Journal of Machine Learning Research 22 (2021) 1-54 Submitted 9/20; Revised 10/21; Published 10/21

Inference for the Case Probability in High-dimensional Logistic Regression

Zijian Guo zijguo@stat.rutgers.edu
Prabrisha Rakshit pr412@scarletmail.rutgers.edu
Department of Statistics
Rutgers University
Piscataway, New Jersey, USA

Daniel S. Herman daniel.herman2@pennmedicine.upenn.edu
Department of Pathology and Laboratory Medicine
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Jinbo Chen jinboche@upenn.edu
Department of Pathology and Laboratory Medicine
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Editor: Jie Peng

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 to support documentation for medical billing, patients' data are frequently not represented with 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 (Parikh et al., 2019; Alhassan et al., 2020; Honga et al., 2019; Faisal et al., 2020). 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 (Castro et al., 2015).

One important example of 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 (Reincke et al., 2012; Lin et al., 2012; Catena et al., 2007). PA is thought, based on epidemiological studies, to affect up to 1% of US adults (Kayser et al., 2016; Hannemann and Wallaschofski, 2011), 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 (Funder et al., 2016). 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. 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 $\{y_i, X_{i\cdot}\}_{1\leq i\leq n}$ 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 $1 \leq i \leq n$,

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

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 throughout 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 functional $x_*^{\intercal}\beta$ 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 (Pearl, 2000; Rosenbaum and Rubin, 1983) and double robustness estimators (Bang and Robins, 2005; Kang and Schafer, 2007).

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 (Bunea, 2008; Bach, 2010; Bühlmann and van de Geer, 2011; Meier et al., 2008; Negahban et al., 2009; Huang and Zhang, 2012). The penalized estimators enjoy desirable estimation accuracy properties. However, these methods do not lend themselves directly to statistical inference for 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 to aid confidence interval construction for individual regression coefficients in high-dimensional linear regression models (van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014). This idea has also been extended to making inference for β_j for $1 \leq j \leq p$ in high-dimensional logistic regression models (van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2018). 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 method (van de Geer et al., 2014) requires an accurate estimator of the high-dimensional vector $[\mathbf{E}\widehat{H}(\beta)]^{-1}x_* \in \mathbb{R}^p$ where $\widehat{H}(\beta)$ denotes 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 since the Hessisan 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 by 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 an 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-corrected 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 establish 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 develop new technical tools to establish the asymptotic normality for the re-weighted estimator; see Section 3.3.

We conduct 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` (Dezeure et al., 2015), the plug-in `WLDP` (Ma et al., 2018) and generalization of the transformation method (Zhu and Bradic, 2018; Tripuraneni and Mackey, 2020) to logistic models. We demonstrate 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-fold.

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 statistically 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.

Our proposed LiVE estimator has been implemented in the R package `SIHR`, which is available from CRAN. More detailed illustration of the R package `SIHR` can be found in Rakshit et al. (2021).

1.3 Comparison with Existing Literature

We have proposed a two-step bias correction procedure to make inference for $h(x_*^\intercal \beta)$ in the high-dimensional logistic model. Specifically, the linearization and variance enhancement techniques are introduced to ensure that our proposed confidence intervals are valid for any $x_* \in \mathbb{R}^p$ and a broad class of design covariance matrix. We shall mention other related works and discuss the connections and differences.

Post-selection inference (Belloni and Chernozhukov, 2013, e.g.) 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 Cai and Guo (2017); Athey et al. (2018); Zhu and Bradic

(2018); Cai et al. (2019). However, these methods cannot be directly applied to make inference for the case probability in the logistic model. Our proposed linearization technique is useful in generalizing the inference methods for linear models to logistic models. 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 condition 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 (van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2018). More detailed comparisons are provided in Section 2.5.

In Section 4, we provide detailed numerical comparisons to the inference methods by van de Geer et al. (2014); Ma et al. (2018); Zhu and Bradic (2018); Tripuraneni and Mackey (2020).

