Inference for the Case Probability in High-dimensional Logistic Regression*

Zijian Guo Prabrisha Rakshit Daniel S. Herman Jinbo Chen

December 13, 2020

Abstract

Labeling patients in electronic health records with respect to their statuses of having a disease or condition, i.e. case or control statuses, has increasingly relied on prediction models using high-dimensional variables derived from structured and unstructured electronic health record data. A major hurdle currently is a lack of valid statistical inference methods for the case probability. In this paper, considering high-dimensional sparse logistic regression models for prediction, we propose a novel bias-corrected estimator for the case probability through the development of linearization and variance enhancement techniques. We establish asymptotic normality of the proposed estimator for any loading vector in high dimensions. We construct a confidence interval for the case probability and propose a hypothesis testing procedure for patient case-control labelling. We demonstrate the proposed method via extensive simulation studies and application to real-world electronic health record data.

Keywords: EHR phenotyping; Case-control; Outcome labelling; Re-weighting; Contraction principle.

1 Introduction

Electronic health record (EHR) data provides an unprecedented resource for clinical and translational research. Since EHRs were initially designed primarily to support documentation for medical billing, patients' data are frequently not represented with

*The research of Z. Guo was supported in part by NSF DMS 1811857, 2015373 and NIH R01GM140463-01, R56-HL-138306-01. The research of P. Rakshit was supported in part by NSF DMS 1811857. The research of D. Herman was supported in part by the University of Pennsylvania Department of Pathology and Laboratory Medicine and a Penn Center for Precision Medicine Accelerator Fund Award. The research of J. Chen was supported in part by NIH R56-HL138306, R01-HL138306 and R01GM140463-01. We would like to acknowledge Dr. Qiyang Han for the helpful discussion on contraction principles and Mr. Rong Ma for sharing the `WLDP` code; We would like to acknowledge the efforts of Xiruo Ding MS and Imran Ajmal MBBS, who were essential to the real data analysis presented. Mr. Ding extracted, wrangled, and engineered the EHR data. Dr. Ajmal performed the chart review for the three clinical phenotypes studied.

sufficient precision and nuance for accurate phenotyping. Therefore, heuristic rules and statistical methods are needed to identify patients with a specific health condition. Logistic regression models have been frequently adopted for this "EHR phenotyping" task [32, 1, 21, 15]. These methods commonly require a curated set of patients who are accurately labeled with regard to the presence or absence of a phenotype (e.g. disease or health condition). To obtain such a dataset, medical experts need to retrospectively review EHR charts and/or prospectively evaluate patients to label them. For many phenotypes, the labor and cost of the label assignment processes limit the achievable sample size, which is typically in the range of 50 to $1,000$. On the other hand, potential predictors in EHRs may include hundreds or thousands of variables derived from billing codes, demographics, disease histories, co-morbid conditions, laboratory test results, prescription codes, and concepts extracted from doctors' notes through methods such as natural language processing. The dimension of these predictors is usually large in comparison to the sample size of the curated dataset [11].

One important example phenotyping goal that would benefit from accurate risk prediction models leveraging large EHR data is primary aldosteronism (PA), the most common identifiable and specifically treatable cause of secondary high blood pressure [34, 27, 12]. PA is thought, based on epidemiological studies, to affect up to 1% of US adults [25, 20], but is diagnosed in many fewer individuals. Endocrine Society Guidelines recommend screening for PA in specific subgroups of hypertension patients, including patients with treatment-resistant high blood pressure or high blood pressure with low blood potassium [19]. While simple, expert-curated heuristics can be used to identify patients that meet PA screening guidelines, it is of great interest to derive more sensitive and specific prediction models by leveraging the larger set of available potential features in the EHR. One goal of the current paper is to use data extracted from the Penn Medicine EHR and develop preliminary prediction models to help identify patients with hypertension and subsets thereof for which PA screening is recommended by guidelines.

1.1 Problem Formulation

We introduce a general statistical problem, which is motivated by EHR phenotyping. We use $\{y_i, X_{i\cdot}\}_{1\le i\le n}$ to denote the available dataset. For the i-th observation, the outcome $y_i \in \{0,1\}$ indicates whether the interest condition (e.g. PA) is present and $X_{i\cdot} \in \mathbb{R}^p$ denotes the observed high-dimensional covariates. Here we assume that $\{y_i, X_{i\cdot}\}_{1\le i\le n}$ are independent and identically distributed and allow the number of covariates p to be larger than the sample size n as often seen in analyzing EHR data. We consider the following high-dimensional logistic regression model, for $1 \le i \le n$,

$$\mathbb{P}(y_i = 1|X_{i\cdot}) = h(X_{i\cdot}^\intercal \beta) \quad \text{with} \quad h(z) = \exp(z)/[1+\exp(z)] \tag{1}$$

where $\beta \in \mathbb{R}^p$ denotes the high-dimensional vector of odds ratio parameters. The high-dimensional vector β is assumed to be sparse througout the paper.

The quantity of interest is the case probability $\mathbb{P}(y_i = 1|X_{i\cdot} = x_*) \equiv h\left(x_*^\intercal \beta\right)$, which is the conditional probability of $y_i = 1$ given $X_{i\cdot} = x_* \in \mathbb{R}^p$. The outcome labeling problem

in EHR phenotyping is formulated as testing the following null hypothesis on the case probability,

$$H_0 : h(x_*^\intercal \beta) < 1/2. \tag{2}$$

Here, the threshold $1/2$ can be replaced by other positive numbers in $(0, 1)$, which are decided by domain scientists. Throughout the paper, we use the threshold $1/2$ to illustrate the main idea of EHR phenotyping.

Although the statistical inference problem is motivated from EHR phenotyping, the proposed inference procedure in the high-dimensional logistic model has a broader scope of applications. The linear contrast $x_*^\intercal \beta$ itself and the conditional probability of being a case are important quantities in statistics. Additionally, the case probability $h(X_{i\cdot}^\intercal \beta)$ is the same as the propensity score in causal inference, which is a central quantity for both matching [33, 35] and double robustness estimators [4, 24].

1.2 Our Results and Contribution

The penalized maximum likelihood estimation methods have been well developed to estimate $\beta \in \mathbb{R}^p$ in the high-dimensional logistic model [8, 3, 7, 29, 30, 22]. The penalized estimators enjoy desirable estimation accuracy properties. However, these methods do not lend themselves directly to statistical inference on the case probability mainly because the bias of the penalized estimator dominates the total uncertainty. Our proposed method is built upon the idea of bias correction that has been first developed for confidence interval construction for individual regression coefficients in high-dimensional linear regression models [39, 23, 40]. This idea has also been extended to making inference for β_j for $1 \le j \le p$ in high-dimensional logistic regression models [39, 31, 28]. However, there is a lack of methods and theories for inference for the case probability $\mathbb{P}(y_i = 1 | X_{i\cdot} = x_*)$, which depends on the high-dimensional loading vector $x_* \in \mathbb{R}^p$ and involves the entire regression vector $\beta \in \mathbb{R}^p$.

We propose a novel two-step bias-corrected estimator of the case probability. In the first step, we estimate β by a penalized maximum likelihood estimator $\widehat{\beta}$ and construct the plug-in estimator $h(x_*^\intercal \widehat{\beta}) = \exp(x_*^\intercal \widehat{\beta})/[1 + \exp(x_*^\intercal \widehat{\beta})]$. In the second step, we correct the bias of this plug-in estimator. The existing bias correction methods [39, 23, 40] compute the projection direction through estimating the high-dimensional vector $[\mathbf{E}\widehat{H}(\beta)]^{-1} x_* \in \mathbb{R}^p$ with $\widehat{H}(\beta)$ denoting the sample Hessian matrix of the negative log-likelihood (see Section 2.1 for its definition). However, it is challenging to extend this idea to inference for the case probability mainly due to the fact that the Hessian matrix $\mathbf{E}\widehat{H}(\beta)$ is complicated in the logistic model and $x_* \in \mathbb{R}^p$ can be an arbitrary high-dimensional vector (with no sparsity structure).

We address these challenges through development of linearization and variance enhancement techniques. The linearization technique is introduced to handle the complex form of the Hessian matrix in the logistic model. Particularly, instead of assigning equal weights, we conduct a weighted average with reweighing $X_{i\cdot}[y_i - h(x_*^\intercal \widehat{\beta})]$ by $1/\mathrm{Var}(y_i \mid X_i)$, which leads to a re-weighted Hessian matrix $n^{-1} \sum_{i=1}^n X_{i\cdot} X_{i\cdot}^\intercal$. We refer to this re-weighting step as "Linearization" since the re-weighted Hessian matrix corresponds

to the Hessian matrix of the least square loss in the linear model. In addition, to develop a inference procedure for any high-dimensional vector x_*, we introduce an extra constraint in constructing the projection direction for bias correction. The additional constraint is to enhance the variance component of the proposed bias-correct estimator such that its variance dominates its bias for any high-dimensional loading vector x_*. We refer to the proposed inference method as Linearization with Variance Enhancement, shorthanded as LiVE.

We have established the asymptotic normality of the proposed LiVE estimator for any high-dimensional loading vector $x_* \in \mathbb{R}^p$. We then construct a confidence interval for the case probability and conduct the hypothesis testing (2) related to the outcome labelling. We have developed new technical tools to establish the asymptotic normality for the re-weighted estimator after applying the linearization technique (See Section 3.3). This analysis has resolved a open question in [28], whether the sample splitting is needed to establish the asymptotic normality of a re-weighted estimator with data-dependent weights.

We have conducted a large set of simulation studies to compare the finite-sample performance of the proposed LiVE estimator with the existing state-of-the art methods: the plug-in Lasso estimator, post-selection method, the plug-in `hdi` [14] and the plug-in `WLDP` [28]. The proposed method outperforms these existing methods in terms of inference properties and computational efficiency. The proposed method is computationally efficient since the proposed method corrects biases in the entire vector all at once. The main computational cost is to fit two high-dimensional penalized regression models while a direct application of the coordinate-wise bias correction procedure [14, 28] requires implementation of $p + 2$ penalized regression problems. See Tables 1 and 2 for details.

We have demonstrated the proposed method using Penn Medicine EHR data to identify patients with hypertension and two subsets thereof that should be screened for PA, per specialty guidelines.

To sum up, the contribution of the current paper is two-folded.

1. We propose a novel bias-corrected estimator of the case probability and establish its asymptotic normality. To our best knowledge, this is the first inference method for the case probability in high dimensions, which is computationally efficient and statically valid for any high-dimensional vector x_*.

2. The theoretical justification on establishing the asymptotic normality of the re-weighted estimators is of independent interest and can be used to handle other inference problems in high-dimensional nonlinear models.

1.3 Existing Literature Comparison

We shall mention other related works and discuss the connections and difference. The estimation problem in the high-dimensional logistic regression has been investigated in the literature, including the ℓ_1 penalized logistic regression [8, 3, 7] and the group penalized regression [29]. The ℓ_1 penalized logistic regression can be taken as special

cases of the results established in [30] and [22], where [30] established general theories on M-estimator and [22] established general theories on penalized convex loss optimization.

Post-selection inference [5] is a commonly used method in constructing confidence intervals, where the first step is to conduct model selection and the second step is to run a low-dimensional logistic model with the selected sub-model. However, such a method typically requires the consistency of model selection in the first step. Otherwise, the constructed confidence intervals are not valid as the uncertainty of model selection in the first step is not properly accounted for. It has been observed in Section 4 that the post-selection method has produced under-covered confidence intervals in finite samples; see Tables 1 and 2 for a detailed comparison.

Inference for a linear combination of regression coefficients in high-dimensional linear model has been investigated in [9, 2, 41, 10]. However, the method cannot be directly applied to solve the same research problem in logistic model due to the more complicated form of Hessian matrix of the log-likelihood function. The linearization technique and also the developed empirical process results in the current paper are useful in adopting the inference methods developed for linear model to the logistic model. The connection established by the linearization is also useful for simplifying the sufficient conditions for estimating the precision matrix or the inverse Hessian matrix. Specifically, the established results in the current paper impose no sparsity conditions on the precision matrix or the inverse Hessian matrix, where such a requirement has typically been imposed in theoretical justifications on inference for individual regression coefficients in the logistic regression setting [39, 31, 28].

The re-weighting idea has been proposed in [28] for inference for the single regression coefficient β_j in the high-dimensional logistic model. However, the current paper targets at the case probability, which can be involved with all regression coefficients $\{\beta_j\}_{1\le j\le p}$. New methods and proof techniques are developed to address the inference problem for the case probability, especially for an arbitrary loading x_*. We have provided both theoretical and numerical comparisons in Sections 3.4 and 4, respectively.

The papers [6, 16, 13] studied inference for treatment effects in high-dimensional regression models while the current paper focuses on inference for a different quantity, the case probability. The papers [36, 37] studied inference in high-dimensional logistic regression and focused on the regime where the dimension p is a fraction of the sample size n. The current paper considered the regime allowing for the dimension p being much larger than the sample size n with imposing additional sparsity conditions on β.

Another related work is the iterated re-weighted least squares (IRLS) [17], which is the standard technique used to maximize the log-likelihood of the logistics modeling. The weighting is used in IRLS to facilitate the optimization problem. In contrast, the weighting used in the current paper is to facilitate the bias-correction for the statistical inference.

1.4 Notation

For a matrix $X \in \mathbb{R}^{n\times p}$, $X_{i\cdot}$, $X_{\cdot j}$ and $X_{i,j}$ denote respectively the i-th row, j-th column, (i,j) entry of the matrix X. $X_{i,-j}$ denotes the sub-row of $X_{i\cdot}$ excluding the j-th entry.

Let $[p] = \{1, 2, \cdots, p\}$. For a subset $J \subset [p]$, for a vector $x \in \mathbb{R}^p$, x_J is the subvector of x with indices in J and x_{-J} is the subvector with indices in J^c. For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{i=1}^q |x_i|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0$ denoting the cardinality of the support of x and $\|x\|_\infty = \max_{1\leq j\leq p} |x_j|$. We use e_i to denote the i-th standard basis vector in $\mathbb{R}^p$. We use $\max |X_{i,j}|$ as a shorthand for $\max_{1\leq i\leq n, 1\leq j\leq p} |X_{i,j}|$. For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalues of 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 $\limsup_{n\to\infty} a_n/b_n = 0$.

2 Methodology

We describe the proposed method for the case probability under the high-dimensional logistic model (1). In Section 2.1, we review the penalized maximum likelihood estimation of β and highlight the challenges of inference for the case probability. Then we introduce the linearization technique in Section 2.2 and the variance enhancement technique in Section 2.3. In Section 2.4, we construct a point estimator and a confidence interval of the case probability and conduct hypothesis testing related to outcome labelling.

2.1 Challenges of Inference for the Case Probability

The negative log-likelihood function for the data $\{(X_{i\cdot}, y_i)\}_{1\leq i\leq n}$ under the logistic regression model (1) is written as $\ell(\beta) = \sum_{i=1}^n \left[\log\left(1 + \exp\left(X_{i\cdot}^\intercal \beta\right)\right) - y_i \cdot \left(X_{i\cdot}^\intercal \beta\right)\right]$. The penalized log-likelihood estimator $\widehat{\beta}$ is defined as [7],

$$\widehat{\beta} = \arg\min_\beta \ell(\beta) + \lambda\|\beta\|_1, \tag{3}$$

with the tuning parameter $\lambda \asymp \sqrt{\log p/n}$. It has been shown that $\widehat{\beta}$ satisfies certain nice estimation accuracy and variable selection properties. However, the plug-in estimator $h(x_*^\intercal \widehat{\beta})$ cannot be directly used for confidence interval construction and hypothesis testing, because its bias can be as large as its variance as demonstrated in later simulation studies. (See Table 3 in the supplement for details.)

