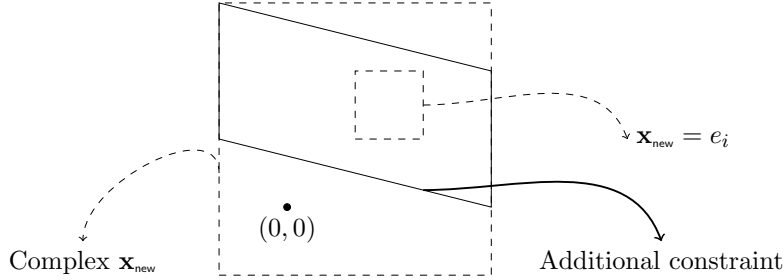


Fig. 3: The solid parallelogram corresponds to the feasible set of (7) and (8) for a dense \mathbf{x}_{new} while the large dashed triangle corresponds to that of (6); the small dashed square corresponds to the feasible set of (6) for $\mathbf{x}_{\text{new}} = \mathbf{e}_i$.

The current paper derives inference procedures for ITE based on high dimensional outcome models. This allows us to include high dimensional confounders to overcome treatment by indication bias frequently encountered in observational studies. The HITS procedure is derived under a supervised linear regression 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, the unlabeled data can be used to significantly increase the inference accuracy [Cai and Guo, 2018c]. The problem of ITE inference in the semi-supervised setting warrants future research.

9. Proofs

In this section, we present the proof for the optimality results, Theorem 3 and Corollary 3 and also the proof for non-vanishing variance, Lemma 1. All other proofs are deferred to Section C in the supplementary material.

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 (LOWER BOUND FOR DETECTION BOUNDARY). *Suppose $s \leq s_u \lesssim \min\{p, \frac{n}{\log p}\}$. Re-order \mathbf{x}_{new} such that $|x_{\text{new},1}| \geq |x_{\text{new},2}| \geq \dots \geq |x_{\text{new},p}|$. For any (q, L) satisfies $q \leq s_u$ and $L \leq \|\mathbf{x}_{\text{new}}\|_0$, the adaptation detection boundary $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ satisfies*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \tau^* \quad (50)$$

where

$$\tau^* = C \frac{1}{\sqrt{n}} \cdot \max \left\{ \sqrt{\sum_{j=1}^s x_{\text{new},j}^2}, \sum_{j=\max\{L-q+2, 1\}}^L |x_{\text{new},j}| \sqrt{(\log(cL/q^2))_+} \right\}. \quad (51)$$

To establish the lower bounds in Theorem 3 and Corollary 3, we simply apply the lower bound for the general detection boundary in (51) to the case of exact loadings. Specifically, τ^* is reduced to the following expression by taking $L = \|\mathbf{x}_{\text{new}}\|_0$ and $q = \min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\}$,

$$\tau^* = \frac{\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} \cdot \max \left\{ \min\{\sqrt{s}, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\}, \min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\} \sqrt{\left(\log \left(\frac{c\|\mathbf{x}_{\text{new}}\|_0}{\min\{s_u, \sqrt{\|\mathbf{x}_{\text{new}}\|_0}\}^2} \right) \right)_+} \right\}. \quad (52)$$

For the case (E1), we have $\|\mathbf{x}_{\text{new}}\|_0 \leq s^2 \leq s_u^2$ and $\tau^* \asymp \frac{\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} \sqrt{\|\mathbf{x}_{\text{new}}\|_0}$; hence the lower bound (38) follows. For the case (E2), if $\gamma_{\text{new}} > 2\gamma_u$, we have $\tau^* \asymp \frac{\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} s_u \sqrt{\log p}$; if $\gamma_{\text{new}} \leq 2\gamma_u$, we have $\tau^* \asymp \frac{\|\mathbf{x}_{\text{new}}\|_\infty}{\sqrt{n}} \sqrt{\|\mathbf{x}_{\text{new}}\|_0}$. Hence, the lower bounds in (39) follow.

By applying Corollary 1, we establish that the detection boundaries $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ in (40) and (42) 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 (74) in the supplement.

9.2. Proof of Lemma 1

The upper bound part of (19), $\sqrt{V} \leq C_0 \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right)$, follows from the following lemma, which is the second high probability inequality of Lemma 4 established in Cai and Guo [2017].

LEMMA 2. With probability larger than $1 - p^{-c}$,

$$\left\| \widehat{\Sigma}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}} \right\|_{\infty} \leq C \|\mathbf{x}_{\text{new}}\|_2 \sqrt{\frac{\log p}{n_k}} \text{ for } k = 1, 2. \quad (53)$$

By Lemma 2, $\Sigma_1^{-1} \mathbf{x}_{\text{new}}$ satisfies the constraints (7) and (8) and hence

$$V \leq \frac{\sigma_1^2}{n_1} \mathbf{x}_{\text{new}}^{\top} \Sigma_1^{-1} \widehat{\Sigma}_1 \Sigma_1^{-1} \mathbf{x}_{\text{new}} + \frac{\sigma_2^2}{n_2} \mathbf{x}_{\text{new}}^{\top} \Sigma_2^{-1} \widehat{\Sigma}_2 \Sigma_2^{-1} \mathbf{x}_{\text{new}}. \quad (54)$$