Belloni et al. (2014); Farrell (2015); Chernozhukov et al. (2018) studied inference for treatment effects in high-dimensional regression models while the current paper focuses on inference for a different quantity, the case probability. Sur and Candès (2019); Sur et al. (2019) 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) (Fox, 2015), which is the standard technique used to maximize the likelihood of the logistic model. 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 $J \subset [p]$ and 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 $\|x\|_q = (\sum_{i=1}^{p} |x_i|^q)^{\frac{1}{q}}$ for $q > 0$ with $\|x\|_0$ denoting the cardinality of the support of x and $\|x\|_\infty = \max_{1\le j\le p} |x_j|$. 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 $\max_{1\le i\le n, 1\le j\le p} |X_{i,j}|$. 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 $a_n \le Cb_n$ 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 for the case probability and conduct hypothesis testing related to outcome labelling. In Section 2.5, we compare with the existing estimators (van de Geer et al., 2014; Ma et al., 2018; Zhu and Bradic, 2018; Tripuraneni and Mackey, 2020; Bickel, 1975).

2.1 Challenges Underlying 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 $\ell(\beta) = \sum_{i=1}^{n} \left[\log\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right) - y_i \cdot (X_{i\cdot}^{\intercal}\beta)\right]$. The penalized log-likelihood estimator $\widehat{\beta}$ is defined as (Bühlmann and van de Geer, 2011),

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

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 C.6 in the supplement for the numerical illustration.

Our proposed method is built on the idea of correcting the bias of the plug-in estimator $x_*^{\intercal}\widehat{\beta}$ and then applying 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-correction idea in high-dimensional linear and logistic models (van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014). 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 used for correcting the bias of $\widehat{\beta}_j$. Define the error $\epsilon_i = y_i - h(X_{i\cdot}^{\intercal}\beta)$ for $1 \leq i \leq n$. We apply the Taylor expansion of the h function and obtain

$$y_i - h(X_{i\cdot}^{\intercal}\widehat{\beta}) = h(X_{i\cdot}^{\intercal}\beta) - h(X_{i\cdot}^{\intercal}\widehat{\beta}) + \epsilon_i = h'(X_{i\cdot}^{\intercal}\widehat{\beta})X_{i\cdot}^{\intercal}(\beta-\widehat{\beta}) + R_i + \epsilon_i$$

with the approximation error $R_i = \int_0^1 (1-t)h''(X_{i\cdot}^{\intercal}\widehat{\beta} + tX_{i\cdot}^{\intercal}(\beta-\widehat{\beta}))dt \cdot (X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta))^2$. Since $h'(x) = h(x)(1-h(x))$ for any $x \in \mathbb{R}$, we simplify the above expression as

$$y_i - h(X_{i\cdot}^{\intercal}\widehat{\beta}) = h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))[X_{i\cdot}^{\intercal}(\beta-\widehat{\beta}) + \Delta_i] + \epsilon_i \quad \text{with} \quad \Delta_i = R_i/h'(X_{i\cdot}^{\intercal}\widehat{\beta}). \tag{5}$$

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 negative log-likelihood $\ell(\beta)$.

To conduct the bias-correction, van de Geer et al. (2014) construct the projection direction $\widehat{u} \in \mathbb{R}^p$ in (4) such that $\widehat{H}(\widehat{\beta})\widehat{u} \approx e_j$ and hence

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

Such an approximation has been shown to be accurate by assuming a sparse $[\mathbf{E}\widehat{H}(\beta)]^{-1}e_j$ (van de Geer et al., 2014). However, $[\mathbf{E}\widehat{H}(\beta)]^{-1}x_*$ can be an arbitrarily dense vector and hence it is challenging to accurately estimate $[\mathbf{E}\widehat{H}(\beta)]^{-1}x_*$ and generalize the bias-correction procedure in van de Geer et al. (2014). Specifically, $[\mathbf{E}\widehat{H}(\beta)]^{-1}$ may be dense for the following two reasons: (1) the columns of $[\mathbf{E}\widehat{H}(\beta)]^{-1}$ are dense; (2) x_* is a dense vector.

In the following two sections, we develop new techniques, which can effectively correct the bias for an arbitrary loading $x_* \in \mathbb{R}^p$ 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,

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

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 invert $\widehat{\Sigma}$, 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 to 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 challenges in the theoretical justification of the proposed method. The corresponding technical challenge will be addressed in Section 3.3 with 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

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

and notice that all three terms depend on our constructed projection direction $\widehat{u} \in \mathbb{R}^p$.

Motivated by the decomposition in (8), we construct $\widehat{u} \in \mathbb{R}^p$ as the solution to the following optimization problem,

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

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

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

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 construction of $\widehat{u}$ in (9) to (11) by connecting it to the error decomposition (8). The objective function in (9) 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), ensure 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 (Javanmard and Montanari, 2014; Zhang and Zhang, 2014) and is shown to be effective when $x_* = e_j$ (Javanmard and Montanari, 2014; Zhang and Zhang, 2014), a sparse x_* (Cai and Guo, 2017) and x_* with a bounded ℓ_2 norm (Athey et al., 2018). 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). Cai et al. (2019) has presented examples where such bias correction method will fail; see Proposition 2 of Cai et al. (2019).