Our proposed method is built on the idea of correcting the bias of the plug-in estimator $x_*^\intercal \widehat{\beta}$ and then apply the h function to estimate the case probability. We conduct the bias correction through estimating the error of the plug-in estimator $x_*^\intercal \widehat{\beta} - x_*^\intercal \beta = x_*^\intercal(\widehat{\beta} - \beta)$. Before proposing the method, we review the existing bias-corrected idea in high-dimensional linear and logistic models [39, 23, 40]. In particular, a bias-corrected estimator of β_j can be constructed as

$$\widehat{\beta}_j + \widehat{u}^\intercal \frac{1}{n} \sum_{i=1}^n X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})) \tag{4}$$

where $\widehat{u} \in \mathbb{R}^p$ is the projection direction constructed for correcting the bias of $\widehat{\beta}_j$. Define the model error $\epsilon_i = y_i - h(X_{i\cdot}^\intercal \beta)$ for $1 \le i \le n$ and the prediction error

$$y_i - h(X_{i\cdot}^\intercal \widehat{\beta}) = h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))[X_{i\cdot}^\intercal(\beta - \widehat{\beta}) + \Delta_i] + \epsilon_i, \tag{5}$$

with the approximation error $\Delta_i = \int_0^1 (1-t) \frac{h''(X_{i\cdot}^\intercal \widehat{\beta} + t X_{i\cdot}^\intercal(\beta - \widehat{\beta}))}{h'(X_{i\cdot}^\intercal \widehat{\beta})} dt \cdot (X_{i\cdot}^\intercal(\widehat{\beta} - \beta))^2$. By multiplying both sides of (5) by X_i and summing over i, we obtain

$$\frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})) = \widehat{H}(\widehat{\beta})(\beta - \widehat{\beta}) + \frac{1}{n}\sum_{i=1}^n \epsilon_i X_{i\cdot} + \frac{1}{n}\sum_{i=1}^n h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))\Delta_i X_{i\cdot}, \tag{6}$$

where $\widehat{H}(\beta) = \frac{1}{n}\sum_{i=1}^n h(X_{i\cdot}^\intercal \beta)(1 - h(X_{i\cdot}^\intercal \beta)) X_{i\cdot} X_{i\cdot}^\intercal$ is the Hessian matrix of the log-likelihood $\ell(\beta)$. The bias-corrected estimator of β_j proposed in [39] essentially constructs the projection direction $\widehat{u} \in \mathbb{R}^p$ in (4) such that $\widehat{H}(\widehat{\beta})\widehat{u} \approx e_j$ and hence

$$\widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^n X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})) \approx \beta_j - \widehat{\beta}_j$$

Such an approximation has been shown to be accurate by assuming a sparse $[\mathbf{E}\widehat{H}(\beta)]^{-1} e_i$ [39]. However, for an arbitrary $x_* \in \mathbb{R}^p$, it is challenging to estimate $[\mathbf{E}\widehat{H}(\beta)]^{-1} x_*$ and generalize the bias-correction procedure in [39] for the following two reasons: (1) if $[\mathbf{E}\widehat{H}(\beta)]^{-1}$ does not have special structures (e.g. sparsity), the algorithm of directly inverting $\widehat{H}(\beta)$ can be unstable in high dimensions; (2) even imposing sparsity structures on $[\mathbf{E}\widehat{H}(\beta)]^{-1}$, it is challenging to estimate $[\mathbf{E}\widehat{H}(\beta)]^{-1} x_* \in \mathbb{R}^p$ accurately if x_* is a dense vector.

In the following two sections, we develop new techniques, which can effectively correct the bias for arbitrary loadings $x_* \in \mathbb{R}^p$ in the high-dimensional logistic regression.

2.2 Linearization: Connecting Logistic to Linear

We introduce a linearization technique to simplify the Hessian matrix. Instead of averaging with equal weights as in (6), we introduce the following re-weighted summation of (5),

$$\frac{1}{n}\sum_{i=1}^n \underbrace{[h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))]^{-1}}_{\text{weight for } i-th \text{ observation}} X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})).$$

In contrast to (6), the above re-weighted summation has the following decomposition:

$$\frac{1}{n}\sum_{i=1}^n [h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))]^{-1} \epsilon_i X_{i\cdot} + \widehat{\Sigma}(\beta - \widehat{\beta}) + \frac{1}{n}\sum_{i=1}^n \Delta_i X_{i\cdot}, \text{ with } \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_{i\cdot} X_{i\cdot}^\intercal.$$

The main advantage of the re-weighting step is that the second component $\widehat{\Sigma}(\beta - \widehat{\beta})$ on the right hand side is multiplication of the sample covariance matrix $\widehat{\Sigma}$ and the vector

difference $\widehat{\beta} - \beta$. In contrast to (6), it is sufficient to deal with $\widehat{\Sigma}$ to estimate the error $\widehat{\beta} - \beta$, instead of the more complicated Hessian matrix $\widehat{H}(\widehat{\beta})$. Since the main purpose of this re-weighting step is to match the re-weighted Hessian matrix as that of the least square loss in the linear models, we refer to this as the "Linearization" technique. We shall point out that, although linearization connects the logistic model to the linear model, it also poses great challenges of studying the theoretical properties of the proposed method. The corresponding technical challenge will be addressed in Section 3.3 by developing suitable empirical process techniques.

2.3 Variance Enhancement: Uniform Procedure for x_*

We apply the linearization technique and correct the bias of the plug-in estimator $x_*^\intercal \widehat{\beta}$ as,

$$\widehat{x_*^\intercal \beta} = x_*^\intercal \widehat{\beta} + \widehat{u}^\intercal \frac{1}{n} \sum_{i=1}^{n} [h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))]^{-1} X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})). \tag{7}$$

with $\widehat{u} \in \mathbb{R}^p$ denoting a projection direction to be constructed. To see how to construct $\widehat{u}$, we decompose the estimation error $\widehat{x_*^\intercal \beta} - x_*^\intercal \beta$ as

$$\frac{1}{n} \sum_{i=1}^{n} [h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))]^{-1} \epsilon_i \widehat{u}^\intercal X_{i\cdot} + (\widehat{\Sigma}\widehat{u} - x_*)^\intercal (\beta - \widehat{\beta}) + \frac{1}{n} \sum_{i=1}^{n} \Delta_i \widehat{u}^\intercal X_{i\cdot}, \tag{8}$$

where all three terms depend on $\widehat{u}$.

Motivated by the above decomposition, we construct $\widehat{u} \in \mathbb{R}^p$ as the solution of the following optimization problem,

$$\widehat{u} = \arg\min_{u \in \mathbf{R}^p} u^\intercal \widehat{\Sigma} u \quad \text{subject to } \|\widehat{\Sigma} u - x_*\|_\infty \le \|x_*\|_2 \lambda_n \tag{9}$$

$$|x_*^\intercal \widehat{\Sigma} u - \|x_*\|_2^2| \le \|x_*\|_2^2 \lambda_n \tag{10}$$

$$\|Xu\|_\infty \le \|x_*\|_2 \tau_n \tag{11}$$

where $\lambda_n \asymp (\log p/n)^{1/2}$ and $\tau_n \asymp (\log n)^{1/2}$. The details on implementing the above algorithm with tuning parameters selection are presented in Section 4.1.

We now provide some explanations on the above algorithm through connecting it to the error decomposition (8). The objective function scaled by $1/n$, $u^\intercal \widehat{\Sigma} u/n$, is of the same order of magnitude as the variance of the first term in the error decomposition (8). The constraints (9) and (11) are introduced to control the second and third terms in the error decomposition (8), respectively. Hence, the objective function, together with the constraints (9) and (11), ensures a projection direction $\widehat{u} \in \mathbb{R}^p$ such that the error $\widehat{x_*^\intercal \beta} - x_*^\intercal \beta$ is controlled to be small. Such an optimization idea has been proposed in the linear model [23, 40] and is shown to be effective when $x_* = e_j$ [23, 40], a sparse x_* [9] and a bounded x_* [2]. We shall emphasize that such an idea cannot be extended to general loadings x_* since the variance level of $\frac{1}{n}\sum_{i=1}^{n} [h(X_{i\cdot}^\intercal \widehat{\beta})(1 - h(X_{i\cdot}^\intercal \widehat{\beta}))]^{-1} \epsilon_i \widehat{u}^\intercal X_{i\cdot}$ is not guaranteed to dominate the other two bias terms in (8), without the additional constraint (10). See Proposition 2 of [10] for the examples.

To resolve this, we introduce the additional constraint (10) such that the variance component $\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i\widehat{u}^{\intercal}X_{i\cdot}$ is the dominating term in the error decomposition (8), for any high-dimensional vector $x_* \in \mathbb{R}^p$. In particular, this constraint enhances the variance component in the error decomposition (8) and hence we refer to the above construction of projection direction $\widehat{u}$ in (9) to (11) as "variance enhancement".

Remark 1. *We have shown in Theorem 1 that, with a high probability, $u^* = \Sigma^{-1}x_*$ belongs to the feasible set defined by* (9), (10) *and* (11). *However, we shall emphasize that, although $\widehat{u}$ defined by the optimization problem* (9) *to* (11) *is targeting at $u^* = \Sigma^{-1}x_*$, the asymptotic normality of the proposed LiVE estimator defined in* (7) *does not require that $\widehat{u}$ is an accurate estimator of u^*. This explains why the proposed bias-corrected estimator is applied to a broad setting since our construction does not require any sparsity condition on Σ^{-1}, x_* or $\Sigma^{-1}x_*$. See Theorem 1 and its proof for details.*

Remark 2. *In the high-dimensional linear model, the variance enhancement idea has been proposed in constructing the bias corrected estimator for $x_*^{\intercal}\beta$ [10]. However, the method developed for linear models in [10] cannot be directly applied to the inference problem for the case probabilities due to the complexity of the Hession matrix, as highlighted in Section 2.2. A valid inference procedure for the case probability depends on both Linearization and Variance Enhancement techniques.*

Remark 3. *The idea of adding the constraint $\|Xu\|_\infty \le \|x_*\|_2\tau_n$ was introduced in [23] to handle the non-Gaussian error in the linear model. In our analysis, this additional constraint is not just introduced to deal with the non-Gaussian error ϵ_i, but also facilitates the empirical process proof. The upper bound τ_n is also different, where equation (54) of [23] has $\|x_*\|_2 = 1$ and $\tau_n \asymp n^{\delta_0}$ with $1/4 < \delta_0 < 1/2$ while τ_n here is required to satisfy $(\log n)^{1/2} \lesssim \tau_n \ll n^{1/2}$. We have set $\tau_n \asymp (\log n)^{1/2}$ throughout the current paper.*

2.4 LiVE: Inference for Case Probabilities

We propose to estimate $x_*^{\intercal}\beta$ by $\widehat{x_*^{\intercal}\beta}$ as defined in (7), with the initial estimator $\widehat{\beta}$ defined in (3) and the projection direction $\widehat{u}$ constructed in (9) to (11).

Subsequently, we estimate the case probability $\mathbb{P}(y_i = 1|X_{i\cdot} = x_*)$ by

$$\widehat{\mathbb{P}}(y_i = 1|X_{i\cdot} = x_*) = h(\widehat{x_*^{\intercal}\beta}) \tag{12}$$

From the above construction, the asymptotic variance of $\widehat{x_*^{\intercal}\beta}$ can be estimated by

$$\widehat{\mathrm{V}} = \widehat{u}^{\intercal}\left[\frac{1}{n^2}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}X_{i\cdot}^{\intercal}\right]\widehat{u}.$$

We construct the confidence interval for the case probability $\mathbb{P}(y_i = 1|X_{i\cdot} = x_*)$ as follows:

$$\mathrm{CI}_\alpha(x_*) = \left[h\left(\widehat{x_*^{\intercal}\beta} - z_{\alpha/2}\widehat{\mathrm{V}}^{1/2}\right), h\left(\widehat{x_*^{\intercal}\beta} + z_{\alpha/2}\widehat{\mathrm{V}}^{1/2}\right)\right], \tag{13}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution. We can conduct the following hypothesis testing related to outcome labeling (2)

$$\phi_\alpha(x_*) = \mathbf{1}\left(\widehat{x_*^\intercal \beta} - z_\alpha \widehat{\mathrm{V}}^{1/2} \geq 0\right). \tag{14}$$

Here, the testing procedure (14) will label the observation as a case if $\widehat{x_*^\intercal \beta}$ is above $z_\alpha \widehat{\mathrm{V}}^{1/2}$; as a control, otherwise. If the goal is to test the null hypothesis $H_0 : h(x_*^\intercal \beta) < c_*$ for $c_* \in (0,1)$, we generalize (14) to $\phi_\alpha^{c_*}(x_*) = \mathbf{1}\left(\widehat{x_*^\intercal \beta} - z_\alpha \widehat{\mathrm{V}}^{1/2} \geq h^{-1}(c_*)\right)$, where h^{-1} is the inverse function of h defined in (1).

3 Theoretical Justification

We provide theoretical justification for the proposed method. In Section 3.1, we present the model conditions and the theoretical properties of the initial estimator $\widehat{\beta}$. In Section 3.2, we establish asymptotic normality of the proposed LiVE estimator and then provide theoretical justification for confidence interval construction and hypothesis testing. In Section 3.3, we present the technical tools of handling the theoretical challenge of linearization.

3.1 Model Conditions and Initial Estimators

We introduce the following modeling assumptions to facilitate theoretical analysis.

(A1) The rows $\{X_{i\cdot}\}_{1 \leq i \leq n}$ are i.i.d. p-dimensional Sub-gaussian random vectors with $\Sigma = \mathbf{E}(X_{i\cdot} X_{i\cdot}^\intercal)$ where Σ satisfies $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$ for some positive constants $C_0 \geq c_0 > 0$; The high-dimensional vector β is assumed to be of sparsity k.

(A2) With probability larger than $1 - p^{-c}$,

$$\min\left\{\frac{\exp(X_{i\cdot}^\intercal \beta)}{1+\exp(X_{i\cdot}^\intercal \beta)}, \frac{1}{1+\exp(X_{i\cdot}^\intercal \beta)}\right\} \geq c_{\min},$$

for $1 \leq i \leq n$ and some small positive constant $c_{\min} \in (0,1)$.

The assumption (A1) imposes the tail condition for the high-dimensional covariates $X_{i\cdot}$ and assumes that the population second order moment matrix is invertible. Assumption (A2) is imposed such that the case probability is uniformly bounded away from 0 and 1 by a small positive constant $c_{\min}$. Condition (A2) requires $X_{i\cdot}^\intercal \beta$ to be bounded for all $1 \leq i \leq n$ with a high probability. Such a condition has been commonly made in in analyzing high-dimensional logistic models [2, 39, 28, 31]. For example, see condition (iv) of Theorem 3.3 of [39] and the overlap assumption (Assumption 6) in [2].

The following proposition establishes the theoretical properties of the penalized maximum likelihood estimator $\widehat{\beta}$ in (3) under model conditions (A1) and (A2).

Proposition 1. *Suppose that Conditions* (A1) *and* (A2) *hold and* $\max_{i,j}|X_{ij}|\, k\lambda_0 \le c$ *with* $\lambda_0 = \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|_\infty$ *and a positive constant* c. *For any positive constant* $\delta_0 > 0$ *and the proposed estimator* $\widehat{\beta}$ *in* (3) *with* $\lambda = (1+\delta_0)\lambda_0$, *then with probability greater than* $1 - p^{-c} - \exp(-cn)$,

$$\|\widehat{\beta}_{S^c} - \beta_{S^c}\|_1 \le (2/\delta_0 + 1)\|\widehat{\beta}_S - \beta_S\|_1 \quad \textit{and} \quad \|\widehat{\beta} - \beta\|_1 \le Ck\lambda_0 \tag{15}$$

where S *denotes the support of* β *and* $C > 0$ *is a positive constant.*

We will choose λ_0 at the scale of $(\log p/n)^{1/2}$ and then Proposition 1 shows that the initial estimator $\widehat{\beta}$ satisfies the following property:

(B) With probability greater than $1 - p^{-c} - \exp(-cn)$ for some constant $c > 0$, the initial estimator $\widehat{\beta}$ satisfies

$$\|\widehat{\beta} - \beta\|_1 \le Ck\,(\log p/n)^{1/2} \quad \text{and} \quad \|\widehat{\beta}_{S^c} - \beta_{S^c}\|_1 \le C_0\|\widehat{\beta}_S - \beta_S\|_1$$

where S denotes the support of β and $C > 0$ and $C_0 > 0$ are positive constants.

As a remark, the asymptotic normality established in next subsection will hold for any initial estimator satisfying condition (B), including the initial estimator (3) used in our algorithm.

3.2 Asymptotic Normality and Statistical Inference

We now establish the limiting distribution for the proposed point estimator $\widehat{x_*^\intercal\beta}$. The limiting distribution for $\widehat{x_*^\intercal\beta}$ will naturally lead to a limiting distribution for $h(\widehat{x_*^\intercal\beta})$.

Theorem 1. *Suppose that Conditions* (A1) *and* (A2) *hold,* $\tau_n \asymp (\log n)^{1/2}$ *defined in* (11) *satisfies* $\tau_n k \log p/\sqrt{n} \to 0$. *Then for any initial estimator* $\widehat{\beta}$ *satisfying condition* (B) *and any constant* $0 < \alpha < 1$,

$$\mathbb{P}\left[\mathrm{V}^{-1/2}\left(\widehat{x_*^\intercal\beta} - x_*^\intercal\beta\right) \ge z_\alpha\right] \to \alpha \tag{16}$$

where

$$\mathrm{V} = \widehat{u}^\intercal\left[\frac{1}{n^2}\sum_{i=1}^n [h(X_{i\cdot}^\intercal\beta)(1 - h(X_{i\cdot}^\intercal\beta))]^{-1} X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u}. \tag{17}$$

With probability greater than $1 - p^{-c} - \exp(-cn)$,

$$c_0\|x_*\|_2/n^{1/2} \le \mathrm{V}^{1/2} \le C_0\|x_*\|_2/n^{1/2}, \tag{18}$$

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

By the above theorem, we can use the delta method to obtain the following limiting distribution of $h(\widehat{x_*^\intercal\beta})$, $\mathbb{P}\left[\left(\rho^2\mathrm{V}\right)^{-1/2}\left(h(\widehat{x_*^\intercal\beta}) - \mathbb{P}(y_i = 1|X_{i\cdot} = x_*)\right) \ge z_\alpha\right] \to \alpha$, where $\rho = h(x_*^\intercal\beta)(1 - h(x_*^\intercal\beta))$.