By Lemma 10 (specifically, the last high probability inequality) of Cai and Guo [2018c], with probability larger than $1 - p^{-c}$, we have

$$\left| \frac{\mathbf{x}_{\text{new}}^{\top} \Sigma_1^{-1} \widehat{\Sigma}_1 \Sigma_1^{-1} \mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^{\top} \Sigma_1^{-1} \mathbf{x}_{\text{new}}} - 1 \right| \lesssim \sqrt{\frac{\log p}{n_1}} \text{ and } \left| \frac{\mathbf{x}_{\text{new}}^{\top} \Sigma_2^{-1} \widehat{\Sigma}_2 \Sigma_2^{-1} \mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^{\top} \Sigma_2^{-1} \mathbf{x}_{\text{new}}} - 1 \right| \lesssim \sqrt{\frac{\log p}{n_2}} \quad (55)$$

Then we establish $\mathbb{P} \left(\sqrt{V} \leq C_0 \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right) \right) \geq 1 - p^{-c}$.

The proof of the lower bound part $\sqrt{V} \geq c_0 \|\mathbf{x}_{\text{new}}\|_2$ is facilitated by the optimization constraint (7). To be specific, we define a proof-facilitating optimization problem,

$$\bar{\mathbf{u}}_1 = \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^{\top} \widehat{\Sigma}_1 \mathbf{u} \quad \text{subject to} \quad \left| \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2^2 \right| \leq \|\mathbf{x}_{\text{new}}\|_2^2 \lambda_1 \quad (56)$$

Note that $\widehat{\mathbf{u}}_1$ satisfies the feasible set of (56) and hence

$$\widehat{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \widehat{\mathbf{u}}_1 \geq \bar{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \bar{\mathbf{u}}_1 \geq \bar{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \bar{\mathbf{u}}_1 + t \left((1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2 - \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \bar{\mathbf{u}}_1 \right) \text{ for any } t \geq 0, \quad (57)$$

where the last inequality follows from the constraint of (56). Note that for any given $t \geq 0$,

$$\bar{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \bar{\mathbf{u}}_1 + t \left((1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2 - \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \bar{\mathbf{u}}_1 \right) \geq \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^{\top} \widehat{\Sigma}_1 \mathbf{u} + t \left((1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2 - \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{u} \right). \quad (58)$$

By solving the minimization problem of the right hand side of (58), we have the minimizer u^* satisfies $\widehat{\Sigma}_1 u^* = \frac{t}{2} \widehat{\Sigma}_1 \mathbf{x}_{\text{new}}$ and hence the minimum of the right hand side of (58) is

$$-\frac{t^2}{4} \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{x}_{\text{new}} + t(1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2.$$

Combined with (57) and (58), we have

$$\widehat{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \widehat{\mathbf{u}}_1 \geq \max_{t \geq 0} \left[-\frac{t^2}{4} \mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{x}_{\text{new}} + t(1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2 \right]. \quad (59)$$

For $t^* = 2 \frac{(1 - \lambda_1) \|\mathbf{x}_{\text{new}}\|_2^2}{\mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{x}_{\text{new}}} > 0$, the minimum of the right hand side of (59) is achieved and hence

$$\widehat{\mathbf{u}}_1^{\top} \widehat{\Sigma}_1 \widehat{\mathbf{u}}_1 \geq \frac{(1 - \lambda_1)^2 \|\mathbf{x}_{\text{new}}\|_2^4}{\mathbf{x}_{\text{new}}^{\top} \widehat{\Sigma}_1 \mathbf{x}_{\text{new}}}. \quad (60)$$

By Lemma 10 (specifically, the last high probability inequality) of Cai and Guo [2018c], with probability larger than $1 - p^{-c}$, we have $\left| \frac{\mathbf{x}_{\text{new}}^\top \hat{\Sigma}_1 \mathbf{x}_{\text{new}}}{\mathbf{x}_{\text{new}}^\top \Sigma_1 \mathbf{x}_{\text{new}}} - 1 \right| \lesssim \sqrt{\log p / n_1}$ and hence $\hat{\mathbf{u}}_1^\top \hat{\Sigma}_1 \hat{\mathbf{u}}_1 \geq c \|\mathbf{x}_{\text{new}}\|_2^2$. Similarly, we can establish $\hat{\mathbf{u}}_2^\top \hat{\Sigma}_2 \hat{\mathbf{u}}_2 \geq c \|\mathbf{x}_{\text{new}}\|_2^2$ and hence

$$\mathbb{P} \left(\sqrt{V} \geq c_0 \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right) \right) \geq 1 - p^{-c}. \quad (61)$$

References

- Kathy S Albain, William E Barlow, Steven Shak, Gabriel N Hortobagyi, Robert B Livingston, I-Tien Yeh, Peter Ravdin, Roberto Bugarini, Frederick L Baehner, Nancy E Davidson, et al. 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, 2010.
- Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.
- Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
- Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.
- Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- Tim Bongartz, Alex J Sutton, Michael J Sweeting, Iain Buchan, Eric L Matteson, and Victor Montori. 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, 2006.
- Ferdinand C Breedveld, Michael H Weisman, Arthur F Kavanaugh, Stanley B Cohen, Karel Pavelka, Ronald van Vollenhoven, John Sharp, John L Perez, and George T Spencer-Green. 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, 2006.
- T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, 45(2):615–646, 2017.
- T Tony Cai and Zijian Guo. Accuracy assessment for high-dimensional linear regression. *The Annals of Statistics*, 46(4):1807–1836, 2018a.