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). 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 rely on $\widehat{u}$ to be an accurate estimator of u^*. This explains why the proposed bias-corrected estimator 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 $x_*^\intercal \beta$ (Cai et al., 2019). However, the method developed for linear models in Cai et al. (2019) 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 $\|Xu\|_\infty \leq \|x_*\|_2 \tau_n$ was first introduced in Javanmard and Montanari (2014) to establish the asymptotic normality for the non-Gaussian error in the linear model. In our analysis, this additional constraint is not just introduced to establish the asymptotic normality for the non-Gaussian error ϵ_i, but also facilitates the empirical process proof. The range of values for τ_n is also different from that in Javanmard and Montanari (2014), where equation (54) of Javanmard and Montanari (2014) has $\|x_*\|_2 = 1$ and $\tau_n \asymp n^{\delta_0}$ with $1/4 < \delta_0 < 1/2$ while τ_n in our paper 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 rest of the 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 $\widehat{u}$ defined 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 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).

2.5 Comparison to Other Estimators

In the following, we discuss the difference between the proposed LiVE method and related methods. A detailed numerical comparison with the methods in van de Geer et al. (2014); Ma et al. (2018); Zhu and Bradic (2018); Tripuraneni and Mackey (2020) is provided in Section 4.

The main distinction is that the existing literature focused on single regression coefficients, instead of the case probability. We shall use $\widetilde{\beta}_j$ to denote the existing coordinate-wise debiased estimator of β_j for $1 \leq j \leq p$ (van de Geer et al., 2014; Ma et al., 2018). The computation cost of the plug-in estimator $x_*^\intercal \widetilde{\beta}$ is much higher than our proposed method, as the proposed method targets at $x_*^\intercal \beta$ directly and requires construction of one projection direction as in (9) to (11). In contrast, the plug-in debiased estimator $x_*^\intercal \widetilde{\beta}$ (van de Geer et al., 2014; Ma et al., 2018) requires construction of p projection directions. See Tables 1 and 2 for a detailed comparison of computation times. We also mention that there exist technical difficulties of controlling the bias of the plug-in debiased estimator; see Section A.1 in the supplement.

The re-weighting idea has been proposed in Ma et al. (2018) for inference for the single regression coefficient β_j in the high-dimensional logistic model. However, it is not straightforward to extend it to make inference for a general linear combination of regression coefficients. A brief summary of the method can be found in Section A.2 in the supplement. Moreover, the analysis in Ma et al. (2018) requires sample splitting, 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. But our empirical process results in Section 3.3 can carry out the analysis for an arbitrary combination of the regression vector and bypass the sample splitting related to the re-weighting step.

The transformation methods have been proposed in Zhu and Bradic (2018); Tripuraneni and Mackey (2020) for high-dimensional linear models. We now extend this method to the logistic model. With $H_* = x_* x_*^{\intercal} / \|x_*\|_2^2$, we have

$$X_{i\cdot}^{\intercal}\beta = X_{i\cdot}^{\intercal} H_* \beta + X_{i\cdot}^{\intercal}(\mathrm{I} - H_*)\beta = \frac{X_{i\cdot}^{\intercal} x_*}{\|x_*\|_2^2}(x_*^{\intercal}\beta) + (X_{i\cdot}^{\intercal} U)(U^{\intercal}\beta)$$

where $\mathrm{I} - H_* = UU^{\intercal}$ and $U^{\intercal}U = \mathrm{I}_{p-1}$. We construct $U \in \mathbb{R}^{p\times(p-1)}$ with its columns denoting the eigenvectors of $\mathrm{I} - H_*$ corresponding to its non-zero eigenvalues. This defines an equivalent logistic regression model

$$\mathbb{P}(y_i = 1 | \check{X}_{i\cdot}) = h(\check{X}_{i\cdot}^{\intercal}\eta) \quad i = 1, 2, \cdots, n \tag{15}$$