The established limiting distribution in Theorem 1 can be used to justify the validity of the proposed confidence interval.

Proposition 2. *Under the same conditions as in Theorem 1, the confidence interval* $\mathrm{CI}_\alpha(x_*)$ *proposed in* (13) *satisfies*

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathbb{P}(y_i = 1|X_{i\cdot} = x_*) \in \mathrm{CI}_\alpha(x_*)\right] \geq 1 - \alpha.$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathbf{L}(\mathrm{CI}_\alpha(x_*)) \geq (1+\delta)\left(\rho^2 \mathrm{V}\right)^{1/2}\right) = 0$$

where $\mathbf{L}(\mathrm{CI}_\alpha(x_*))$ *denotes the length of the confidence interval* $\mathrm{CI}_\alpha(x_*)$, $\delta > 0$ *is any positive constant,* V *is defined in* (17) *and* $\rho = h(x_*^\intercal\beta)(1 - h(x_*^\intercal\beta))$.

A few remarks are in order for Theorem 1 and Proposition 2. Firstly, the asymptotic normality in Theorem 1 is established without imposing any condition on the high-dimensional vector $x_* \in \mathbb{R}^p$. The variance enhancement construction of the projection direction $\widehat{u}$ in (9) to (11) is crucial to establishing such a uniform result over any $x_* \in \mathbb{R}^p$. Specifically, with the additional constraint (10), we can establish the lower bound of the asymptotic variance in (18), which guarantees the variance component of (8) to dominate the remaining bias.

Secondly, to establish the asymptotic normality result, we do not imposes any sparsity condition on the precision matrix Σ^{-1}. This has weakened the sparsity assumptions on Σ^{-1} in the literature; see Section 3.4 for details. Thirdly, the sparsity condition on β is nearly the weakest condition, which is necessary to construct adaptive confidence interval. With $\tau_n \asymp (\log n)^{1/2}$, a sufficient sparsity condition on β is $k \ll n^{1/2}/[\log p\,(\log n)^{1/2}]$. The optimality results in [9] have shown that the ultra-sparse condition $k \ll n^{1/2}/\log p$ is necessary and sufficient for constructing adaptive confidence intervals for β_j in the linear setting with unknown Σ^{-1}. The required sparsity condition $k \ll n^{1/2}/[\log p\,(\log n)^{1/2}]$ for the case probability inference in the non-linear logistic model almost achieves the weakest sparsity condition for constructing adaptive confidence intervals in the linear models.

Theorem 1 also justifies the validity of the proposed testing procedure. To study the testing procedure, we introduce the following parameter space for $\theta = (\beta, \Sigma)$,

$$\Theta(k) = \{\theta = (\beta, \Sigma) : \|\beta\|_0 \leq k, c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0\}$$

for some positive constants $C_0 \geq c_0 > 0$. We consider the following null parameter space $\mathcal{H}_0 = \{\theta = (\beta, \Sigma) \in \Theta(k) : x_*^\intercal\beta \leq 0\}$ and the local alternative parameter space

$$\mathcal{H}_1(\mu) = \left\{\theta = (\beta, \Sigma) \in \Theta(k) : x_*^\intercal\beta = \mu/n^{1/2}\right\}, \quad \text{for some } \mu > 0.$$

Proposition 3. *Under the same conditions as in Theorem 1, for each* $\theta \in \mathcal{H}_0$, *the proposed testing procedure* $\phi_\alpha(x_*)$ *in* (14) *satisfies* $\limsup_{n\to\infty} \mathbb{P}_\theta\left[\phi_\alpha(x_*) = 1\right] \leq \alpha$. *For* $\theta \in \mathcal{H}_1(\mu)$, *we have*

$$\limsup_{n\to\infty} \left|\mathbb{P}_\theta\left[\phi_\alpha(x_*) = 1\right] - [1 - \Phi^{-1}(z_\alpha - \mu/(n\mathrm{V})^{1/2})]\right| = 0, \tag{19}$$

where Φ^{-1} *is the inverse of the cumulative function of standard normal distribution.*

The proposed hypothesis testing procedure is shown to have a well-controlled type I error rate. Regarding the power for testing against $H_1(\mu)$, we note that (18) implies $c_0\|x_*\|_2 \leq (n\mathrm{V})^{1/2} \leq C_0\|x_*\|_2$ and hence the power of the proposed test in (19) is nontrivial if $\mu \geq C\|x_*\|_2$ holds for a large positive constant C. If $\mu/\|x_*\|_2 \to \infty$ or equivalently $n^{1/2}x_*^\intercal\beta/\|x_*\|_2 \to \infty$, then the asymptotic power will be 1 in the asymptotic sense. It has also been observed in Section 4 that the finite sample performance of the proposed procedure depends on the sample size n and the ℓ_2 norm $\|x_*\|_2$.

3.3 Analysis Related to Reweighting in Linearization

In the following, we provide more insights on how to establish the asymptotic normality and summarize new technical tools for analyzing the re-weighted estimator in the linearization procedure. Regarding the decomposition (8), the first term captures the stochastic error due to the model error ϵ_i, the second term is a bias component arising from estimating $\Sigma^{-1}x_*$, and the third term appears due to the nonlinearity of the logistic regression model. The following proposition controls the second and third terms.

Proposition 4. *Suppose that Conditions* (A1) *and* (A2) *hold. For any estimator* $\widehat{\beta}$ *satisfying Condition* (B)*, then with probability larger than* $1 - p^{-c} - \exp(-cn)$*,*

$$n^{1/2}\left|(\widehat{\Sigma}\widehat{u} - x_*)^\intercal(\widehat{\beta} - \beta)\right| \leq n^{1/2}\|x_*\|_2\lambda_n\|\widehat{\beta} - \beta\|_1 \lesssim \|x_*\|_2 k\log p \cdot n^{-1/2}, \tag{20}$$

and

$$n^{1/2}|\widehat{u}^\intercal\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}\Delta_i| \leq \tau_n\|x_*\|_2 k\log p \cdot n^{-1/2} \tag{21}$$

With the above proposition and the decomposition (8), the main next step is to establish the asymptotic normality of re-weighted summation of the model errors,

$$\widehat{u}^\intercal\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^\intercal\widehat{\beta})(1 - h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1}X_{i\cdot}\epsilon_i. \tag{22}$$

Because of the dependence between the weight $[h(X_{i\cdot}^\intercal\widehat{\beta})(1 - h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1}$ and the model error ϵ_i, it is challenging to establish the asymptotic normality of this re-weighted summation (22).

Remark 4. *In comparison, in the linear model case or the logistic model without re-weighting [39, 23, 40], such a challenge does not exist since the corresponding stochastic error term is of the form* $\widehat{u}^\intercal\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\epsilon_i$ *and the direction* $\widehat{u}$*, defined in* **(author?)** *[39, 23, 40], is either directly independent of* ϵ_i *or can be replaced by* $u^* = \Sigma^{-1}x_*$ *(by assuming* $\Sigma^{-1}x_*$ *to be sparse). The similar techniques cannot be applied to establish the asymptotic normality of the term* (22) *in the re-weighted estimator.*

The dependence between the weights and the model error in (22) requires a careful analysis to establish the asymptotic normality. We decouple the correlation between $\widehat{\beta}$ and ϵ_i through the following expression,

$$\begin{aligned}\widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^n [h(X_{i\cdot}^\intercal\widehat{\beta})(1-h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1}X_{i\cdot}\epsilon_i = \widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^n [h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}X_{i\cdot}\epsilon_i \\ + \widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^n \Big([h(X_{i\cdot}^\intercal\widehat{\beta})(1-h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1} - [h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}\Big) X_{i\cdot}\epsilon_i.\end{aligned} \tag{23}$$

The first component on the right hand side of the above summation is not involved with the estimator $\widehat{\beta}$, so that the standard probability argument can be applied to establish the asymptotic normality. The second component on the right hand side of (23) captures the error of estimating β by $\widehat{\beta}$. We will provide a sharp control of this error term through the development of suitable empirical process theory. The following lemma addresses the approximation error in the decoupling decomposition (23).

Lemma 1. *Suppose that Conditions* (A1) *and* (A2) *hold and the initial estimator* $\widehat{\beta}$ *satisfies Condition* (B)*, then with probability greater than* $1-p^{-c}-\exp(-cn)-1/t_0$*,*

$$\left|\widehat{u}^\intercal \frac{1}{n^{1/2}}\sum_{i=1}^n \Big([h(X_{i\cdot}^\intercal\widehat{\beta})(1-h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1} - [h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}\Big) X_{i\cdot}\epsilon_i\right| \le Ct_0\tau_n\|x_*\|_2\frac{k\log p}{n^{1/2}} \tag{24}$$

where τ_n *is defined in* (11)*,* $t_0>1$ *is a large positive constant and* $c>0$ *and* $C>0$ *are positive constants.*

The main step of establishing the above lemma is to apply a contraction principle for i.i.d. symmetric random variables taking values $\{-1,1,0\}$. See Lemma 7 for the precise statement. This extends the existing results of contraction principles on i.i.d Rademacher random variables [26]. This lemma and the related analysis are particularly useful for carefully characterizing the approximation error in (24) and can be of independent interest in establishing asymptotic normality of other re-weighted estimators in high dimensions. The proof of Lemma 1 is presented in Section 6.

We remark that the bound in Lemma 1 is more refined than applying standard inequalities. In particular, we can apply Hölder inequality to upper bound the left hand side of (24) by $\frac{1}{n^{1/2}}\sum_{i=1}^n |\widehat{u}^\intercal X_{i\cdot}\epsilon_i| \max_{1\le i\le p}\left|\Big([h(X_{i\cdot}^\intercal\widehat{\beta})(1-h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1} - [h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}\Big)\right|$. By the Taylor expansion and assuming $[h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}$ to have a finite first order derivative, we can further upper bound the left hand side of (24) by

$$\frac{1}{n^{1/2}}\sum_{i=1}^n |\widehat{u}^\intercal X_{i\cdot}\epsilon_i| \max_{1\le i\le p}|X_{i\cdot}^\intercal(\widehat{\beta}-\beta)| \lesssim n^{1/2}\max_{1\le i\le p}|X_{i\cdot}^\intercal(\widehat{\beta}-\beta)| \lesssim k\log p, \tag{25}$$

where the first inequality follows from $\frac{1}{n}\sum_{i=1}^n |\widehat{u}^\intercal X_{i\cdot}\epsilon_i| \to \mathbf{E}\left(|\widehat{u}^\intercal X_{i\cdot}\epsilon_i|\right)$ and the second inequality follows from the error bound for $\|\widehat{\beta}-\beta\|_1$. By comparing (25) and (24), we have seen the analysis in (24) leads to the additional convergence factor $1/\sqrt{n}$.

3.4 Comparison to Existing Estimators

In the following, we make a few remarks on comparing the proposed LiVE method with the existing inference results for high-dimensional sparse logistic regression models. A detailed numerical comparison with the state-of-the-art methods [39, 28] are provided in Section 4. The main distinction is that the existing literature focused on single regression coefficients, instead of the case probability. Since the proposed method is directly targeting at the case probability, it enjoys both theoretical and numerical advantage in comparison to the existing inference methods.

Firstly, there exists technical difficulties to establish the asymptotic normality of the plug-in estimators $x_*^\intercal\widetilde{\beta}$ with $\widetilde{\beta} \in \mathbb{R}^p$ denoting any coordinate-wise bias-corrected estimator proposed in [39, 31, 28]. To see this, we can apply the results in [39, 31, 28] to show that for $1 \le j \le p$, $\widetilde{\beta}_j = \beta_j + \mathrm{Var}(\widetilde{\beta}_j) + \mathrm{Bias}(\widetilde{\beta}_j)$ where $\sqrt{n}\mathrm{Var}(\widetilde{\beta}_j)$ is asymptotically normal and $\mathrm{Bias}(\widetilde{\beta}_j)$ is a small bias component. Then we have, for the error decomposition $x_*^\intercal\widetilde{\beta} - x_*^\intercal\beta = \sum_{j=1}^p x_{*,j}\mathrm{Var}(\widetilde{\beta}_j) + \sum_{j=1}^p x_{*,j}\mathrm{Bias}(\widetilde{\beta}_j)$, the component $\sqrt{n}\sum_{j=1}^p x_{*,j}\mathrm{Var}(\widetilde{\beta}_j)$ is asymptotically normal with a standard error of order $\|x_*\|_2$ and the bias $\sum_{j=1}^p x_{*,j}\mathrm{Bias}(\widetilde{\beta}_j)$ is upper bounded by $\|x_*\|_1 k \log p/n$, with a high probability. If $\|x_*\|_1$ is much larger than $\|x_*\|_2$, the upper bound for the bias $\sum_{j=1}^p x_{*,j}\mathrm{Bias}(\widetilde{\beta}_j)$ is not necessarily dominated by the standard error of $\sum_{j=1}^p x_{*,j}\mathrm{Var}(\widetilde{\beta}_j)$, even if $k \ll \sqrt{n}/\log p$. We shall point out that, the upper bound for the bias depends on $\|x_*\|_1$ instead of $\|x_*\|_2$ mainly because the coordinate-wise inference results constrained the bias $\mathrm{Bias}(\widetilde{\beta}_j)$ separately instead of directly constraining $\sum_{j=1}^p x_{*,j}\mathrm{Bias}(\widetilde{\beta}_j)$ as a total. This makes it challenging to establish asymptotic normality of the plug-in estimators for any high-dimensional loading x_*. Due to the same reason, the computation cost of the plug-in estimator $x_*^\intercal\widetilde{\beta}$ can be much higher than the proposed method, as the proposed method targets at $x_*^\intercal\beta$ directly and requires construction of one projection direction. In contrast, the plug-in debiased estimator $x_*^\intercal\widetilde{\beta}$ requires construction of p projection directions. See Tables 1 and 2 for details.

Secondly, the proposed re-weighting in the linearization step is helpful in removing sparsity conditions imposed on the inverse of the Hessian matrix $\mathbf{E}\left(h(X_{i\cdot}^\intercal\beta)\left(1-h(X_{i\cdot}^\intercal\beta)\right)X_{i\cdot}X_{i\cdot}^\intercal\right)$ or the precision matrix Σ^{-1} [39, 31, 28]. Thirdly, a related re-weighting idea has been proposed in [28] for the inference on individual regression coefficients. It is not obvious how to extend the specific estimator in [28] to handle the inference for arbitrary linear combination of all regression coefficients. Additionally, the theoretical justification in [28] required sample splitting for establishing the asymptotic normality, where half of the data was used for constructing an initial estimator of the regression coefficient vector and the other half was used for bias correction. It has been left as an open question in [28] whether sample splitting can be avoided for general re-weighting estimators. Our newly developed empirical process results in Lemma 1 are key to establishing the asymptotic normality without sample splitting. As a consequence, the proposed estimator makes use of the full sample in both initial estimation and bias-correction steps and hence retains the efficiency.

4 Numerical Studies

Throughout the simulation, we range the sample size n across $\{200, 400, 600\}$ and fix $p = 501$, including the intercept and 500 predictors. We use $X_{i,1} = 1$ to represent the intercept and generate the covariates $\{X_{i,-1}\}_{1\le i\le n}$ from the multivariate normal distribution with zero mean and covariance matrix $\Sigma = \{0.5^{1+|j-l|}\}_{1\le j,l\le 500}$. We generate both exact sparse and approximate sparse regression vector β and present them in Sections 4.2 and 4.3, respectively. Conditioning on the high-dimension covariates $X_{i\cdot}$, the binary outcome is generated by $y_i \sim \text{Bernoulli}\left(h\left(X_{i\cdot}^\intercal \beta\right)\right)$, for $1 \le i \le n$. All simulation results are averaged over 500 replications. To illustrate that the proposed method works for arbitrary x_*, we generate the following loadings x_*.

- **Loading 1**: We set $x_{\text{basis},1} = 1$ and generate $x_{\text{basis},-1} \in \mathbb{R}^{500}$ following $N(0, \Sigma)$ with $\Sigma = \{0.5^{1+|j-l|}\}_{1\le j,l\le 500}$ and generate x_* as

$$x_{*,j} = \begin{cases} x_{\text{basis},j} & \text{for } 1 \le j \le 11 \\ r \cdot x_{\text{basis},j} & \text{for } 12 \le j \le 501 \end{cases} \tag{26}$$

  where the ratio r varies across $\{1, 1/25\}$. For $r = 1$, x_* is the same as x_{basis} while for $r = 1/25$, x_* is a shrunk version of x_{basis}.

- **Loading 2**: $x_{\text{basis},1}$ is set as 1 and $x_{\text{basis},-1} \in \mathbb{R}^{500}$ is generated as following $N(0, \Sigma)$ with $\Sigma = \{(-0.75)^{1+|j-l|}\}_{1\le j,l\le 500}$ and x_* is generated using (26) with $r = 1$ or $r = 1/25$.

The loadings are only generated once and kept the same across all 500 replications.