- T Tony Cai and Zijian Guo. Supplement to “accuracy assessment for high-dimensional linear regression”. 2018b.
- T Tony Cai and Zijian Guo. Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *arXiv preprint arXiv:1806.06179*, 2018c.
- Leonard H Calabrese, Cassandra Calabrese, and Elizabeth Kirchner. 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, 2016.
- Emmanuel Candès and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- Lorraine A Chantry, Adnan M Nagrial, Clare Watson, Amber L Johns, Mona Martyn-Smith, Skye Simpson, Scott Mead, Marc D Jones, Jaswinder S Samra, Anthony J Gill, et al. Precision medicine for advanced pancreas cancer: the individualized molecular pancreatic cancer therapy (impact) trial. *Clinical Cancer Research*, 21(9):2029–2037, 2015.
- A Chatzikyriakidou, I Georgiou, PV Voulgari, AI Venetsanopoulou, and AA Drosos. Combined tumour necrosis factor- α and tumour necrosis factor receptor genotypes could predict rheumatoid arthritis patients’ response to anti-tnf- α therapy and explain controversies of studies based on a single polymorphism. *Rheumatology (Oxford, England)*, 46(6):1034, 2007.
- Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.
- François Danion, Laetitia Sparsa, Laurent Arnaud, Ghada Alsaleh, François Lefebvre, Vincent Gies, Thierry Martin, Cédric Lukas, Jean Durckel, Marc Ardizzone, et al. 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, 2017.
- David A Eberhard, Bruce E Johnson, Lukas C Amler, Audrey D Goddard, Sherry L Heldens, Roy S Herbst, William L Ince, Pasi A Jänne, Thomas Januario, David H Johnson, et al. 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, 2005.
- Paul Emery, Ferdinand C Breedveld, Stephen Hall, Patrick Durez, David J Chang, Deborah Robertson, Amitabh Singh, Ronald D Pedersen, Andrew S Koenig, and Bruce Freundlich. 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, 2008.

- William E Evans and Mary V Relling. Moving towards individualized medicine with pharmacogenomics. *Nature*, 429(6990):464, 2004.
- Steve Eyre, John Bowes, Dorothée Diogo, Annette Lee, Anne Barton, Paul Martin, Alexandra Zhernakova, Eli Stahl, Sebastien Viatte, Kate McAllister, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, 44(12):1336, 2012.
- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- Vivian S Gainer, Andrew Cagan, Victor M Castro, Stacey Duey, Bhaswati Ghosh, Alyssa P Goodson, Sergey Goryachev, Reeta Metta, Taowei David Wang, Nich Watanasin, et al. 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, 2016.
- Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.
- Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
- Isaac S Kohane, Susanne E Churchill, and Shawn N Murphy. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2):181–185, 2012.
- Nicholas B La Thangue and David J Kerr. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nature Reviews Clinical Oncology*, 8(10):587–596, 2011.
- Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in r. *The Journal of Machine Learning Research*, 16(1):553–557, 2015.
- Katherine P Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62(8):1120–1127, 2010.
- Chunyu Liu, Franak Batliwalla, Wentian Li, Annette Lee, Ronenn Roubenoff, Evan Beckman, Houman Khalili, Aarti Damle, Marlena Kern, Robert M Plenge, et al. Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-tnf treatment in rheumatoid arthritis. *Molecular medicine*, 14(9-10):575, 2008.
- Hojin Moon, Hongshik Ahn, Ralph L Kodell, Songjoon Baek, Chien-Ju Lin, and James J Chen. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine*, 41(3):197–207, 2007.

- Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876, 2013.
- FS Ong, K Das, J Wang, H Vakil, JZ Kuo, WL Blackwell, SW Lim, MO Goodarzi, KE Bernstein, JI Rotter, et al. Personalized medicine and pharmacogenetic biomarkers: progress in molecular oncology testing. *Expert Review of Molecular Diagnostics*, 12(6):593–602, 2012.
- L Padyukov, J Lampa, M Heimbürger, S Ernestam, T Cederholm, I Lundkvist, P Andersson, Y Hermansson, A Harju, L Klareskog, et al. Genetic markers for the efficacy of tumour necrosis factor blocking therapy in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 62(6):526–529, 2003.
- Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180–1210, 2011.
- Seth D Seegobin, Margaret HY Ma, Chanaka Dahanayake, Andrew P Cope, David L Scott, Cathryn M Lewis, and Ian C Scott. 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, 2014.
- George Simon, Anupama Sharma, Xueli Li, Todd Hazelton, Frank Walsh, Charles Williams, Alberto Chiappori, Eric Haura, Tawee Tanvetyanon, Scott Antonia, et al. Feasibility and efficacy of molecular analysis-directed individualized therapy in advanced non-small-cell lung cancer. *Journal of Clinical Oncology*, 25(19):2741–2746, 2007.
- Rui Song, Michael Kosorok, Donglin Zeng, Yingqi Zhao, Eric Laber, and Ming Yuan. On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4(1):59–68, Feb 2015. ISSN 2049-1573. doi: 10.1002/sta4.78. URL <http://dx.doi.org/10.1002/sta4.78>.
- Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.
- Peter C Taylor and Marc Feldmann. Anti-tnf biologic agents: still the therapy of choice for rheumatoid arthritis. *Nature Reviews Rheumatology*, 5(10):578, 2009.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
- Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
- Désirée van der Heijde, Lars Klareskog, Vicente Rodriguez-Valverde, Catalin Codreanu, Horatiu Bolosiu, Jose Melo-Gomes, Jesus Tornero-Molina, Joseph Wajdula, Ronald Pedersen, Saeed Fatenejad, et al. 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, 2006.

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

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

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 (37). 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 $s \leq s_u \lesssim \frac{n}{\log p}$. We calibrate s, s_u and the decaying of \mathbf{x}_{new} by γ, γ_u and δ , respectively, as defined in (35) and (37).*

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \frac{1}{\sqrt{n}} \cdot (1 + \sqrt{\log s} \cdot \mathbf{1}(\delta = \frac{1}{2})) \quad (62)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \gtrsim \begin{cases} c_p \frac{s_u^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta} & \text{if } \gamma_u < \frac{1}{2} \\ \sqrt{\frac{p^{1-2\delta}}{n}} & \text{if } \gamma_u \geq \frac{1}{2} \end{cases} \quad (63)$$

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

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 $s \leq s_u \lesssim \frac{\sqrt{n}}{\log p}$.*

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{\|\mathbf{x}_{\text{new}}\|_2}{\sqrt{n}} \quad (64)$$

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}_{new} is a slow decaying loading with $0 \leq \delta < \frac{1}{2}$, 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 $\gamma < \gamma_u \leq \frac{1}{2}$, then*

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \frac{s_u^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}. \quad (65)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \sqrt{\frac{p^{1-2\delta}}{n}} \gg \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \frac{s^{1-2\delta}}{\sqrt{n}} (\log p)^{\frac{1}{2}-\delta}. \quad (66)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \asymp \tau_{\text{mini}}(s, \mathbf{x}_{\text{new}}) \asymp \sqrt{\frac{p^{1-2\delta}}{n}}. \quad (67)$$

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}_{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 corresponds 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}_{new} , which parallels the corollary 4 in the main paper for the case of exact loading,

COROLLARY 7. *Suppose that $s \leq s_u \lesssim \frac{\sqrt{n}}{\log p}$ and $\gamma_u \geq \frac{1}{2}$. 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}_{new} satisfying (37).*

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}_{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}_{new} is ordered as follows,

$$|\mathbf{x}_{\text{new},1}| \geq |\mathbf{x}_{\text{new},2}| \geq \cdots \geq |\mathbf{x}_{\text{new},p}|. \quad (68)$$

For $k = 1, 2$, we define $\tilde{\beta}_{k,j}$ to be the de-biased estimator introduced by Javanmard and Montanari [2014] with the corresponding covariance matrix of $\tilde{\beta}_{k,\cdot} \in \mathbb{R}^p$ as $\widehat{\text{Var}}^k$. Define $\mathbf{d} = \beta_1 - \beta_2$. For $1 \leq j \leq p$, we introduce the hard-thresholding estimator $\hat{\mathbf{d}}_j$ as

$$\hat{\mathbf{d}}_j = (\tilde{\beta}_{1,j} - \tilde{\beta}_{2,j}) \times \mathbf{1} \left\{ \left| \frac{\tilde{\beta}_{1,j} - \tilde{\beta}_{2,j}}{\sqrt{\widehat{\text{Var}}_{j,j}^1 + \widehat{\text{Var}}_{j,j}^2}} \right| \geq \sqrt{2.01 \log p} \right\}$$