with $\eta_1 = x_*^{\intercal}\beta$, $\eta_{-1} = U^{\intercal}\beta$ and $\check{X}_{i,1} = X_{i,\cdot}^{\intercal} x_* / \|x_*\|_2^2$, $\check{X}_{i,-1} = X_{i,\cdot}^{\intercal} U$. We apply our proposed method to the newly defined logistic regression model and obtain the bias-corrected estimator $\widehat{\eta}_1$ and its variance estimator $\widehat{\mathrm{V}}_U$. The CI and the testing procedure related to $x_*^{\intercal}\beta$ are given by

$$\mathrm{CI}_\alpha(x_*) = \left[h(\widehat{\eta}_1 - z_{\alpha/2}\widehat{\mathrm{V}}_U^{1/2}), h(\widehat{\eta}_1 + z_{\alpha/2}\widehat{\mathrm{V}}_U^{1/2})\right], \ \phi_\alpha(x_*) = \mathbf{1}\left(\widehat{\eta}_1 - z_\alpha \widehat{\mathrm{V}}_U^{1/2} \geq 0\right). \tag{16}$$

In Section 4.3, we provide a comparison between (16) and our proposed method. We observe that the transformation method suffers from a larger bias and does not achieve the desired coverage when x_* is relatively dense; see Table 3 for details.

Furthermore, it can be challenging to analyze the transformed model (15) since η is not necessarily sparse (even if β is sparse). To guarantee a sparse η, certain special structures (e.g. sparsity) need to be imposed on the loading x_*. More detailed discussions on Tripuraneni and Mackey (2020) are provided in Section A.3 in the supplement.

Our proposed debiased estimator is closely related to the one-step estimator in Bickel (1975). Our bias correction step corresponds to the following estimating equation

$$\mathbf{E}\psi(y_i, X_{i\cdot}, \beta) = 0 \quad \text{with} \quad \psi(y_i, X_{i\cdot}, \beta) = \frac{X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\beta))}{h(X_{i\cdot}^{\intercal}\beta)(1 - h(X_{i\cdot}^{\intercal}\beta))}, \tag{17}$$

which is a weighted version of the estimating equation $\mathbf{E}X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\beta)) = 0$. We shall point out that the weight in (17) depends on both the data and the unknown parameter β.

In comparison to Bickel (1975), we propose our bias-correction step by taking a Taylor expansion of $X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\beta))$ instead of $\psi(y_i, X_{i\cdot}, \beta)$; see equation (5). Note that the derivative of $\psi(y_i, X_{i\cdot}, \beta)$ with respect to β has a complicated form due to the fact that the weight in (17) also involves β. Hence, it is not straightforward to express our weighted bias-corrected as a one-step estimator since the weight in the estimating equation (17) also depends on β.

3. Theoretical Justification

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 k non-zero entries.

(A2) With probability larger than $1 - p^{-c}$, $\min\{h(X_{i\cdot}^{\intercal}\beta), 1 - h(X_{i\cdot}^{\intercal}\beta)\} \geq c_{\min}$ for $1 \leq i \leq n$ and some small positive constant $c_{\min} \in (0, 1)$.

Condition (A1) imposes the tail condition for the high-dimensional covariates $X_{i\cdot}$ and assumes that the population second order moment matrix is invertible. Condition (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 (Athey et al., 2018; van de Geer et al., 2014; Ma et al., 2018; Ning and Liu, 2017). For example, see condition (iv) of Theorem 3.3 of van de Geer et al. (2014) and the overlap assumption (Assumption 6) in Athey et al. (2018). As a remark, Condition (A2) may be stringent for certain applications; we test the robustness of our proposed method to the violation of (A2) in Section 4.6.

The following proposition states the theoretical properties of the penalized estimator $\widehat{\beta}$ in (3), which have been established in Negahban et al. (2009); Huang and Zhang (2012).

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

$$\|\widehat{\beta}_{S^c} - \beta_{S^c}\|_1 \leq (2/\delta_0 + 1)\|\widehat{\beta}_S - \beta_S\|_1 \quad \textit{and} \quad \|\widehat{\beta} - \beta\|_1 \leq Ck\lambda_0 \tag{18}$$

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 $c > 0$,

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

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

The asymptotic normality established in next subsection will hold for any initial estimator satisfying condition (B), including our initial estimator defined in (3).

3.2 Asymptotic Normality and Statistical Inference

We now establish the limiting distribution for the proposed point estimator $\widehat{x_*^\intercal\beta}$.