4.1 Algorithm Implementation

We provide details on how to implement the LiVE estimator defined in (7). The initial estimator $\widehat{\beta}$ defined in (3) is computed using the R-package `cv.glmnet` [18] with the tuning parameter λ chosen by cross-validation. To compute the projection direction $\widehat{u} \in \mathbb{R}^p$, we implement the following constrained optimization,

$$\begin{aligned} \widehat{u} = \ \arg\min_{u\in\mathbf{R}^p} u^\intercal \widehat{\Sigma} u \quad \text{subject to } & \|\widehat{\Sigma} u - x_*\|_\infty \le \|x_*\|_2 \lambda_n \\ & |x_*^\intercal \widehat{\Sigma} u - \|x_*\|_2^2| \le \|x_*\|_2^2 \lambda_n \end{aligned} \tag{27}$$

In comparison to the optimization problem in (9) to (11), the numerical implementation does not include constraint (11) as such a condition is mainly imposed to facilitating the technical proof. We have observed that construction of $\widehat{u}$ in (27) is effective in correcting the bias and constructing valid confidence intervals across different settings. We solve the dual problem of (27),

$$\widehat{v} = \arg\min_{v\in\mathbb{R}^{p+1}} \frac{1}{4} v^\intercal H^\intercal \widehat{\Sigma} H v + b^\intercal H v + \mu\|v\|_1 \text{ with } H = [b, \mathbb{I}_{p\times p}], b = \frac{1}{\|x_*\|_2} x_*$$

and then solve the primal problem (27) as $\widehat{u} = -\left(\widehat{v}_{-1} + \widehat{v}_1 b\right)/2$. In this dual problem, when $\widehat{\Sigma}$ is singular and the tuning parameter $\mu > 0$ gets sufficiently close to 0, the dual problem cannot be solved as the minimum value converges to negative infinity. Hence, we choose the smallest $\mu > 0$ such that the dual problem has a finite minimum value. The exact code for implementing the algorithm is available at `https://github.com/zijguo/Logistic-Model-Inference`.

We compare the proposed LiVE estimator with the following state-of-the-art methods.

- **Plug-in Lasso.** We estimate $x_*^\intercal \beta$ by $x_*^\intercal \widehat{\beta}$ with $\widehat{\beta}$ denoting the penalized logistic regression estimator in (3).
- **Post-selection method.** First select important predictors through penalized logistic regression estimator $\widehat{\beta}$ in (3) and then fit a standard logistic regression with the selected predictors. The post-selection estimator $\widehat{\beta}_{\mathrm{PL}} \in \mathbf{R}^p$ is used to estimate $x_*^\intercal \beta$ by $x_*^\intercal \widehat{\beta}_{\mathrm{PL}}$. The variance of this post-selection estimator $x_*^\intercal \widehat{\beta}_{\mathrm{PL}}$ can be obtained by the inference results in the classical low-dimensional logistic regression, denoted by $\widehat{\mathrm{V}}_{\mathrm{PL}}$.
- **Plug-in `hdi` [14].** The R package `hdi` is implemented to obtain the coordinate debiased Lasso estimator $\widehat{\beta}_{\mathtt{hdi}} \in \mathbf{R}^p$ and the plug-in estimator $x_*^\intercal \widehat{\beta}_{\mathtt{hdi}}$ is used to estimate $x_*^\intercal \beta$, with the variance estimator as $\widehat{\mathrm{V}}_{\mathtt{hdi}}$.
- **Plug-in `WLDP` [28].** We compute the debiased lasso estimator $\widehat{\beta}_{\mathtt{WLDP}} \in \mathbf{R}^p$ by the Weighted LDP algorithm in Table 1 of [28]. The plug-in estimator of $x_*^\intercal \beta$ and the associated variance are given by $x_*^T \widehat{\beta}_{\mathtt{WLDP}}$ and $\widehat{\mathrm{V}}_{\mathtt{WLDP}}$ respectively.

We compare the above estimators with the proposed LiVE estimator in (7) in terms of Root Mean Square Error (RMSE), standard error and bias. Since the plug-in Lasso estimator is not useful for CI construction due to its large bias, we compare with Post-selection method, plug-in `hdi` and `WLDP`, from the perspectives of CI construction and hypothesis testing (2). Recall that the proposed CI and testing procedure for (2) are implemented as in (13) and (14), respectively. The inference procedures based on post-selection method, plug-in `hdi` and plug-in weighted `WLDP` are defined as,

$$\mathrm{CI}_\alpha(x_*) = \left[h(x_*^\intercal \widetilde{\beta} - z_{\alpha/2}\widetilde{\mathrm{V}}^{1/2}), h(x_*^\intercal \widetilde{\beta} + z_{\alpha/2}\widetilde{\mathrm{V}}^{1/2})\right],\ \phi_\alpha(x_*) = \mathbf{1}\left(x_*^\intercal \widetilde{\beta} - z_\alpha \widetilde{\mathrm{V}}^{1/2} \geq 0\right).$$

with replacing $(\widetilde{\beta}, \widetilde{\mathrm{V}})$ with $\left(\widehat{\beta}_{\mathrm{PL}}, \widehat{\mathrm{V}}_{\mathrm{PL}}\right)$, $\left(\widehat{\beta}_{\mathtt{hdi}}, \widehat{\mathrm{V}}_{\mathtt{hdi}}\right)$ and $\left(\widehat{\beta}_{\mathtt{WLDP}}, \widehat{\mathrm{V}}_{\mathtt{WLDP}}\right)$, respectively.

4.2 Exact Sparse β

We generate the sparse regression vector β as $\beta_1 = 0$, $\beta_j = (j-1)/20$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq 501$. Since the R package `hdi` and the `WLDP` algorithm only report the debiased estimators together with their variance estimators for the regression coefficients excluding the intercept, the intercept β_1 is set as 0 to have a fair comparison. For this β, the case probabilities $\mathbb{P}(y_i = 1 \mid X_{i\cdot} = x_*) = h(x_*^\intercal \beta)$ for Loading 1 and Loading 2 turn out to be 0.732 and 0.293, respectively. Here, the scale parameter r in (26) controls the magnitude of the noise variables in x_*. As r decreases, $\|x_*\|_2$ decreases but the case

probability $h(x_*^\intercal\beta)$ remains the same for all choices of r since only the values of $x_{*,j}$ for $1 \leq j \leq 11$ affect $x_*^\intercal\beta$.

In Table 1, we compare the proposed LiVE method with post-selection, `hdi` and `WLDP`, in terms of CI construction and hypothesis testing. The coverage and lengths of CIs are reported under the columns indexed with "Cov" and "Len", respectively. The CIs constructed by LiVE and `hdi` have coverage over different scenarios and the lengths are reduced when a larger sample is used to construct the CI. `WLDP` suffers from the issue of over-coverage and the post-selection method suffers from under-coverage.

Regarding the testing procedure, we report the empirical rejection rate (ERR) of the proposed testing procedures, where ERR is defined as the proportion of null hypothesis in (2) being rejected out of the 500 replications. Under the null hypothesis, ERR is an empirical measure of the type I error; under the alternative hypothesis, ERR is an empirical measure of the power. For Loading 1 (alternative hypothesis), the empirical power increases with sample sizes, for all methods. For the case that $\|x_*\|_2$ is relatively small, the proposed LiVE method has a power above 0.90 when the sample size reaches 400. For settings with large $\|x_*\|_2$, the power is not as high mainly due to the high variance of the bias-corrected estimator. This is consistent with the theoretical results established in Proposition 3. On the contrary, even with sample size as large as 600 and relatively small $\|x_*\|_2$, the test based on `WLDP` does not have a good power. For loading 2 (null hypothesis), the proposed LiVE method, `hdi` and `WLDP` have type I error controlled across all sample sizes while post selection does not have it controlled for the sample size at $n = 200$.

We have investigated the computational efficiency of all methods and reported the averaged time of implementing each method under the column indexed with "t" (the units are seconds). The proposed LiVE method is computationally efficient and can be finished within 25 seconds on average. The `hdi` algorithm provides valid CIs but requires around an hour to achieve the same goal for $n = 600$ and $p = 501$. The main reason is that the `hdi` is not designed for inference for case probabilities and requires the implementation of p high-dimensional penalization algorithm for bias-correction.

We report Root Mean Squared Error (RMSE), the bias and also the standard deviation of the proposed LiVE estimator, plug-in Lasso, post-selection, `hdi` and `WLDP` in Table 3 in Section B.1 of the supplementary material. It is observed that the plug-in Lasso estimator cannot be used for confidence interval construction as its bias component is as large as its variance component and the uncertainty of the bias component is hard to quantify.

Post selection inference methods can produce incorrect inference due to the fact that the model selection uncertainty is not quantified. In the selection step, the post-selection method can select either a larger model or a smaller model compared to the true one. As reported in Table 1, post selection is under-coverage since, in this specific simulation setting, post-selection tends to select a relatively large set of variables and this results in perfect separation of cases and controls in the re-fitting step. In Section B.3 of the supplementary material, we show another setting where the post-selection method selects a smaller model and leads to a substantial omitted variable bias.

Exact Sparse Regression Setting																	
Loading 1 (Case probability $= 0.732$**)**																	
			LiVE				Post Selection				`hdi`				`WLDP`			
$\|x_*\|_2$	r	n	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t
16.1	1	200	0.98	0.05	0.88	5	0.68	0.54	0.42	1	0.97	0.06	0.93	370	1.00	0.00	1.00	34
		400	0.97	0.10	0.81	14	0.71	0.57	0.38	2	0.96	0.10	0.87	751	1.00	0.00	1.00	56
		600	0.95	0.13	0.74	23	0.70	0.68	0.32	6	0.94	0.10	0.83	3212	1.00	0.00	1.00	118
1.90	$\frac{1}{25}$	200	0.96	0.62	0.34	5	0.80	0.77	0.31	1	0.92	0.86	0.31	371	1.00	0.36	0.58	34
		400	0.94	0.92	0.23	14	0.83	0.93	0.24	2	0.92	0.96	0.23	751	1.00	0.45	0.53	54
		600	0.95	0.95	0.19	22	0.82	0.95	0.20	5	0.95	0.97	0.19	3211	1.00	0.47	0.50	118
Loading 2 (Case probability $= 0.293$**)**																		
			LiVE				Post Selection				`hdi`				`WLDP`			
$\|x_*\|_2$	r	n	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t
16.6	1	200	0.94	0.02	0.95	4	0.66	0.20	0.62	1	0.92	0.04	0.93	370	0.98	0.00	0.97	32
		400	0.95	0.01	0.91	13	0.82	0.07	0.66	2	0.94	0.02	0.92	743	1.00	0.00	0.98	56
		600	0.96	0.03	0.90	22	0.79	0.03	0.61	5	0.95	0.03	0.89	3132	1.00	0.00	0.98	115
5.38	$\frac{1}{25}$	200	0.93	0.02	0.81	4	0.66	0.17	0.62	1	0.93	0.01	0.78	369	1.00	0.00	0.72	32
		400	0.96	0.00	0.67	13	0.83	0.03	0.67	2	0.95	0.00	0.68	742	0.99	0.00	0.65	56
		600	0.95	0.01	0.61	21	0.82	0.03	0.59	5	0.94	0.01	0.58	3131	0.98	0.00	0.60	115

Table 1: The top and bottom parts correspond to Loading 1 and 2, respectively. "r" represents the shrinkage parameter in (26). The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs, respectively; the column indexed with "ERR" represent the empirical rejection rate of the testing procedure; "t" represents the averaged computation time (in seconds). The columns under "LiVE" ,"Post Selection", "`hdi`" and "`WLDP`" correspond to the proposed estimator, the post selection estimator, the plug-in debiased estimator using `hdi` and `WLDP`, respectively.

4.3 Approximate Sparse β

In most practical settings, the regression vector β might not be exact sparse but can have some large regression coefficients and most others are small but not exactly zero. To simulate these practical settings, we consider the setting where the regression vector β is generated as a decaying loading: $\beta_1 = 0$ and $\beta_j = (j-1)^{-\text{decay}}$ for $2 \le j \le 501$, with decay $\in \{1, 2\}$. We illustrate the results using the loading x_* generated in (26). With $r = 1$, for decay $= 1$ and decay $= 2$, the case probabilities $\mathbb{P}(y_i = 1 \mid X_{i\cdot} = x_*)$ are 0.645 and 0.488, respectively; with $r = 1/25$, for decay $= 1$ and decay $= 2$, the case probabilities are 0.523 and 0.481, respectively. Note that as the regression coefficient is decaying, the shrinking parameter r in (26) plays a role in determining the case probability.

The inference results are reported in Table 2 for decay $\in \{1, 2\}$. The main observations are consistent with those for the exact sparse β, where only the proposed LiVE method and `hdi` have proper coverage across different scenarios while the CI by post selection is under-coverage and the CI by `WLDP` is over-coverage and conservative. The proposed method is computationally more efficient than `hdi`: for $n = 600$, the average computation time for the proposed algorithm is 23 seconds while `hdi` with a similar performance requires more than $3,200$ seconds.

For decay $= 1$ and $r = 1$, the case probability (0.645) is above 0.5, the proposed

LiVE method and `hdi` achieve the correct coverage level but the testing procedures have low powers. This matches with Proposition 3, that is, the power of the proposed testing procedure tends to be low for the observation x_* with very large $\|x_*\|_2$. For decay $= 1$ and $r = 1/25$, the case probability is 0.523 and this represents an alternative in the indistinguishable region and the power of the proposed testing procedure is low as expected. For decay $= 2$, the testing procedures based on the proposed LiVE, `hdi` and `WLDP` have type I error controlled for both $r = 1$ and $r = 1/25$ while the post selection method suffers from an inflated Type I error for the setting $r = 1$. The estimation results are reported in Table 4 in Section B.2 of the supplementary material. The plug-in Lasso estimator suffers from a large bias and cannot be used for confidence interval construction.

The consistent performance of the proposed inference method suggests that the proposed method not only works for the exact sparse setting, but also for the approximately sparse setting, which is simulated here to better approximate the practical data sets.

Approximate Sparse Regression Seting																			
decay $= 1$																			
				LiVE				Post Selection				`hdi`				`WLDP`			
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t
16.1	1	0.645	200	0.96	0.05	0.93	5	0.58	0.26	0.45	1	0.96	0.06	0.93	370	1.00	0.00	1.00	34
			400	0.96	0.04	0.85	14	0.60	0.31	0.41	2	0.97	0.07	0.90	751	1.00	0.00	1.00	56
			600	0.97	0.05	0.80	23	0.62	0.37	0.37	6	0.96	0.07	0.86	3212	1.00	0.00	1.00	118
1.09	$\frac{1}{25}$	0.523	200	0.96	0.06	0.40	5	0.69	0.13	0.31	1	0.96	0.05	0.39	371	1.00	0.00	0.75	34
			400	0.96	0.11	0.28	14	0.58	0.16	0.24	2	0.94	0.11	0.28	751	1.00	0.00	0.68	54
			600	0.97	0.07	0.24	22	0.71	0.09	0.21	5	0.96	0.04	0.24	3211	1.00	0.00	0.65	118
decay $= 2$																			
				LiVE				Post Selection				`hdi`				`WLDP`			
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t
16.1	1	0.488	200	0.95	0.04	0.91	5	0.66	0.14	0.37	1	0.94	0.03	0.92	370	1.00	0.00	1.00	34
			400	0.96	0.03	0.86	14	0.60	0.18	0.29	2	0.95	0.04	0.88	751	1.00	0.00	1.00	56
			600	0.96	0.03	0.78	23	0.67	0.15	0.27	6	0.93	0.03	0.85	3212	1.00	0.00	1.00	118
1.09	$\frac{1}{25}$	0.481	200	0.96	0.03	0.35	5	0.87	0.05	0.22	1	0.95	0.03	0.38	371	1.00	0.00	0.75	34
			400	0.93	0.04	0.27	14	0.83	0.06	0.15	2	0.93	0.03	0.27	751	1.00	0.00	0.68	54
			600	0.96	0.02	0.22	22	0.73	0.02	0.12	5	0.94	0.02	0.23	3211	1.00	0.00	0.65	118

Table 2: The top and bottom parts correspond to decay $= 1$ and decay $= 2$, respectively. "r" and"Prob" represent the shrinkage parameter in (26) and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs; the column indexed with "ERR" represent the empirical rejection rate of the testing procedure; "t" represents the averaged computation time (in seconds). The columns under "LiVE" ,"Post Selection", "`hdi`" and "`WLDP`" correspond to the proposed estimator, the post selection estimator, the plug-in debiased estimator using `hdi` and `WLDP` respectively.

5 Real Data Analysis

We applied the proposed methods to develop preliminary models for predicting three related disease conditions, hypertension, hypertension that appears to be resistant to standard treatment (henceforth "R-hypertension"), and hypertension with unexplained low blood potassium (henceforth "LP-hypertension"). The data were extracted from the Penn Medicine clinical data repository, including demographics, laboratory results, medication prescriptions, vital signs, and encounter meta information. The analysis cohort consisted of 348 patients who were at least 18 years old, had at least 5 office visits over at least three distinct years between 2007 and 2017, and at least 2 office visits were at one of the 37 primary care practices. Patient charts were reviewed by a dedicated physician to determine each of the three outcome statuses, and unclear cases were secondarily reviewed by an additional expert clinician. The prevalence of the three outcome variables were 39.4%, 8.1%, and 4.6%, respectively. Longitudinal EHR variables, which had varied values over multiple observations, were summarized by minimum, maximum, mean, median, standard deviation, and/or skewness, and these summary statistics were used as predictors after appropriate normalization. Highly right-skewed variables were log-transformed. We included 198 predictors in the final analyses, after removing those with missing values.