Define the vector ξ as a sparsified version of \mathbf{x}_{new} ,

$$\xi_j = x_{\text{new},j} \text{ for } 1 \leq j \leq q, \quad \xi_j = 0 \text{ for } q+1 \leq j \leq p, \quad (69)$$

and also

$$\hat{\xi}_j = 0 \text{ for } 1 \leq j \leq q, \quad \hat{\xi}_j = \mathbf{x}_{\text{new},j} \sqrt{\widehat{\text{Var}}_{j,j}^1 + \widehat{\text{Var}}_{j,j}^2} \text{ for } q+1 \leq j \leq p, \quad (70)$$

where q is an integer to be specified later. We define the estimator $\widehat{\xi^\top \mathbf{d}}$ as in (11) with \mathbf{x}_{new} replaced with the sparsified loading ξ . In particular, the projection directions $\check{\mathbf{u}}_k$ for $k = 1, 2$ are constructed as

$$\check{\mathbf{u}}_k = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \mathbf{u}^\top \hat{\Sigma}_k \mathbf{u} \quad \text{subject to} \quad \left\| \hat{\Sigma}_k \mathbf{u} - \xi \right\|_\infty \leq \|\xi\|_2 \lambda_k$$

$$\left| \xi^\top \hat{\Sigma}_k \mathbf{u} - \|\xi\|_2^2 \right| \leq \|\xi\|_2^2 \lambda_k,$$

where $\lambda_k \asymp \sqrt{\log p / n_k}$. Define the sum $S(s_u) = \sum_{j=1}^{s_u} |\hat{\xi}_{(j)}|$, where $\hat{\xi}_{(j)}$ as the j -th largest absolute value of $\hat{\xi}$. We propose the following estimator

$$\check{\Delta}_{\text{new}} = \widehat{\xi^\top \mathbf{d}} + \sum_{j=q+1}^p x_{\text{new},j} \hat{\mathbf{d}}_j. \quad (71)$$

We construct the CI as

$$\text{CI}(q, s_u) = (\check{\Delta}_{\text{new}} - \rho_{\alpha/2}(q, s_u), \check{\Delta}_{\text{new}} + \rho_{\alpha/2}(q, s_u)) \quad (72)$$

where $\rho_{\alpha/2}(q, s_u) = z_{\alpha/2} \sqrt{\tilde{V}} + S(s_u) \sqrt{2.01 \log p}$, with \tilde{V} defined as

$$\tilde{V} = \frac{\hat{\sigma}_1^2}{n_1} \check{\mathbf{u}}_1^\top \hat{\Sigma}_1 \check{\mathbf{u}}_1 + \frac{\hat{\sigma}_2^2}{n_2} \check{\mathbf{u}}_2^\top \hat{\Sigma}_2 \check{\mathbf{u}}_2. \quad (73)$$

We propose the following decision rule,

$$\phi(q, s_u) = \mathbf{1}(\check{\Delta}_{\text{new}} - \rho_\alpha(q, s_u) > 0). \quad (74)$$

where $\rho_\alpha(q, s_u) = z_\alpha \sqrt{\check{V}} + S(s_u) \sqrt{2.01 \log p}$,

For the estimator defined in (71), we have the following error decomposition,

$$\check{\Delta}_{\text{new}} - \Delta_{\text{new}} = \widehat{\xi^\top \mathbf{d}} - \xi^\top \mathbf{d} + \sum_{j=q+1}^p x_{\text{new},j}(\widehat{\mathbf{d}}_j - \mathbf{d}_j), \quad (75)$$

It follows from Theorem 2 that

$$\frac{\widehat{\xi^\top \mathbf{d}} - \xi^\top \mathbf{d}}{\sqrt{\check{V}}} \rightarrow N(0, 1) \quad (76)$$

In addition, with probability larger than $1 - p^{-c}$,

$$\left| \sum_{j=q+1}^p x_{\text{new},j}(\widehat{\mathbf{d}}_j - \mathbf{d}_j) \right| \leq S(s_u) \sqrt{2.01 \log p} \quad (77)$$

Combining (75), (76) and (77), we establish the coverage property of the confidence interval $\text{CI}(q, s_u)$ proposed in (72) and also control the type I error of the testing procedure $\phi(q, s_u)$ defined in (74). It remains to control the length of $\rho_\alpha(q, s_u)$. We focus on the decaying loading $|x_{\text{new},j}| \asymp j^{-\delta}$ for $0 \leq \delta < \frac{1}{2}$. Following from (97), we have

$$\sqrt{\check{V}} \asymp \|\xi\|_2 \lesssim \sqrt{\sum_{j=1}^q |x_{\text{new},j}|^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \lesssim q^{\frac{1}{2}-\delta} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (78)$$

and

$$S(s_u) = \sum_{j=1}^{s_u} |\widehat{\xi}_{(j)}| \lesssim \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sum_{j=q+1}^{q+s_u} |x_{\text{new},j}| \lesssim s_u \cdot (q + cs_u)^{-\delta} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (79)$$

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

$$|\rho_\alpha(q, s_u)| \lesssim (s_u 2 \log p)^{\frac{1}{2}-\delta} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (80)$$

For the exact loading, we can also take $q = 0$.