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* $0 < \alpha < 1$,

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

where

$$\mathrm{V} = \widehat{u}^\intercal\left[\frac{1}{n^2}\sum_{i=1}^{n}[h(X_{i\cdot}^\intercal\beta)(1-h(X_{i\cdot}^\intercal\beta))]^{-1}X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u}. \tag{19}$$

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

$$c_0\|x_*\|_2/n^{1/2} \leq \mathrm{V}^{1/2} \leq C_0\|x_*\|_2/n^{1/2}, \tag{20}$$

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

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* (19) *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 for 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 (20), which guarantees that the variance component of (8) dominates the remaining bias.

Secondly, to establish the asymptotic normality result, we do not impose any sparsity condition on the precision matrix Σ^{-1}. This has weakened 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} (van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2018). Thirdly, with $\tau_n \asymp (\log n)^{1/2}$, the required sparsity condition on β is $k \ll n^{1/2}/[\log p\,(\log n)^{1/2}]$. Such sparsity conditions are imposed for confidence interval construction for both high-dimensional linear models and logistic models (Javanmard and Montanari, 2014; Zhang and Zhang, 2014; van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2018). Regarding confidence interval construction for β_j in high-dimensional linear models, Cai and Guo (2017) have shown that the ultra-sparse condition $k \ll n^{1/2}/\log p$ is necessary and sufficient for constructing a confidence interval of length $1/\sqrt{n}$. Recently, Cai et al. (2021) extended this result to inference for single regression coefficients in the high-dimensional logistic regression.

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*

$$\limsup_{n\to\infty} \left|\mathbb{P}_\theta\left[\phi_\alpha(x_*) = 1\right] - \left[1 - \Phi^{-1}(z_\alpha - \mu/(n\mathrm{V})^{1/2})\right]\right| = 0, \tag{21}$$

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. The asymptotic power expression in (21) holds for any μ. Since (20) implies $c_0\|x_*\|_2 \leq (n\mathrm{V})^{1/2} \leq C_0\|x_*\|_2$, the power of the proposed test in (21) 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 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 $\|x_*\|_2$; see Tables 1 and 2 for details.

3.3 Analysis Related to Reweighting in Linearization

In the following, we provide more insights on how to establish the asymptotic normality and summarize technical tools for analyzing the re-weighted estimator obtained by 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), *with probability larger than $1-p^{-c}-\exp(-cn)$ for some positive constant $c > 0$,*

$$n^{1/2}\left|(\widehat{\Sigma}\widehat{u} - x_*)^\intercal(\widehat{\beta} - \beta)\right| \leq n^{1/2}\|x_*\|_2\lambda_n\|\widehat{\beta} - \beta\|_1 \lesssim \|x_*\|_2 k \log p \cdot n^{-1/2}, \tag{22}$$

and

$$n^{1/2}|\widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}\Delta_i| \leq \tau_n\|x_*\|_2 k \log p \cdot n^{-1/2} \tag{23}$$

Together with (8), it remains to establish the asymptotic normality of the following term,

$$\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}\epsilon_i. \tag{24}$$

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 (24).

We decouple the correlation between $\widehat{\beta}$ and ϵ_i through the following expression,

$$\begin{aligned}\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}\epsilon_i &= \widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\beta)(1-h(X_{i\cdot}^{\intercal}\beta))]^{-1}X_{i\cdot}\epsilon_i \\ &+ \widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\left([h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}-[h(X_{i\cdot}^{\intercal}\beta)(1-h(X_{i\cdot}^{\intercal}\beta))]^{-1}\right)X_{i\cdot}\epsilon_i.\end{aligned} \tag{25}$$

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 (25) captures the error incurred on estimating β by $\widehat{\beta}$. We now provide a sharp control of this error term by suitable empirical process theory.

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

$$\left|\widehat{u}^{\intercal}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left([h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}-[h(X_{i\cdot}^{\intercal}\beta)(1-h(X_{i\cdot}^{\intercal}\beta))]^{-1}\right)X_{i\cdot}\epsilon_i\right| \leq Ct_0\tau_n\|x_*\|_2\frac{k\log p}{n^{1/2}} \tag{26}$$

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 in 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 on contraction principles for i.i.d Rademacher random variables (Koltchinskii, 2011). This lemma and the related analysis are particularly useful for carefully characterizing the approximation error in (26) 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.