In our analysis, we randomly sampled 30 patients as the test sample, then their predictor vectors were treated as x_*. A prediction model for each outcome variable was developed using the remaining 318 patients and then applied to the test sample to obtain bias-corrected estimates of the case probabilities using our method. The left and right columns in Figure 1 present results on two independent test samples, where the three rows within each column correspond to the three outcome variables. In each panel, the x-axis represents the predicted probability generated by our method, and the y-axis represents the true outcome status (1 or 0). In all six panels, the predicted probabilities by the LiVE method for true cases tended to be high and for true controls tended to be low. This illustrates that the LiVE estimator in (12) is predictive for the true outcome status.

Figure 2 presented confidence intervals constructed using our method for the case probabilities shown in the top two panels in the right column in Figure 1, corresponding to prediction of hypertension and resistant hypertension. The length of the constructed confidence intervals appeared to vary since each patient in the test sample had different observed predictors x_*. This observation is consistent with the established theory in Theorem 1, which states that the length of confidence interval depends on $\|x_*\|_2$. More interestingly, the constructed confidence intervals appeared to be informative of the outcome statuses for the majority of the test patients. For hypertension, 80% of the confidence intervals lied either above or below 50%; For R-hypertension, 83% of the confidence intervals lie either above or below 50%.

We further divide the 30 randomly sampled observations into two subgroups by their true status and then investigate the performance of constructed confidence intervals for the subgroup of observations being cases and the other subgroup of observations

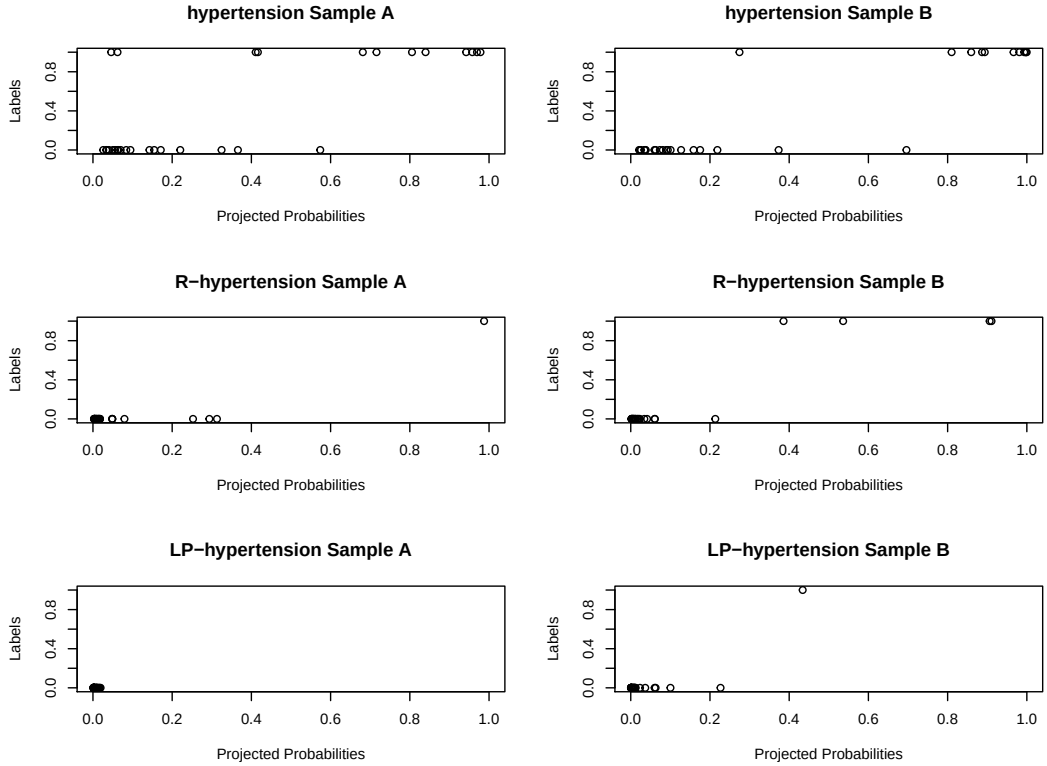

Figure 1: Performance for predicting three phenotypes in two random sub-samples.

being controls. On the left panel of Figure 2, the observations with indexes between 1 and 11 correspond to cases (observations with hypertension) while the remaining 19 observations correspond to observations without hypertension. Out of the 11 observations with hypertension, six constructed CIs are predictive with the whole interval above 0.5, one is misleading as the interval is below 0.5 and the remaining four are not predictive as the CIs come cross 0.5; Out of the 19 patients without hypertension, 17 constructed CIs are below 0.5 and hence predictive but the remaining two are not. On the right hand side of Figure 2, the observations with indexes between 1 and 4 correspond to observations with R-hypertension while the remaining 26 observations correspond to the observations without R-hypertension. Out of the four observations with R-hypertension, only one constructed CI is predictive and the other three are not; out of the 26 observations without R-hypertension, 24 are predictive and the other two are not. Overall, the constructed CIs are predictive for the outcome for 77% (hypertension), 83%(R-hypertension), and 77% (LP-hypertension) of subjects, where a constructed CI is predictive if either the constructed CI lies above 0.5 for the true case or below 0.5 for the true control. This demonstrated the practical usefulness of the developed models for evaluating the outcome status of patients, the labor-intensive chart review may be avoided for the majority of patients.

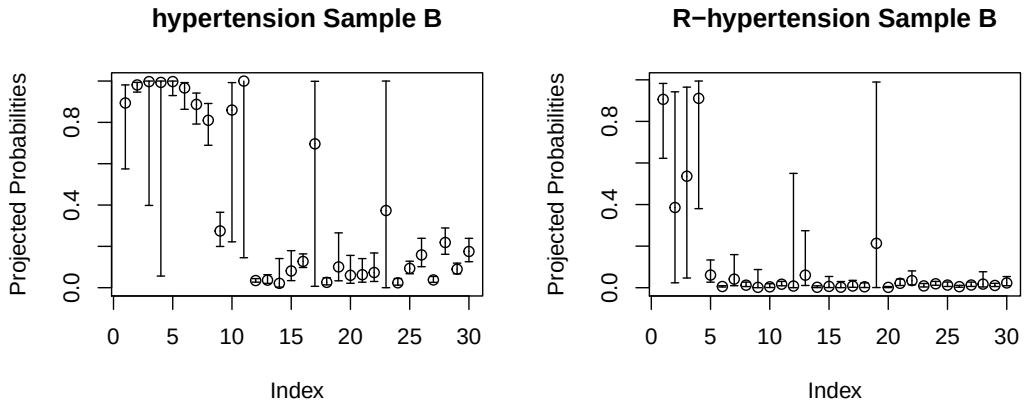

Figure 2: Confidence interval construction: on the left panel, indexes 1 to 11 correspond to observations with hypertension; indexes 12 to 30 correspond to those without hypertension. On the right panel, indexes 1 to 4 correspond to observations with R-hypertension; indexes 5 to 30 correspond to those without R-hypertension.

The additional results corresponding to the remaining four panels are presented in Figure 3 in the supplementary materials. The observation is similar to that reported in Figure 2.

6 Proof

We provide the proof of Theorem 1 in Section 6.1 and that of Lemma 1 in Section 6.2. The remaining of the proof is postponed to Section A in the supplementary material.

We introduce the following events and also two lemmas to facilitate the proof.

$$\mathcal{A}_1 = \left\{ \max_{1\le i\le n,\ 1\le j\le p} |X_{ij}| \le C\sqrt{\log n + \log p} \right\},\ \ \mathcal{A}_2 = \left\{ \min_{\|\eta\|_2=1, \|\eta_{S^c}\|_1 \le C\|\eta_S\|_1} \frac{1}{n}\sum_{i=1}^{n} \left(X_{i\cdot}^\intercal \eta\right)^2 \ge c\lambda_{\min}\left(\Sigma\right) \right\}$$
$$\mathcal{A}_3 = \left\{ \min_{1\le i\le n} \frac{\exp\left(X_{i\cdot}^\intercal \beta\right)}{\left[1+\exp\left(X_{i\cdot}^\intercal \beta\right)\right]^2} \ge c_{\min}^2 \right\},\ \ \mathcal{A}_4 = \left\{ \lambda_0 = \left\| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i X_i \right\|_\infty \le C\sqrt{\frac{\log p}{n}} \right\}$$
$$\mathcal{A}_5 = \left\{ \|\widehat{\beta} - \beta\|_2 \le C\sqrt{\frac{k\log p}{n}} \right\},\ \ \mathcal{A}_6 = \left\{ \|(\widehat{\beta}-\beta)_{S^c}\|_1 \le C_0 \|(\widehat{\beta}-\beta)_S\|_1 \right\} \tag{28}$$

where S denotes the support of the high-dimensional vector β. The following lemma 2 controls the probability of these defined events and the proof is omitted as it is similar to Lemma 4 in [9].

Lemma 2. *Suppose Conditions* (A1) *and* (A2) *hold, then*

$$\mathbb{P}\left(\cap_{i=1}^{4}\mathcal{A}_i\right) \ge 1 - \exp(-cn) - p^{-c} \tag{29}$$

and on the event $\cap_{i=1}^4 \mathcal{A}_i$*, the events* $\mathcal{A}_5$ *and* $\mathcal{A}_6$ *hold.*

The following Lemma is about the Taylor expansion of logit function and the corresponding proof is presented in Section A.4 in the supplementary material.

Lemma 3. *For* $h(x) = \frac{\exp(x)}{1+\exp(x)}$*, we have*

$$\left(h'(B)\right)^{-1}\left(h(x) - h(B)\right) = (x-a) + \int_0^1 (1-t)(x-a)^2 \frac{h''(a+t(x-a))}{h'(B)} dt. \tag{30}$$

where

$$h'(x) = \frac{\exp(x)}{(1+\exp(x))^2} \quad \textit{and} \quad h''(x) = \frac{2\exp(2x)}{(1+\exp(x))^3}$$

We further have

$$\exp\left(-|x-a|\right) \le \frac{h'(x)}{h'(B)} \le \exp\left(|x-a|\right) \quad \textit{and} \quad \left|\frac{h'(x)}{h'(B)} - 1\right| \le \exp\left(|x-a|\right) \tag{31}$$

and

$$\left|\int_0^1 (1-t)(x-a)^2 \frac{h''(a+t(x-a))}{h'(B)} dt\right| \le \exp(|x-a|)(x-a)^2 \tag{32}$$

6.1 Proof of Theorem 1

Proof of (18). We first note that, on the event $\mathcal{A}_3$,

$$\widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n X_{i\cdot} X_{i\cdot}^\intercal\right] \widehat{u} \le \mathrm{V} \le \frac{1}{c_{\min}^2} \widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n X_{i\cdot} X_{i\cdot}^\intercal\right] \widehat{u}. \tag{33}$$

To control the upper bound part $\sqrt{\mathrm{V}} \le \frac{C_0\|x_*\|_2}{n}$, we define the following events

$$\begin{aligned}
\mathcal{B}_1 &= \left\{\left\|\widehat{\Sigma}\Sigma^{-1}x_* - x_*\right\|_\infty \le \|x_*\|_2 \lambda_n\right\} \\
\mathcal{B}_2 &= \left\{\left|x_*^\intercal \widehat{\Sigma}\Sigma^{-1}x_* - \|x_*\|_2^2\right| \le \|x_*\|_2^2 \lambda_n\right\} \\
\mathcal{B}_3 &= \left\{\|X\Sigma^{-1}x_*\|_\infty \le \|x_*\|_2 \tau_n\right\}
\end{aligned} \tag{34}$$

On the event $\cap_{i=1}^3 \mathcal{B}_i$, then $u = \Sigma^{-1}x_*$ satisfies the constraints (9), (10) and (11). As a consequence, the feasible set is non-empty on the event $\cap_{i=1}^3 \mathcal{B}_i$ and we further obtain an upper bound for the minimum value, that is,

$$\mathrm{V} \le x_*^\intercal \Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}x_*/n. \tag{35}$$

The following lemma controls the probability of the above events,

Lemma 4. *Suppose Condition* (A1) *holds and* $\lambda_n \asymp \sqrt{\frac{\log p}{n}}$ *and* $\tau_n \lesssim n^{\delta}$ *for* $0 < \delta < \frac{1}{2}$, *then*

$$\mathbb{P}\left(\cap_{i=1}^{3}\mathcal{B}_i\right) \geq 1 - n^{-c} - p^{-c}. \tag{36}$$

The proof of the lower bound part $\sqrt{\mathrm{V}} \geq \frac{c_0\|x_*\|_2}{n}$ is facilitated by the optimization constraint (9). We define a proof-facilitating optimization problem,

$$\begin{aligned} \widetilde{u} = \ & \arg\min_{u\in\mathbb{R}^p} u^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)u \\ & \text{subject to } |x_*^{\intercal}\widehat{\Sigma}u - \|x_*\|_2^2| \leq \|x_*\|_2^2\lambda_n \end{aligned} \tag{37}$$

Note that $\widehat{u}$ satisfies the feasible set of (37) and hence

$$\begin{aligned} & \widehat{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widehat{u} \geq \widetilde{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widetilde{u} \\ & \geq \widetilde{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widetilde{u} + t\left((1-\lambda_n)\|x_*\|_2^2 - x_*^{\intercal}\widehat{\Sigma}\widetilde{u}\right) \text{ for any } t \geq 0, \end{aligned} \tag{38}$$

where the last inequality follows from the constraint of (37). Note that for a given $t \geq 0$, we have

$$\begin{aligned} & \widetilde{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widetilde{u} + t\left((1-\lambda_n)\|x_*\|_2^2 - x_*^{\intercal}\widehat{\Sigma}\widetilde{u}\right) \\ & \geq \min_{u\in\mathbb{R}^p} u^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)u + t\left((1-\lambda_n)\|x_*\|_2^2 - x_*^{\intercal}\widehat{\Sigma}u\right). \end{aligned} \tag{39}$$

By solving the minimization problem of the right hand side of (39), we have the minimizer u^* satisfies $\widehat{\Sigma}u^* = \frac{t}{2}\widehat{\Sigma}x_*$ and hence the minimized value of the right hand side of (39) is

$$-\frac{t^2}{4}x_*^{\intercal}\widehat{\Sigma}x_* + t(1-\lambda_n)\|x_*\|_2^2.$$

Combined with (38) and (39), we have

$$\widehat{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widehat{u} \geq \max_{t\geq 0}\left[-\frac{t^2}{4}x_*^{\intercal}\widehat{\Sigma}x_* + t(1-\lambda_n)\|x_*\|_2^2\right]. \tag{40}$$

For $t^* = 2\frac{(1-\lambda_n)\|x_*\|_2^2}{x_*^{\intercal}\widehat{\Sigma}x_*} > 0$, the minimum of the right hand side of (40) is achieved and hence establish

$$\widehat{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widehat{u} \geq \frac{(1-\lambda_n)^2\|x_*\|_2^4}{x_*^{\intercal}\widehat{\Sigma}x_*}. \tag{41}$$

Proof of (16). The limiting distribution in (16) follows from the decomposition (8), the variance control in (18), Lemma 1 and Proposition 4 and the following lemma.

Lemma 5. *Suppose that Conditions* (A1) *and* (A2) *hold and* τ_n *defined in* (11) *satisfies* $(\log n)^{1/2} \lesssim \tau_n \ll n^{1/2}$*, then*

$$\frac{1}{V^{1/2}}\widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}X_{i\cdot}\epsilon_i \to N(0,1) \tag{42}$$

where V *is defined in* (17)*.*

6.2 Proof of Lemma 1

To start the proof, we recall that $h(z) = \frac{\exp(z)}{1+\exp(z)}$ and define the functions g_i for $1 \le i \le n$

$$g_i(t_i) = \left(\left(\frac{\exp\left(X_{i\cdot}^\intercal\beta + t_i\right)}{\left(1+\exp\left(X_{i\cdot}^\intercal\beta + t_i\right)\right)^2}\right)^{-1} - \left(\frac{\exp\left(X_{i\cdot}^\intercal\beta\right)}{\left(1+\exp\left(X_{i\cdot}^\intercal\beta\right)\right)^2}\right)^{-1}\right)\widehat{u}^\intercal X_{i\cdot},$$

and the space for $\delta \in \mathbb{R}^p$ as

$$\mathcal{C} = \left\{\delta : \|\delta_{S^c}\|_1 \le c\|\delta_S\|_1,\ \|\delta\|_2 \le C^*\sqrt{\frac{k\log p}{n}}\right\}. \tag{43}$$

for some positive constants $c > 0$ and $C^* > 0$. We further define

$$\mathcal{T} = \left\{t = (t_1, \cdots, t_n) : t_i = X_{i\cdot}^\intercal\delta \text{ where } \delta \in \mathcal{C}\right\}, \tag{44}$$

We can rewrite the main component of the left hand side of (24) as

$$\begin{aligned}
&\left|\widehat{u}^\intercal\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left(\frac{\exp(X_{i\cdot}^\intercal\widehat{\beta})}{\left(1+\exp(X_{i\cdot}^\intercal\widehat{\beta})\right)^2}\right)^{-1} - \left(\frac{\exp\left(X_{i\cdot}^\intercal\beta\right)}{\left(1+\exp\left(X_{i\cdot}^\intercal\beta\right)\right)^2}\right)^{-1}\right)X_{i\cdot}\epsilon_i\right| \cdot \mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3\cap\mathcal{A}_5\cap\mathcal{A}_6} \\
&\le \sup_{\delta\in\mathcal{C}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(X_{i\cdot}^\intercal\delta)\cdot\mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right| = \sup_{t\in\mathcal{T}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(t_i)\cdot\mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right|
\end{aligned} \tag{45}$$

where $\mathcal{C}$ is defined in (43) and $\mathcal{T}$ is defined in (44). In the following, we control the last part of (45) via applying the symmetrization technique [38] stated in Lemma 6 and the contraction principle in Lemma 7. The proofs of Lemma 6 and Lemma 7 are presented in Sections A.5 and A.6 in the supplementary materials, respectively.