C. Additional Proofs

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

C.1. Proof of Theorem 4

Suppose that we observe a random variable Z which has a distribution \mathbf{P}_θ 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}_{π_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

$$\chi^2(f_1, f_0) = \int \frac{(f_1(z) - f_0(z))^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1 \quad (81)$$

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

$$L_1(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}. \quad (82)$$

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

$$\inf_{\theta \in \mathcal{F}_1} \mathbb{E}_\theta \phi \leq L_1(f_{\pi_1}, f_{\pi_0}) + \sup_{\theta \in \mathcal{F}_0} \mathbb{E}_\theta \phi \quad (83)$$

In addition, suppose that $L_1(f_{\pi_1}, f_{\pi_0}) < 1 - \alpha - \eta$ 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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbb{T}(\theta_0) - \mathbb{T}(\theta_1)|. \quad (84)$$

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

$$\mathbb{E}_{\pi_1} \phi - \mathbb{E}_{\pi_0} \phi \leq L_1(f_{\pi_1}, f_{\pi_0}). \quad (85)$$

Then (83) follows from

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

where the first and last inequalities follows from the definition of inf and sup and the second inequality follows from (85). The lower bound in (84) follows from the definition of $\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}})$ and the fact that

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

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,

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \frac{1}{\sqrt{n}} \sum_{j=\max\{L-q+2, 1\}}^L |x_{\text{new},j}| \sqrt{(\log(cL/q^2))_+}. \quad (86)$$

For this case, we assume that $q \leq \sqrt{cL}$; otherwise the lower bound in (86) is trivial. To simplify the notation of the proof, we fix $\beta_2 = 0$ and denote $\beta_1 = \boldsymbol{\eta}$. In addition, we set $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$ and $\sigma_2 = 1$. Without loss of generality, we set $x_{\text{new},i} \geq 0$ and $x_{\text{new},i} \geq x_{\text{new},i+1}$. 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 $L_1(f_{\pi_1}, f_{\pi_0})$
- (c) to maximize the functional distance $\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbf{T}(\theta_0) - \mathbf{T}(\theta_1)|$

To establish the lower bound (86), we construct the following parameter spaces,

$$\begin{aligned} \mathcal{F}_0 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta}, 1, \mathbf{I} \\ \mathbf{0}, 1, \mathbf{I} \end{pmatrix} : \boldsymbol{\eta}_1 = \rho \cdot \frac{\sum_{j=L-q+2}^L x_{\text{new},j}}{x_{\text{new},1}}, \|\boldsymbol{\eta}_{-1}\|_0 = q-1, \boldsymbol{\eta}_j \in \{0, -\rho\} \text{ for } 2 \leq j \leq L \right\} \\ \mathcal{F}_1 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta}, 1, \mathbf{I} \\ \mathbf{0}, 1, \mathbf{I} \end{pmatrix} : \boldsymbol{\eta}_1 = \rho \cdot \frac{\sum_{j=L-q+2}^L x_{\text{new},j}}{x_{\text{new},1}}, \boldsymbol{\eta}_{-1} = \mathbf{0} \right\} \end{aligned} \quad (87)$$

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

$$\mathcal{F}_0 \subset \mathcal{H}_0(s_u) \quad \text{and} \quad \mathcal{F}_1 \subset \mathcal{H}_1(s, \tau) \quad \text{for} \quad \tau = \rho \cdot \sum_{j=L-q+2}^L x_{\text{new},j} \quad (88)$$

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

$$L_1(f_{\pi_1}, f_{\pi_0}) \leq \sqrt{\chi^2(f_{\pi_0}, f_{\pi_1})}. \quad (89)$$

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 (89). 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 $\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbf{T}(\theta_0) - \mathbf{T}(\theta_1)|$, where $\mathbf{T} = \Delta_{\text{new}}$ and hence we have

$$\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbf{T}(\theta_0) - \mathbf{T}(\theta_1)| \gtrsim \sqrt{\frac{2\log[L/q^2]}{n}} \cdot \sum_{j=L-q+2}^L x_{\text{new},j} \quad (90)$$

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

$$\tau_{\text{adap}}(s_u, s, \mathbf{x}_{\text{new}}) \geq \frac{1}{\sqrt{n}} \cdot \sqrt{\sum_{j=1}^s x_{\text{new},j}^2}. \quad (91)$$

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 $\beta_2 = 0$ and denote $\beta_1 = \boldsymbol{\eta}$. In addition, we set $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$ and $\sigma_2 = 1$. Without loss of generality, we set $x_{\text{new},i} \geq 0$ and $x_{\text{new},i} \geq x_{\text{new},i+1}$. To establish the lower bound (86), we construct the following parameter space,

$$\begin{aligned} \mathcal{F}_0 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} 0, 1, \mathbf{I} \\ \mathbf{0}, 1, \mathbf{I} \end{pmatrix} \right\} \\ \mathcal{F}_1 &= \left\{ \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta}, 1, \mathbf{I} \\ \mathbf{0}, 1, \mathbf{I} \end{pmatrix} : \boldsymbol{\eta}_j = \rho \cdot \frac{x_{\text{new},j}}{\sqrt{\sum_{j=1}^s x_{\text{new},j}^2}} \quad \text{for } 1 \leq j \leq s \right\} \end{aligned} \quad (92)$$

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

$$\mathcal{F}_0 \subset \mathcal{H}_0(s_u) \quad \text{and} \quad \mathcal{F}_1 \subset \mathcal{H}_1(s, \tau) \quad \text{for} \quad \tau = \rho \cdot \sqrt{\sum_{j=1}^s x_{\text{new},j}^2} \quad (93)$$