Lemma 6. *Suppose that* y_i' *is an independent copy of* y_i *and* ϵ_i' *is defined as* $y_i' - \mathbf{E}(y_i' \mid X_i)$*. For all convex nondecreasing functions* $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$*, then*

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} g_i(t_i)\epsilon_i\right|\right) \le \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} g_i(t_i)\xi_i\right|\right), \tag{46}$$

where

$$\xi_i = \epsilon_i - \epsilon_i' = y_i - y_i'.$$

The following lemma is a modification of Theorem 2.2 in [26], where the result in [26] was only developed for i.i.d Rademacher variables ξ_i. The following lemma is more general in the sense that $\xi_1, \xi_2, \cdots, \xi_n$ are only required to be independent and satisfy the probability distribution (49).

Lemma 7. *Let $t = (t_1, \cdots, t_n) \in \mathcal{T} \subset R^n$ and let $\phi_i : \mathbb{R} \to \mathbb{R}, i = 1, \cdots, n$ be functions such that $\phi_i(0) = 0$ and*

$$|\phi_i(u) - \phi_i(v)| \leq |u - v|, u, v \in \mathbb{R}. \tag{47}$$

For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, then

$$\mathbf{E}\Phi\left(\frac{1}{2}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\phi_i(t_i)\xi_i\right|\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}t_i\xi_i\right|\right), \tag{48}$$

where $\{\xi_i\}_{1\leq i\leq n}$ are independent random variables with the probability density function

$$\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) \in [0, \frac{1}{2}],\ \mathbb{P}(\xi_i = 0) = 1 - 2\mathbb{P}(\xi_i = 1). \tag{49}$$

We will apply Lemmas 6 and 7 and provide a sharp control for $\sup_{t\in\mathcal{T}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(t_i)\cdot \mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right|$ in (45). For $t, s \in \mathcal{T} \subset \mathbb{R}^n$, then there exist $\delta^t, \delta^s \in \mathcal{C} \subset \mathbb{R}^p$ such that

$$t_i - s_i = X_{i\cdot}^{\intercal}\left(\delta^t - \delta^s\right) \quad \text{and} \quad t_i = X_{i\cdot}^{\intercal}\delta^t \quad \text{for } 1 \leq i \leq n.$$

Hence on the event $\mathcal{A}_1$,

$$\max\left\{\max_{1\leq i\leq n}|t_i - s_i|, \max_{1\leq i\leq n}|t_i|\right\} \leq Ck\sqrt{\frac{\log p}{n}}\sqrt{\log p + \log n} \leq 1. \tag{50}$$

where the last inequality follows as long as $\sqrt{n} \geq k\log p\left(1 + \sqrt{\frac{\log n}{\log p}}\right)$

Define $q(x) = \left(\frac{\exp(x)}{(1+\exp(x))^2}\right)^{-1}$ and then

$$\begin{aligned} g_i(s_i) - g_i(t_i) &= \left(q\left(X_{i\cdot}^{\intercal}\beta + s_i\right) - q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)\right)\widehat{u}^{\intercal}X_{i\cdot} \\ &= \left(\frac{q\left(X_{i\cdot}^{\intercal}\beta + s_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)} - 1\right)\frac{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta\right)}q\left(X_{i\cdot}^{\intercal}\beta\right)\widehat{u}^{\intercal}X_{i\cdot} \end{aligned} \tag{51}$$

By (31), we have

$$\left|\frac{q\left(X_{i\cdot}^{\intercal}\beta + s_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)} - 1\right| \leq |\exp(|s_i - t_i|) - 1| \leq e|s_i - t_i|, \tag{52}$$

where the last inequality holds as long as $|s_i - t_i|$ is sufficiently small, as verified in (50). Similarly, we establish that $\frac{q(X_{i\cdot}^\intercal \beta + t_i)}{q(X_{i\cdot}^\intercal \beta)} \leq e$. Combined with (51) and (52), we obtain

$$|g_i(s_i) - g_i(t_i)| \leq \frac{1}{c_{\min}^2} e^2 \, |s_i - t_i| \, |\widehat{u}^\intercal X_{i\cdot}| \leq \frac{1}{c_{\min}^2} e^2 \, |s_i - t_i| \, \|x_*\|_2 \tau_n, \tag{53}$$

where the last inequality follows from the constraint (11). By applying (53), we have

$$\frac{1}{L_n} |g_i(t_i) - g_i(s_i)| \cdot \mathbf{1}_{\mathcal{A}_1 \cap \mathcal{A}_3} \leq |t_i - s_i| \quad \text{where} \quad L_n = \frac{e^2}{c_{\min}^2} \|x_*\|_2 \tau_n. \tag{54}$$

Define $\phi_i(t_i) = \frac{1}{L_n} g_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1}$. We then apply (46) and (48) with $\Phi(x) = x$ and obtain

$$\mathbf{E}_{\xi|X} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n \phi_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1 \cap \mathcal{A}_3} \xi_i \right| \leq 2 \mathbf{E}_{\xi|X} \sup_{\delta \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \delta^\intercal X_{i\cdot} \xi_i \right|$$

and hence

$$\mathbf{E} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n \phi_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1 \cap \mathcal{A}_3} \xi_i \right| \leq 2 \mathbf{E} \sup_{\delta \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \delta^\intercal X_{i\cdot} \xi_i \right|.$$

Note that

$$\mathbf{E} \sup_{\delta \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \delta^\intercal X_{i\cdot} \xi_i \right| \leq \sup_{\delta \in \mathcal{C}} \|\delta\|_1 \mathbf{E} \left\| \frac{1}{n} \sum_{i=1}^n X_{i\cdot} \xi_i \right\|_\infty \leq \sup_{\delta \in \mathcal{C}} \|\delta\|_1 \sqrt{\frac{2 \log p}{n}} \|X_{i\cdot}\|_{\psi_2},$$

where the last inequality follows from the fact that $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_{i\cdot} \xi_i$ is sub-gaussian random variable with sub-gaussian norm $\|X_{i\cdot}\|_{\psi_2}$. Combined with the fact that $\sup_{\delta \in \mathcal{C}} \|\delta\|_1 \leq Ck\sqrt{\frac{\log p}{n}}$, we establish $\mathbf{E} \sup_{\delta \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \delta^\intercal X_{i\cdot} \xi_i \right| \leq C \frac{k \log p}{n} \|X_{i\cdot}\|_{\psi_2}$ and hence

$$\mathbf{E} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n \phi_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1} \xi_i \right| \leq C \frac{k \log p}{n} \|X_{i\cdot}\|_{\psi_2}.$$

By Chebyshev's inequality, we have

$$\mathbb{P} \left(\sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n g_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1} \xi_i \right| \geq Ct \|x_*\|_2 \tau_n \frac{k \log p}{n} \|X_{i\cdot}\|_{\psi_2} \right) \leq \frac{1}{t}$$

By (45), we establish that (24) holds with probability larger than $1 - (\frac{1}{t} + p^{-c} + \exp(-cn))$.

References

[1] Zakhriya Alhassan, David Budgen, Riyad Alshammari, and Noura Al Moubayed. Predicting current glycated hemoglobin levels in adults from electronic health records: Validation of multiple logistic regression algorithm. *JMIR Medical Informatics*, 2020.

[2] 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.

[3] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[4] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[5] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[6] 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.

[7] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[8] Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+ \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[9] T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

[10] Tianxi Cai, Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

[11] Victor M Castro, Jessica Minnier, Shawn N Murphy, Isaac Kohane, Susanne E Churchill, Vivian Gainer, Tianxi Cai, Alison G Hoffnagle, Yael Dai, Stefanie Block, Sydney R Weill, Mireya Nadal-Vicens, Alisha R Pollastri, J Niels Rosenquist, Sergey Goryachev, Dost Ongur, Pamela Sklar, Roy H Perlis, Jordan W Smoller, and International Cohort Collection for Bipolar Disorder Consortium. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *The American Journal of Psychiatry*, 172:363–372, 2015.

[12] Cristiana Catena, GianLuca Colussi, Roberta Lapenna, Elisa Nadalini, Alessandra Chiuch, Pasquale Gianfagna, and Leonardo A. Sechi. Long-term cardiac effects of adrenalectomy or mineralocorticoid antagonists in patients with primary aldosteronism. *Hypertension*, 50:911–918, 2007.

[13] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[14] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[15] Muhammad Faisal, Andy Scally, Robin Howes, Kevin Beatson, Donald Richardson, and Mohammed A Mohammed. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. *Health Informatics Journal*, 26:34–44, 2020.

[16] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[17] John Fox. *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[18] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 2010.

[19] John W. Funder, Robert M. Carey, Franco Mantero, M. Hassan Murad, Martin Reincke, Hirotaka Shibata, Michael Stowasser, and William F. Young Jr. The management of primary aldosteronism: Case detection, diagnosis, and treatment: An endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology & Metabolism*, 101:1889–1916, 2016.

[20] A Hannemann and H Wallaschofski. Prevalence of primary aldosteronism in patient's cohorts and in population-based studies–a review of the current literature. *Hormone and Metabolic Research*, 44:157–162, 2011.

[21] Na Honga, Andrew Wena, Daniel J. Stonea, Shintaro Tsujia, Paul R. Kingsburya, Luke V. Rasmussenb, Jennifer A. Pachecob, Prakash Adekkanattuc, Fei Wangc, Yuan Luob, Jyotishman Pathakc, Hongfang Liua, and Guoqian Jiang. Developing a fhir-based ehr phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *Journal of Biomedical Informatics*, 2019.

[22] Jian Huang and Cun-Hui Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13(Jun):1839–1864, 2012.

[23] 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.

[24] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[25] Sabine C Kayser, Tanja Dekkers, Hans J Groenewoud, Gert Jan van der Wilt, J Carel Bakx, Mark C van der Wel, Ad R Hermus, Jacques W Lenders, and Jaap Deinum. Study heterogeneity and estimation of prevalence of primary aldosteronism: A systematic review and meta-regression analysis. *The Journal of Clinical Endocrinology & Metabolism*, 101:2826–2835, 2016.

[26] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d?Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[27] Yen-Hung Lin, Lian-Yu Lin, Aaron Chen, Xue-Ming Wu, Jen-Kuang Lee, Ta-Chen Su, Vin-Cent Wu, Shih-Chieh Chueh, Wei-Chou Lin, Men-Tzung Lo, Pa-Chun Wang, Yi-Lwun Ho, Kwan-Dun Wu, and TAIPAI Study Group. Adrenalectomy improves increased carotid intima-media thickness and arterial stiffness in patients with aldosterone producing adenoma. *Atherosclerosis*, 221:154–159, 2012.

[28] Rong Ma, T Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *arXiv preprint arXiv:1805.06970*, 2018.

[29] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[30] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[31] Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

[32] Ravi B. Parikh, Christopher Manz, Corey Chivers, Susan Harkness Regli, Jennifer Braun, Michael E. Draugelis, Lynn M. Schuchter, Lawrence N. Shulman, Amol S. Navathe, Mitesh S. Pate, and Nina R. OConnor. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Network Open*, 2019.

[33] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[34] Martin Reincke, Evelyn Fischer, Sabine Gerum, Katrin Merkle, Sebastian Schulz, Anna Pallauf, Marcus Quinkler, Gregor Hanslik, Katharina Lang, Stefanie Hahner, Bruno Allolio, Christa Meisinger, Rolf Holle, Felix Beuschlein, Martin Bidlingmaier, Stephan Endres, and German Conn's Registry-Else Krner-Fresenius-Hyperaldosteronism Registry. Observational study mortality in treated primary aldosteronism: the german conn's registry. *Hypertension*, 60:618–624, 2012.

[35] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[36] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[37] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175(1-2):487–558, 2019.

[38] Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge UP, 2006.

[39] 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.

[40] 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.

[41] 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 Additional Proofs

A.1 Proof of Proposition 4

Proof of (20) The first inequality of (20) follows from Holder's inequality and the second inequality follows from Condition (B).
Proof of (21) By Cauchy inequality, we have

$$\sqrt{n}\left|\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\Delta_i\right| \leq \max_{1\leq i\leq n}|\widehat{u}^{\intercal}X_{i\cdot}|\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\Delta_i| \leq \tau_n\|x_*\|_2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\Delta_i| \tag{55}$$

By Lemma 3, we have $|\Delta_i| \leq \exp\left(|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)|\right)\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2$. On the event $\mathcal{A} = \cap_{i=1}^{6}\mathcal{A}_i$, we have

$$\begin{aligned}&\sum_{i=1}^{n}|\Delta_i| \leq \sum_{i=1}^{n}\exp\left(|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)|\right)\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\\ &\leq \exp\left(\max|X_{ij}|\cdot\|\widehat{\beta}-\beta\|_1\right)\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \leq C\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2.\end{aligned} \tag{56}$$

where the second inequality follows from Holder inequality and the last inequality follows from the fact that $\sqrt{n} \gg k\log p\left(1+\sqrt{\frac{\log n}{\log p}}\right)$. On the event $\mathcal{A}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \leq C\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{k\log p}{n}. \tag{57}$$

Together with (55) and (56), we establish that, on the event $\mathcal{A}$,

$$\sqrt{n}\left|\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\Delta_i\right| \leq C\tau_n\|x_*\|_2\frac{k\log p}{\sqrt{n}}. \tag{58}$$

A.2 Proof of Lemma 5

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2}\right)^{-1}\widehat{u}^{\intercal}X_{i\cdot}\epsilon_i \to N(0,\mathrm{V}) \tag{59}$$

Define

$$W_i = \frac{1}{\sqrt{\mathrm{V}}}\left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2}\right)^{-1}\widehat{u}^{\intercal}X_{i\cdot}\epsilon_i \tag{60}$$

Conditioning on X, then $\{W_i\}_{1\le i\le n}$ are independent random variables with $\mathbf{E}(W_i \mid X_i) = 0$ and $\sum_{i=1}^n \mathrm{Var}(W_i \mid X_i) = n^2$. To establish (59), it is sufficient to check the Lindeberg's condition, that is, for any constant $\bar{\epsilon} > 0$,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \mathbf{E}\left(W_i^2 \mathbf{1}_{\{|W_i|\ge \bar{\epsilon}\sqrt{n}\}}\right) = 0. \tag{61}$$

Note that

$$\max_{1\le i\le n} \left| \frac{1}{\sqrt{\mathrm{V}}} \left(\frac{\exp\left(X_{i\cdot}^\intercal \beta\right)}{\left(1+\exp\left(X_{i\cdot}^\intercal \beta\right)\right)^2} \right)^{-1} \widehat{u}^\intercal X_{i\cdot} \epsilon_i \right| \le 2 \frac{\max_{1\le i\le n} |\widehat{u}^\intercal X_{i\cdot}|}{\sqrt{V}} \le \frac{2\tau_n \|x_*\|_2}{\sqrt{V}} \le \bar{\epsilon}\sqrt{n} \tag{62}$$

where the first inequality follows from the fact that $\left(\frac{\exp\left(X_{i\cdot}^\intercal \beta\right)}{\left(1+\exp\left(X_{i\cdot}^\intercal \beta\right)\right)^2} \right)^{-1} \epsilon_i \le \frac{2}{c_{\min}}$, the second inequality follows from $|\widehat{u}^\intercal X_{i\cdot}| \le \tau_n \|x_*\|_2$ and the last inequality follows from (18) and the condition $\tau_n \ll \sqrt{n}$. Then (61) follows from (62) and by Lindeberg's central limit theorem, we establish (59).

A.3 Proof of Proposition 1

For $t \in (0,1)$, by the definition of $\widehat{\beta}$, we have

$$\ell(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_1 \le \ell(\widehat{\beta} + t(\beta - \widehat{\beta})) + \lambda\|\widehat{\beta} + t(\beta - \widehat{\beta})\|_1 \le \ell(\widehat{\beta} + t(\beta - \widehat{\beta})) + (1-t)\lambda\|\widehat{\beta}\|_1 + t\lambda\|\beta\|_1 \tag{63}$$

where $\ell(\beta) = \frac{1}{n}\sum_{i=1}^n \left(\log\left(1+\exp\left(X_{i\cdot}^\intercal\beta\right)\right) - y_i \cdot \left(X_{i\cdot}^\intercal \beta\right)\right)$. Then we have

$$\frac{\ell(\widehat{\beta}) - \ell(\widehat{\beta} + t(\beta - \widehat{\beta}))}{t} + \lambda\|\widehat{\beta}\|_1 \le \lambda\|\beta\|_1 \quad \text{for any } t \in (0,1) \tag{64}$$

and taking the limit $t \to 0$, we have

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{\exp(X_{i\cdot}^\intercal \widehat{\beta})}{1+\exp(X_{i\cdot}^\intercal \widehat{\beta})} - y_i \right) X_{i\cdot}^\intercal (\widehat{\beta} - \beta) + \lambda\|\widehat{\beta}\|_1 \le \lambda\|\beta\|_1 \tag{65}$$

Note that

$$\left(\frac{\exp(X_{i\cdot}^\intercal \widehat{\beta})}{1+\exp(X_{i\cdot}^\intercal \widehat{\beta})} - y_i \right) X_{i\cdot}^\intercal (\widehat{\beta} - \beta) = \left(-\epsilon_i + \left(\frac{\exp(X_{i\cdot}^\intercal \widehat{\beta})}{1+\exp(X_{i\cdot}^\intercal \widehat{\beta})} - \frac{\exp\left(X_{i\cdot}^\intercal \beta\right)}{1+\exp\left(X_{i\cdot}^\intercal \beta\right)} \right) \right) X_{i\cdot}^\intercal (\widehat{\beta} - \beta)$$

$$= -\epsilon_i X_{i\cdot}^\intercal (\widehat{\beta} - \beta) + \int_0^1 \frac{\exp\left(X_{i\cdot}^\intercal \beta + t X_{i\cdot}^\intercal (\widehat{\beta} - \beta)\right)}{\left[1+\exp\left(X_{i\cdot}^\intercal \beta + t X_{i\cdot}^\intercal (\widehat{\beta} - \beta)\right)\right]^2} \left(X_{i\cdot}^\intercal (\widehat{\beta} - \beta)\right)^2 dt \tag{66}$$

By (31), we have

$$\begin{aligned}
&\frac{\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)\right]^2} \geq \frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\exp\left(-t\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right) \\
&\geq \frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\exp\left(-t\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)
\end{aligned} \tag{67}$$

Combined with (66), we have

$$\begin{aligned}
&\int_0^1 \frac{\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 dt \\
&\geq \frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\int_0^1\exp\left(-t\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)dt \\
&= \frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}
\end{aligned} \tag{68}$$

Together with (65), we have

$$\begin{aligned}
&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\left(\frac{1}{n}\sum_{i=1}^n\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\right)+\lambda\|\widehat{\beta}\|_1 \\
&\leq \lambda\|\beta\|_1+\frac{1}{n}\sum_{i=1}^n\epsilon_i X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\leq\lambda\|\beta\|_1+\lambda_0\|\widehat{\beta}-\beta\|_1.
\end{aligned} \tag{69}$$

By the fact that $\|\widehat{\beta}\|_1=\|\widehat{\beta}_S\|_1+\|\widehat{\beta}_{S^c}-\beta_{S^c}\|_1$ and $\|\beta\|_1-\|\widehat{\beta}_S\|_1\leq\|\widehat{\beta}_S-\beta_S\|_1$, then we have

$$\begin{aligned}
&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\left(\frac{1}{n}\sum_{i=1}^n\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\right) \\
&+\delta_0\lambda_0\|\widehat{\beta}_{S^c}-\beta_{S^c}\|_1\leq(2+\delta_0)\,\lambda_0\|\widehat{\beta}_S-\beta_S\|_1
\end{aligned} \tag{70}$$

Then we deduce (15) and

$$\begin{aligned}
&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\left(\frac{1}{n}\sum_{i=1}^n\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2\right) \\
&\leq(2+\delta_0)\,\lambda_0\|\widehat{\beta}_S-\beta_S\|_1.
\end{aligned} \tag{71}$$

Lemma 8. *On the event $\mathcal{A}_2 \cap \mathcal{A}_3$, then*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^2}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \geq c\lambda_{\min}\left(\Sigma\right)\|\widehat{\beta}-\beta\|_2^2 \tag{72}$$

Then (71) is further simplified as

$$\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}c\lambda_{\min}\left(\Sigma\right)\|\widehat{\beta}-\beta\|_2^2 \leq (2+\delta_0)\,\lambda_0\|\widehat{\beta}_S-\beta_S\|_1 \tag{73}$$

Case 1: Assume that

$$\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right| \leq c_1 \quad \text{for some } c_1>0 \tag{74}$$

then we have

$$\begin{aligned}&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|} = \int_0^1 \exp\left(-t\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)dt\\ &\geq \int_0^1 \exp\left(-tc_1\right)dt = \frac{1-\exp\left(-c_1\right)}{c_1}\end{aligned} \tag{75}$$

Define $c_2 = \frac{c\lambda_{\min}(\Sigma)}{2+\delta_0}\frac{1-\exp(-c_1)}{c_1}$, then we have

$$c_2\|\widehat{\beta}-\beta\|_2^2 \leq \lambda_0\|\widehat{\beta}_S-\beta_S\|_1 \leq \sqrt{k}\lambda_0\|\widehat{\beta}_S-\beta_S\|_2 \tag{76}$$

and hence

$$\|\widehat{\beta}-\beta\|_2 \lesssim \frac{1}{\lambda_{\min}}\sqrt{k}\lambda_0 \quad \text{and} \quad \|\widehat{\beta}-\beta\|_1 \leq k\lambda_0 \tag{77}$$

Case 2: Assume that (74) does not hold, then

$$\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|} \geq \frac{1-\exp(-c_1)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|} \tag{78}$$

Together with (73), we have

$$c_2c_1\|\widehat{\beta}-\beta\|_2^2 \leq \lambda_0\|\widehat{\beta}_S-\beta_S\|_1\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right| \tag{79}$$

By $\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right| \leq \max|X_{ij}|\,\|\widehat{\beta}-\beta\|_1$ and (15), we further have

$$\begin{aligned}\lambda_0\|\widehat{\beta}_S-\beta_S\|_1\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right| &\leq \frac{2+2\delta_0}{\delta_0}\max|X_{ij}|\,\lambda_0\|\widehat{\beta}_S-\beta_S\|_1^2\\ &\leq \frac{2+2\delta_0}{\delta_0}\max|X_{ij}|\,k\lambda_0\|\widehat{\beta}_S-\beta_S\|_2^2,\end{aligned} \tag{80}$$

where the last inequality follows from Cauchy inequality. Combining (79) and (80), we have shown that if (74) does not hold, then

$$\max |X_{ij}| \frac{2+2\delta_0}{\delta_0} k\lambda_0 \geq c_2 c_1, \tag{81}$$

Since this contradicts the assumption that $\max |X_{ij}| \, k\lambda_0 < \frac{c_2 c_1 \delta_0}{2+2\delta_0}$, we establish (77) and hence (15).

A.4 Proof of Lemma 3

We first introduce the following version of Taylor expansion.

Lemma 9. *If $f''(x)$ is continuous on an open interval $\mathcal{I}$ that contains a, and $x \in \mathcal{I}$, then*

$$f(x) - f(a) = f'(a)(x-a) + \int_0^1 (1-t)(x-a)^2 f''(a+t(x-a))dt \tag{82}$$

By applying Lemma 9, we have

$$h(x) - h(a) = h'(a)(x-a) + \int_0^1 (1-t)(x-a)^2 h''(a+t(x-a))dt$$

Divide both sides by $(h'(a))^{-1}$, we establish (30). The inequality (31) follows from

$$\frac{h'(x)}{h'(a)} = \exp(x-a)\frac{[1+\exp(a)]^2}{[1+\exp(x)]^2} \leq \exp(x-a)\exp(2(a-x)_+) = \exp(|x-a|)$$

and

$$\frac{h'(a)}{h'(x)} \leq \exp(|x-a|)$$

The control of (32) follows from the following inequality,

$$\begin{aligned}\frac{h''(a+t(x-a))}{h'(a)} &= \frac{2\exp(2a+2t(x-a))}{(1+\exp(a+t(x-a)))^3} \cdot \frac{(1+\exp(a))^2}{\exp(a)} \\ &\leq 2\exp(t(x-a)) \cdot \frac{(1+\exp(a))^2}{(1+\exp(a+t(x-a)))^2} \leq 2\exp(t|x-a|)\end{aligned}$$

A.5 Proof of Lemma 6

We start with the conditional expectation $\mathbf{E}_{y|X}\Phi\left(\sup_{t\in\mathcal{T}}|\sum_{i=1}^n g_i(t_i)\epsilon_i|\right)$ and note that

$$\mathbf{E}_{y|X}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)\epsilon_i\right|\right) = \mathbf{E}_{y|X}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)\epsilon_i - \mathbf{E}_{y'|X}\sum_{i=1}^n g_i(t_i)\epsilon_i'\right|\right).$$

Since $\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)\epsilon_i - \mathbf{E}_{y'|X}\sum_{i=1}^n g_i(t_i)\epsilon_i'\right| \le \mathbf{E}_{y'|X}\sup_{t\in\mathcal{T}}|\sum_{i=1}^n g_i(t_i)(\epsilon_i-\epsilon_i')|$ and Φ is a non-decreasing function, we have

$$\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)\epsilon_i - \mathbf{E}_{y'|X}\sum_{i=1}^n g_i(t_i)\epsilon_i'\right|\right) \le \Phi\left(\mathbf{E}_{y'|X}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)(\epsilon_i-\epsilon_i')\right|\right).$$

Since Φ is a convex function, we have

$$\mathbf{E}_{y|X}\Phi\left(\mathbf{E}_{y'|X}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)(\epsilon_i-\epsilon_i')\right|\right) \le \mathbf{E}_{(y,y')|X}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n g_i(t_i)(\epsilon_i-\epsilon_i')\right|\right).$$

Integration of both sides of the above inequality leads to (46).

A.6 Proof of Lemma 7

The proof follows from that of Theorem 2.2 in [26] and some modification is necessary to extend the results to the general random variables $\xi_1, \xi_2, \cdots, \xi_n$ which are independent and follow the probability distribution (49).

We start with proving the following inequality for a function $A : \mathcal{T} \to \mathbb{R}$,

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[A(t) + \sum_{i=1}^n \phi_i(t_i)\xi_i]\right) \le \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[A(t) + \sum_{i=1}^n t_i\xi_i]\right), \tag{83}$$

We first prove the special case $n = 1$, which is reduced to be the following inequality,

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + \phi(t_2)\xi_0]\right) \le \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + t_2\xi_0]\right), \tag{84}$$

where $\mathcal{T} \subset \mathbb{R}^2$ and $\mathbb{P}(\xi_0 = 1) = \mathbb{P}(\xi_0 = -1) \in [0, \frac{1}{2}]$ and $\mathbb{P}(\xi_0 = 0) = 1 - 2\mathbb{P}(\xi = 1)$. It suffices to verify (84) by establishing the following inequality,

$$\begin{aligned}
&\mathbb{P}(\xi_0 = 1)\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + \phi(t_2)]\right) + \mathbb{P}(\xi_0 = -1)\Phi\left(\sup_{t\in\mathcal{T}}[t_1 - \phi(t_2)]\right) + \mathbb{P}(\xi_0 = 0)\Phi\left(\sup_{t\in\mathcal{T}}[t_1]\right) \\
&\le \mathbb{P}(\xi_0 = 1)\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + t_2]\right) + \mathbb{P}(\xi_0 = -1)\Phi\left(\sup_{t\in\mathcal{T}}[t_1 - t_2]\right) + \mathbb{P}(\xi_0 = 0)\Phi\left(\sup_{t\in\mathcal{T}}[t_1]\right)
\end{aligned}$$

This is equivalent to show

$$\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + \phi(t_2)]\right) + \Phi\left(\sup_{t\in\mathcal{T}}[t_1 - \phi(t_2)]\right) \le \Phi\left(\sup_{t\in\mathcal{T}}[t_1 + t_2]\right) + \Phi\left(\sup_{t\in\mathcal{T}}[t_1 - t_2]\right) \tag{85}$$

The above inequality follows from the same line of proof as that in [26]. It remains to

prove the lemma by applying induction and (84), that is,

$$\mathbf{E}_{(\xi_1,\cdots,\xi_n)|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n}\phi_i(t_i)\xi_i]\right)=\mathbf{E}_{(\xi_1,\cdots,\xi_{n-1})|X}\mathbf{E}_{\xi_n|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n}\phi_i(t_i)\xi_i]\right)$$
$$\leq \mathbf{E}_{(\xi_1,\cdots,\xi_{n-1})|X}\mathbf{E}_{\xi_n|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n-1}\phi_i(t_i)\xi_i+t_n\xi_n]\right)$$
$$=\mathbf{E}_{\xi_n|X}\mathbf{E}_{(\xi_1,\cdots,\xi_{n-1})|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n-1}\phi_i(t_i)\xi_i+t_n\xi_n]\right)$$

Continuing the above equation, we establish $\mathbf{E}_{(\xi_1,\cdots,\xi_n)|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n}\phi_i(t_i)\xi_i]\right)\leq \mathbf{E}_{(\xi_1,\cdots,\xi_n)|X}\Phi\left(\sup_{t\in\mathcal{T}}[A(t)+\sum_{i=1}^{n}t_i\xi_i]\right)$. Integration with respect to X leads to (83). In the following, we will apply (83) to establish (48). Note that

$$\mathbf{E}\Phi\left(\frac{1}{2}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\phi_i(t_i)\xi_i\right|\right)=\mathbf{E}\Phi\left(\frac{1}{2}\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\phi_i(t_i)\xi_i\right)_+\!\!+\frac{1}{2}\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\phi_i(t_i)(-\xi_i)\right)_+\right)$$
$$\leq\frac{1}{2}\left[\mathbf{E}\Phi\left(\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\phi_i(t_i)\xi_i\right)_+\right)+\mathbf{E}\Phi\left(\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\phi_i(t_i)(-\xi_i)\right)_+\right)\right]$$

By applying (83) to the function $u\to\Phi(u_+)$, which is convex and non-decreasing, we have $\mathbf{E}\Phi\left(\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\phi_i(t_i)\xi_i\right)_+\right)\leq\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}t_i\xi_i\right)\leq\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}|\sum_{i=1}^{n}t_i\xi_i|\right)$. Then we establish (48).

B Additional Simulation Study

B.1 Exact Sparse

We consider the exact sparse regression setup (4.2) and report the results on estimation accuracy in the following. In Table 3, we compare the proposed estimator, the post selection estimator, the plug-in `hdi` and `WLDP` and Lasso estimator in terms of Root Mean Squared Error (RMSE), bias and standard error. Through comparing the proposed and plug-in Lasso estimators, we observe that the bias component is reduced at the expense of increasing the variance. Although the bias component is reduced, the total RMSE is not necessarily decreasing after correcting the bias, since the increased variance can lead to a larger RMSE in total. The increase in variance is proportional to the loading norm $\|x_*\|_2$; specifically, if the loading norm is large, we may suffer from a larger total RMSE after bias-correction; if the loading norm is relatively small, the variance only increases slightly and the total RMSE decreases due to the reduction of the bias; see loading 1 with $r=1/25$. This matches with the theoretical results presented in Theorem 1.

Exact Sparse Regression Setting																	
Loading 1 (Case probability $= 0.732$**)**																	
			LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.1	1	200	0.33	-0.10	0.32	0.26	-0.02	0.26	0.38	-0.11	0.37	0.37	-0.04	0.37	0.14	-0.11	0.09
		400	0.26	-0.04	0.26	0.21	-0.02	0.21	0.31	-0.06	0.31	0.31	0.01	0.31	0.11	-0.08	0.07
		600	0.24	-0.05	0.24	0.20	-0.02	0.20	0.30	-0.08	0.29	0.30	-0.03	0.30	0.08	-0.06	0.06
1.09	$\frac{1}{25}$	200	0.10	-0.03	0.09	0.14	0.04	0.14	0.07	0.02	0.07	0.11	0.10	0.06	0.13	-0.11	0.07
		400	0.07	-0.02	0.07	0.09	0.03	0.09	0.06	0.02	0.06	0.09	0.08	0.05	0.10	-0.08	0.06
		600	0.06	-0.02	0.06	0.07	0.02	0.07	0.05	0.01	0.05	0.08	0.07	0.04	0.08	-0.06	0.05
Loading 2 (Case probability $= 0.293$**)**																	
			LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.6	1	200	0.45	0.16	0.42	0.43	0.19	0.40	0.44	0.17	0.40	0.35	0.12	0.32	0.29	0.20	0.22
		400	0.38	0.10	0.37	0.33	0.08	0.32	0.39	0.10	0.37	0.29	0.08	0.28	0.23	0.13	0.19
		600	0.36	0.04	0.35	0.25	0.03	0.25	0.34	0.05	0.34	0.22	0.01	0.22	0.22	0.13	0.19
5.38	$\frac{1}{25}$	200	0.29	0.05	0.29	0.39	0.14	0.37	0.28	0.07	0.27	0.22	-0.02	0.22	0.27	0.17	0.21
		400	0.21	0.01	0.21	0.30	0.06	0.29	0.21	0.03	0.21	0.20	-0.03	0.20	0.22	0.13	0.18
		600	0.22	0.03	0.22	0.24	0.01	0.24	0.20	0.03	0.20	0.17	0.01	0.17	0.21	0.11	0.18

Table 3: The top and bottom parts correspond to Loading 1 and 2, respectively. "r" represents the shrinkage parameter in (26). The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively. The columns under "LiVE", "Post Selection", "`hdi`", "`WLDP`" and "Lasso" correspond to the proposed estimator, the post model selection estimator, the plug-in `hdi`, the plug-in `WLDP` and the plug-in Lasso estimator respectively.