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

$$\chi^2(f_{\pi_1}, f_{\pi_0}) \leq \exp(2n\rho^2) - 1 \quad (94)$$

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 $\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbf{T}(\theta_0) - \mathbf{T}(\theta_1)|$, where $\mathbf{T} = \Delta_{\text{new}}$ and hence we have

$$\min_{\theta_0 \in \mathcal{F}_0, \theta_1 \in \mathcal{F}_1} |\mathbf{T}(\theta_0) - \mathbf{T}(\theta_1)| \gtrsim \frac{1}{\sqrt{n}} \cdot \sqrt{\sum_{j=1}^s x_{\text{new},j}^2}. \quad (95)$$

C.2. Proof of Theorem 5 and Corollary 6

The lower bound is an application of the general detection boundary (51), which is translated to the following lower bound,

$$\tau \leq \tau^* = \frac{1}{\sqrt{n}} \cdot \max \left\{ \sqrt{\sum_{j=1}^s j^{-2\delta}}, \sum_{j=\max\{L-q+2, 1\}}^L j^{-\delta} \sqrt{(\log(cL/q^2))_+} \right\}. \quad (96)$$

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

$$\sum_{j=l_1}^{l_2} j^{-\delta} \in \begin{cases} \frac{1}{\delta-1} [(l_1-1)^{1-\delta} - l_2^{1-\delta}, l_1^{1-\delta} - (l_2+1)^{1-\delta}] & \text{for } \delta > 1 \\ [\log \frac{l_2+1}{l_1}, \log \frac{l_2}{l_1-1}] & \text{for } \delta = 1 \\ \frac{1}{1-\delta} [(l_2+1)^{1-\delta} - l_1^{1-\delta}, l_2^{1-\delta} - (l_1-1)^{1-\delta}] & \text{for } \delta < 1 \end{cases} \quad (97)$$

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 (62) follows.

For the case (D2), we first consider $\gamma_u \geq \frac{1}{2}$, we take $q = \sqrt{p}$ in (96) and have

$$\sum_{j=\max\{p-q+2, 1\}}^p j^{-\delta} \geq \frac{1}{1-\delta} \left((p+1)^{1-\delta} - (p-\sqrt{p}-2)^{1-\delta} \right) \asymp (p-c\sqrt{p})^{-\delta} \sqrt{p} \asymp p^{\frac{1}{2}-\delta}. \quad (98)$$

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

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

Combined with (98), we establish the lower bound (63). 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 (64), (66) and (67) 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 (65) and $\tau_{\text{mini}}(s, \mathbf{x}_{\text{new}})$ in (66) are achieved by the hypothesis testing procedure introduced in (74).

C.3. Proof of Theorems 1 and 2

By combining the error decompositions for $k = 1$ and $k = 2$ in (16), we have

$$\begin{aligned} \widehat{\mathbf{x}}_{\text{new}}^\top \widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{x}}_{\text{new}}^\top \widehat{\boldsymbol{\beta}}_2 - \mathbf{x}_{\text{new}}^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) &= \widehat{\mathbf{u}}_1^\top \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{1,i\epsilon_{1,i}} - \widehat{\mathbf{u}}_2^\top \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2,i\epsilon_{2,i}} \\ &\quad + \left(\widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) - \left(\widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2 - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \end{aligned} \quad (99)$$

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

$$\left| \left(\widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \right| \leq \left\| \widehat{\boldsymbol{\Sigma}}_k \widehat{\mathbf{u}}_k - \mathbf{x}_{\text{new}} \right\|_\infty \cdot \|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_1 \lesssim \|\mathbf{x}_{\text{new}}\|_2 \sqrt{\frac{\log p}{n_k}} \cdot \|\boldsymbol{\beta}_k\|_0 \sqrt{\frac{\log p}{n_k}},$$

where the second inequality follows from the optimization constraint (7) and the condition (B1). Hence, we establish that, with probability larger than $1 - g(n_1, n_2)$,

$$\begin{aligned} &\left| \left(\widehat{\boldsymbol{\Sigma}}_1 \widehat{\mathbf{u}}_1 - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) - \left(\widehat{\boldsymbol{\Sigma}}_2 \widehat{\mathbf{u}}_2 - \mathbf{x}_{\text{new}} \right)^\top (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \right| \\ &\lesssim \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{\|\boldsymbol{\beta}_1\|_0 \log p}{n_1} + \frac{\|\boldsymbol{\beta}_2\|_0 \log p}{n_2} \right). \end{aligned} \quad (100)$$

Note that

$$\mathbb{E}_{|\mathbb{X}} \left(\widehat{\mathbf{u}}_1^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i\epsilon_{1,i}} - \widehat{\mathbf{u}}_2^\top \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i\epsilon_{2,i}} \right)^2 = V, \quad (101)$$

where V is defined in (12). By (19), with probability larger than $1 - p^{-c} - \frac{1}{t^2}$, then

$$\left| \widehat{\mathbf{u}}_1^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i\epsilon_{1,i}} - \widehat{\mathbf{u}}_2^\top \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i\epsilon_{2,i}} \right| \lesssim t \|\mathbf{x}_{\text{new}}\|_2 \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \right) \quad (102)$$

Combing (100) and (102), we establish Theorem 1.