B.2 Approximate Sparse

We report the estimation accuracy results in the approximate sparse regression setup (4.3) with decay $\in \{1, 2\}$. Table 4 summarizes the estimation accuracy results for Loading 1 and the results are similar to the exact sparse setting in Table 3. Table 4 shows again the plug-in Lasso estimator cannot be used for confidence interval construction owing to its large bias.

B.3 Comparison of Proposed Method with Post Selection Method

We now consider a challenging setting for post-selection and compare the post-selection method with the proposed LiVE method. We set the values for the 500 regression coefficients as

$$\beta_j = \begin{cases} (j-1)/20 & \text{for } 2 \le j \le 11 \quad ; \quad j \ne 9, 10 \\ 0.01 & \text{for } j = 9, 10 \\ 0 & \text{for } 12 \le j \le 501 \end{cases}$$

We set the intercept β_1 to be 0. The loading x_* is generated as follows :
Loading 3: We set $x_{\text{basis},1} = 1$ and generate $x_{\text{basis},-1} \in \mathbb{R}^{500}$ following $N(0, \Sigma)$ with

Approximate Sparse Regression Setting																		
decay $= 1$																		
				LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	Prob	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.1	1	0.645	200	0.37	-0.03	0.37	0.31	-0.09	0.31	0.38	-0.04	0.38	0.38	-0.02	0.38	0.18	-0.14	0.11
			400	0.29	-0.05	0.28	0.28	-0.10	0.27	0.32	-0.06	0.31	0.32	0.02	0.32	0.16	-0.13	0.09
			600	0.24	-0.03	0.24	0.23	-0.05	0.23	0.28	-0.02	0.28	0.29	0.07	0.29	0.15	-0.13	0.08
1.09	$\frac{1}{25}$	0.523	200	0.10	-0.01	0.10	0.18	-0.03	0.18	0.10	0.02	0.10	0.12	0.05	0.11	0.07	-0.03	0.06
			400	0.08	-0.01	0.08	0.13	-0.01	0.13	0.08	0.01	0.08	0.09	0.04	0.09	0.05	-0.03	0.04
			600	0.06	-0.02	0.06	0.09	-0.03	0.08	0.05	0.01	0.05	0.05	0.05	0.05	0.05	-0.03	0.04
decay $= 2$																		
				LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	Prob	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.1	1	0.488	200	0.36	0.02	0.36	0.25	-0.02	0.25	0.38	0.01	0.38	0.41	0.02	0.41	0.08	0.00	0.08
			400	0.31	0.01	0.31	0.19	-0.01	0.19	0.32	-0.01	0.32	0.38	0.00	0.38	0.06	0.00	0.06
			600	0.26	-0.03	0.26	0.15	0.00	0.15	0.32	-0.02	0.32	0.26	-0.03	0.26	0.04	0.00	0.04
1.09	$\frac{1}{25}$	0.481	200	0.09	0.01	0.09	0.10	-0.01	0.10	0.10	0.01	0.10	0.11	0.01	0.11	0.04	0.01	0.04
			400	0.07	0.01	0.07	0.05	0.00	0.05	0.08	-0.01	0.08	0.09	-0.01	0.09	0.03	0.01	0.03
			600	0.06	0.01	0.06	0.05	-0.02	0.05	0.06	-0.01	0.06	0.06	-0.01	0.06	0.03	0.01	0.03

Table 4: The top and bottom parts correspond to decay = 1 and decay = 2, respectively."r" and "Prob" represent the shrinkage parameter in (26) and case probability respectively. The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively. The columns under "LiVE", "Post Selection", "`hdi`", "`WLDP`" and "Lasso" correspond to the proposed estimator, the post model selection estimator, the plug-in `hdi`, the plug-in `WLDP` and the Plug-in Lasso estimator respectively.

$\Sigma = \{0.5^{1+|j-l|}\}_{1\leq j,l\leq 500}$ and generate x_* as

$$x_{*,j} = \begin{cases} x_{\text{basis},j} & \text{for } 1 \leq j \leq 11 \quad ; \quad j \neq 9, 10 \\ 10 & \text{for } j = 9, 10 \\ \frac{1}{25} \cdot x_{\text{basis},j} & \text{for } 12 \leq j \leq 501 \end{cases}$$

We construct the new β and x_* as we believe this is a challenging setting for post selection. The insignificant regression coefficients β_9, β_{10} make the corresponding covariates $X_{\cdot,9}$ and $X_{\cdot,10}$ unlikely to be selected by Lasso in the first step. However, with enlarged entries $x_{*,9}, x_{*,10}$, the corresponding covariates comprises a major part of the magnitude of the case probability $h(x_*^\intercal \beta)$, thereby leading to a large omitted variable bias when these relevant covariates are not selected by Lasso. We have observed in Table 5 that the post selection estimator has a large omitted variable bias and also produces a under-covered confidence interval.

B.4 Exact sparse with intercept

To explore the performance of the inference procedures in presence of an intercept, we set the values for the 501 regression coefficients as in Section 4.2 i.e., $\beta_j = (j-1)/20$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq 501$ and consider two values for β_1, $\beta_1 = -1$ and $\beta_1 = 1$ leading to two different target case probabilities 0.501 and 0.881 respectively. We investigate the finite sample performance of the inference methods for Loading 1.

Loading 3 (Case Probability $= 0.578$**)**														
			LiVE						Post Selection					
$\|x_{\text{new}}\|_2$	r	n	Cov	ERR	Len	RMSE	Bias	SE	Cov	ERR	Len	RMSE	Bias	SE
14.19	$\frac{1}{25}$	200	0.91	0.11	0.87	0.36	0.05	0.35	0.54	0.29	0.29	0.24	0.05	0.23
		400	0.95	0.08	0.85	0.29	0.01	0.29	0.73	0.25	0.32	0.21	0.07	0.20
		600	0.96	0.10	0.81	0.27	0.01	0.27	0.75	0.30	0.30	0.21	0.07	0.19

Table 5: Comparison of the proposed method and the post selection method for Loading 3 with regard to sample sizes $\{200, 400, 600\}$. "r" represents the shrinkage parameter in (26). The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs; the column indexed with "ERR" represent the empirical rejection rate of the testing procedure; the columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively. The columns under "LiVE" and "Post Selection" correspond to the proposed estimator and post model selection estimator respectively.

We report the simulation results based on 500 replications in Tables 6 and 7. Table 6 shows the proposed inference procedure continue to produce valid confidence intervals and the confidence intervals have shorter lengths for a larger sample size or a smaller ℓ_2 norm $\|x_*\|_2$. In comparison, `hdi` is under-coverage in general while the over-coverage issue of `WLDP` is still persistent. For $\beta_1 = -1$, the case probability represents an alternative in the indistinguishable region and hence the testing procedures do not have power in general while for $\beta_1 = 1$, the case probability is well above 0.5 and corresponds to an alternative to the null hypothesis, thereby the ERR, an empirical measure of power, increases with a larger sample size for all the testing procedures except for the one based on `WLDP`. It should be mentioned here that the comparison of our proposed method with `hdi` and `WLDP` in the setting with intercepts is unfair since `hdi` and `WLDP` are not designed to handle case probability and their output does not handle inference for the intercept. However, in practical applications, the intercept is an important term in capturing the case probabilities in logistic model.

Exact Sparse Regression Setting with Intercept														
$\beta_1 = -1$ **(Case Probability** $= 0.501$**)**														
			LiVE			Post Selection			`hdi`			`WLDP`		
$\|x_*\|_2$	r	n	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len
16.1	1	200	0.97	0.01	0.93	0.65	0.22	0.49	0.91	0.11	0.97	1.00	0.00	1.00
		400	0.96	0.02	0.85	0.61	0.20	0.41	0.90	0.13	0.89	1.00	0.00	1.00
		600	0.96	0.03	0.80	0.71	0.21	0.38	0.91	0.12	0.83	1.00	0.00	1.00
1.09	$\frac{1}{25}$	200	0.93	0.02	0.42	0.76	0.10	0.41	0.23	0.84	0.34	0.94	0.25	0.61
		400	0.97	0.01	0.30	0.88	0.09	0.33	0.15	0.98	0.21	0.90	0.31	0.54
		600	0.98	0.02	0.22	0.94	0.02	0.26	0.07	0.99	0.21	0.88	0.34	0.51
$\beta_1 = 1$ **(Case Probability** $= 0.881$**)**														
			LiVE			Post Selection			`hdi`			`WLDP`		
$\|x_*\|_2$	r	n	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len
16.1	1	200	0.97	0.07	0.91	0.65	0.80	0.28	0.91	0.07	0.98	1.00	0.00	1.00
		400	0.97	0.14	0.79	0.61	0.84	0.24	0.93	0.07	0.91	1.00	0.00	1.00
		600	0.96	0.18	0.74	0.71	0.88	0.20	0.91	0.06	0.86	1.00	0.00	1.00
1.09	$\frac{1}{25}$	200	0.93	0.97	0.24	0.65	0.99	0.16	0.51	0.79	0.43	1.00	0.17	0.63
		400	0.94	0.98	0.15	0.71	1.00	0.11	0.39	0.94	0.24	1.00	0.30	0.55
		600	0.91	1.00	0.12	0.85	1.00	0.11	0.21	0.96	0.20	1.00	0.32	0.52

Table 6: The top and bottom parts correspond to $\beta_1 = -1$ and $\beta_1 = 1$, respectively. "r" represents the shrinkage parameter in (26). The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs; the column indexed with "ERR" represent the empirical rejection rate of the testing procedure. The columns under "LiVE" ,"Post Selection", "`hdi`" and "`WLDP`" correspond to the proposed estimator, the post model selection estimator, the plug-in debiased estimator using `hdi` and `WLDP` respectively.

Exact Sparse Regression Setting with Intercept																	
$\beta_1 = -1$ **(Case Probability** $= 0.501$**)**																	
			LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.1	1	200	0.35	0.01	0.35	0.31	0.01	0.31	0.40	0.15	0.38	0.41	0.19	0.36	0.14	-0.09	0.12
		400	0.32	-0.01	0.32	0.27	-0.01	0.27	0.37	0.11	0.35	0.38	0.15	0.35	0.11	-0.07	0.09
		600	0.26	0.02	0.26	0.20	0.02	0.20	0.31	0.16	0.27	0.32	0.10	0.25	0.08	-0.05	0.07
1.09	$\frac{1}{25}$	200	0.12	-0.04	0.11	0.21	-0.04	0.21	0.27	0.26	0.08	0.31	0.31	0.07	0.13	-0.10	0.08
		400	0.08	-0.02	0.08	0.12	-0.02	0.12	0.23	0.23	0.05	0.30	0.30	0.04	0.09	-0.07	0.06
		600	0.05	-0.02	0.05	0.08	0.02	0.07	0.22	0.22	0.04	0.29	0.29	0.04	0.08	0.06	0.05
$\beta_1 = 1$ **(Case Probability** $= 0.881$**)**																	
			LiVE			Post Selection			`hdi`			`WLDP`			Lasso		
$\|x_*\|_2$	r	n	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
16.1	1	200	0.33	-0.15	0.30	0.18	-0.02	0.18	0.48	-0.28	0.38	0.45	-0.25	0.38	0.13	-0.10	0.08
		400	0.22	-0.07	0.21	0.17	-0.01	0.17	0.41	-0.20	0.33	0.35	-0.15	0.31	0.09	-0.07	0.05
		600	0.23	-0.06	0.21	0.10	-0.01	0.10	0.38	-0.20	0.30	0.36	-0.20	0.30	0.07	-0.06	0.04
1.09	$\frac{1}{25}$	200	0.06	-0.02	0.06	0.08	0.03	0.07	0.16	-0.14	0.08	0.12	-0.09	0.07	0.11	-0.10	0.06
		400	0.04	-0.01	0.04	0.06	0.04	0.05	0.16	-0.15	0.06	0.10	-0.09	0.05	0.08	-0.06	0.04
		600	0.03	-0.01	0.03	0.03	0.02	0.03	0.16	-0.16	0.05	0.11	-0.10	0.04	0.06	-0.05	0.03

Table 7: The top and bottom parts correspond to $\beta_1 = -1$ and $\beta_1 = 1$, respectively."r" represents the shrinkage parameter in (26). The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively. The columns under "LiVE", "Post Selection", "`hdi`", "`WLDP`" and "Lasso" correspond to the proposed estimator, the post model selection estimator, the plug-in `hdi`, the plug-in `WLDP` and the Plug-in Lasso estimator respectively.

C Additional Real Data Analysis

Figure 3 presented confidence intervals constructed using our method for the predicted probabilities shown in all six panels in Figure 1, corresponding to prediction of hypertension, resistant hypertension and high blood pressure with unexplained low blood potassium across two random subsamples.

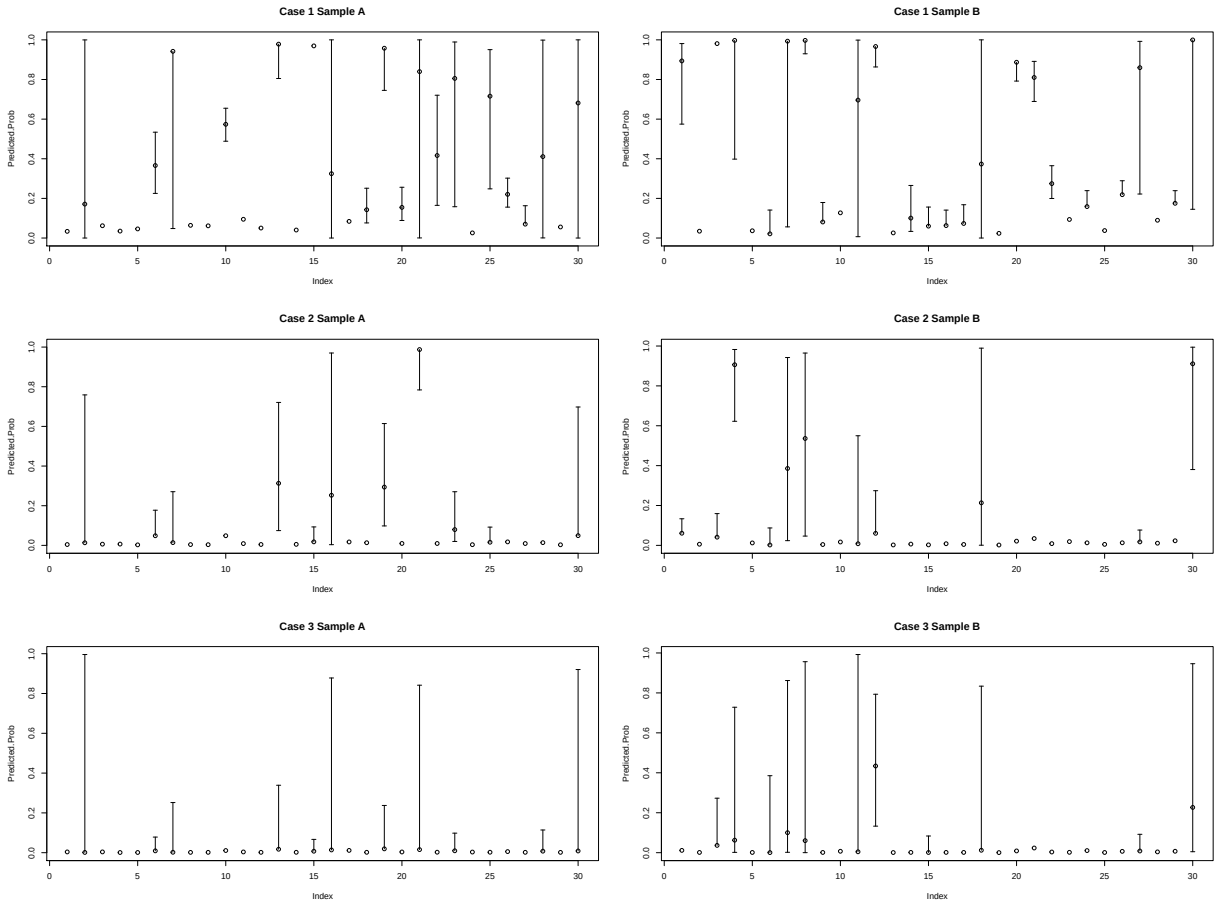

Figure 3: Confidence interval construction for the random subsamples