By assuming the condition (A3), we show that conditioning on \mathbb{X} , $\widehat{\mathbf{u}}_1^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i\epsilon_{1,i}} - \widehat{\mathbf{u}}_2^\top \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i\epsilon_{2,i}} \sim N(0, V)$ where V is defined in (12). After normalization, we have

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

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

$$\frac{1}{\sqrt{V}} \left(\widehat{\mathbf{u}}_1^\top \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_{1,i\epsilon_{1,i}} - \widehat{\mathbf{u}}_2^\top \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_{2,i\epsilon_{2,i}} \right) \sim N(0, 1) \quad (103)$$

Combing (100) with (19), we show that with probability larger than $1 - p^{-c} - g(n_1, n_2)$,

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

Together with (103), we establish Theorem 2.

C.4. Proof of Proposition 1

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

$$\begin{aligned} L(\mathbf{u}, \tau, \eta, \tau_0, \eta_0) &= \mathbf{u}^\top \hat{\Sigma} \mathbf{u} + \tau^\top \left(\hat{\Sigma} \mathbf{u} - \mathbf{x}_{\text{new}} - \|\mathbf{x}_{\text{new}}\|_2 \lambda \mathbf{1} \right) + \eta^\top \left(\mathbf{x}_{\text{new}} - \hat{\Sigma} \mathbf{u} - \|\mathbf{x}_{\text{new}}\|_2 \lambda \mathbf{1} \right) \\ &+ \tau_0 \left(\frac{\mathbf{x}_{\text{new}}^\top}{\|\mathbf{x}_{\text{new}}\|_2} \hat{\Sigma} \mathbf{u} - (1 + \lambda) \|\mathbf{x}_{\text{new}}\|_2 \right) + \eta_0 \left((1 - \lambda) \|\mathbf{x}_{\text{new}}\|_2 - \frac{\mathbf{x}_{\text{new}}^\top}{\|\mathbf{x}_{\text{new}}\|_2} \hat{\Sigma} \mathbf{u} \right) \end{aligned} \quad (104)$$

where $\tau \in \mathbb{R}^p$, $\eta \in \mathbb{R}^p$ and $\{\tau_i, \eta_i\}_{0 \leq i \leq p}$ are positive constants. Then we derive the dual function $g(\tau, \eta, \tau_0, \eta_0) = \arg \min_{\mathbf{u}} L(\mathbf{u}, \tau, \eta, \tau_0, \eta_0)$. 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

$$2\hat{\Sigma} \mathbf{u}^* + \hat{\Sigma} \left[(\tau - \eta) + (\tau_0 - \eta_0) \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2} \right] = 0. \quad (105)$$

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

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

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

$$\max_{\eta_i \geq \max\{0, -\gamma_i\}} \gamma_i + 2\eta_i = |\gamma_i|. \quad (106)$$

By applying (106), we establish

$$\begin{aligned} \max_{\gamma, \eta_0} g(\gamma, \eta, \gamma_0, \eta_0) &= -\frac{1}{4} \left[\gamma + \gamma_0 \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2} \right]^\top \hat{\Sigma} \left[\gamma + \gamma_0 \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2} \right] \\ &- \mathbf{x}_{\text{new}}^\top \left(\gamma + \gamma_0 \frac{\mathbf{x}_{\text{new}}}{\|\mathbf{x}_{\text{new}}\|_2} \right) - \lambda \|\mathbf{x}_{\text{new}}\|_2 (|\gamma_0| + \|\gamma\|_1) \end{aligned} \quad (107)$$

Then it is equivalent to solve the dual problem defined in (47). By (105), we establish (46).

C.5. Proof of Corollaries 1 and 2

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

C.6. Proof of Proposition 2

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

If the non-zero coordinates of the loading \mathbf{x}_{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).

D. Additional Simulation Results

In this session, we present additional simulation results for the decaying loading case, where \mathbf{x}_{new} is generated as,

$$x_{\text{new},j} = \text{Ratio} \cdot j^{-\delta},$$

where $\delta \in \{0, 0.1, 0.25, 0.5\}$ and $\text{Ratio} \in \{0.25, 0.375, 0.5\}$. The results are reported in Table 6 and the observation is consistent with that reported in Section 6.2 in the main paper.

δ	Ratio	n	ERR		Coverage		Len		HITS			Lasso			Deb		
			HITS	Deb	HITS	Deb	HITS	Deb	RM	Bias	SE	RM	Bias	SE	RM	Bias	SE
0	0.25	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
		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
	0.375	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
		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
	0.5	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
		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
	0.25	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
		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
0.1	0.375	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
		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
	0.5	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
		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
	0.25	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
		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
	0.375	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
		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
0.25	0.5	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
		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
	0.25	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
		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
	0.375	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
		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
	0.5	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
		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 $x_{\text{new},j} = \text{Ratio} * j^{-\delta}$. Reported also are the RM, 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”).