Remark 4 *In comparison, in case of the linear model or the logistic model without re-weighting (van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014), such a challenge does not exist since the corresponding 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 van de Geer et al. (2014); Javanmard and Montanari (2014); Zhang and Zhang (2014), is either directly independent of ϵ_i or can be replaced by $u^*=\Sigma^{-1}x_*$ (by assuming $\Sigma^{-1}x_*$ to be sparse).*

4. Numerical Studies

4.1 Algorithm Implementation and Method Comparision

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` (Friedman et al., 2010) 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,

$$\widehat{u} = \arg\min_{u \in \mathbf{R}^p} u^{\intercal}\widehat{\Sigma}u \quad \text{subject to } \|\widehat{\Sigma}u - x_*\|_\infty \leq \|x_*\|_2\lambda_n, \quad |x_*^{\intercal}\widehat{\Sigma}u - \|x_*\|_2^2| \leq \|x_*\|_2^2\lambda_n. \tag{27}$$

This construction does not include the constraint (11), which is mainly imposed to facilitating the theoretical proof. We have conducted an additional check in simulations and observed that our constructed $\widehat{u}$ in (27) satisfies $\|X\widehat{u}\|_\infty \leq C\sqrt{\log n}\|x_*\|_2$; see Section C.2 in the supplementary material for details.

We solve the dual problem of (27),

$$\widehat{v} = \arg\min_{v \in \mathbb{R}^{p+1}} \frac{1}{4}v^{\intercal}H^{\intercal}\widehat{\Sigma}Hv + b^{\intercal}Hv + \lambda_n\|v\|_1 \text{ with } H = [b, \mathbb{I}_{p\times p}]\,, b = \frac{1}{\|x_*\|_2}x_* \tag{28}$$

and then solve the primal problem (27) as $\widehat{u} = -\left(\widehat{v}_{-1} + \widehat{v}_1 b\right)/2$. We refer to Proposition 2 in Cai et al. (2019) for the the detailed derivation of the dual problem (28). In this dual problem, when $\widehat{\Sigma}$ is singular and the tuning parameter $\lambda_n > 0$ gets sufficiently close to 0, the dual problem cannot be solved as the minimum value converges to negative infinity. Hence, we choose the smallest $\lambda_n > 0$ such that the dual problem has a finite minimum value. The tuning parameter λ_n selected in this manner is at the scale of $\sqrt{\log p/n}$. We investigate the ratio $\lambda_n/\sqrt{\log p/n}$ in Section C.1 in the supplement.

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

- **Plug-in Lasso.** Estimate $x_*^{\intercal}\beta$ by $x_*^{\intercal}\widehat{\beta}$ with $\widehat{\beta}$ denoting the penalized 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 $\widehat{\beta}_{\rm PL} \in \mathbf{R}^p$ is used to estimate $x_*^{\intercal}\beta$ by $x_*^{\intercal}\widehat{\beta}_{\rm PL}$. The variance of this post-selection estimator $x_*^{\intercal}\widehat{\beta}_{\rm PL}$ can be obtained by the inference results in the classical low-dimensional logistic regression, denoted by $\widehat{\rm V}_{\rm PL}$.
- **Plug-in `hdi` (Dezeure et al., 2015).** The R package `hdi` is implemented to obtain the coordinate debiased Lasso estimator $\widehat{\beta}_{\tt hdi} \in \mathbf{R}^p$ and the plug-in estimator $x_*^{\intercal}\widehat{\beta}_{\tt hdi}$ is used to estimate $x_*^{\intercal}\beta$, with the variance estimator as $\widehat{\rm V}_{\tt hdi}$.
- **Plug-in `WLDP` (Ma et al., 2018).** We compute the debiased lasso estimator $\widehat{\beta}_{\tt WLDP} \in \mathbf{R}^p$ by the Weighted LDP algorithm in Table 1 of Ma et al. (2018). The plug-in estimator of $x_*^{\intercal}\beta$ and the associated variance are given by $x_*^{\intercal}\widehat{\beta}_{\tt WLDP}$ and $\widehat{\rm V}_{\tt WLDP}$ respectively.
- **Generalization of Transformation Method (Zhu and Bradic, 2018) in** (16)**.**

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`, `WLDP` and the transformation method, from the perspectives of CI construction and hypothesis testing (2). Recall that our 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 $(\widetilde{\beta}, \widetilde{\mathrm{V}})$ by $(\widehat{\beta}_{\mathrm{PL}}, \widehat{\mathrm{V}}_{\mathrm{PL}})$, $(\widehat{\beta}_{\texttt{hdi}}, \widehat{\mathrm{V}}_{\texttt{hdi}})$ and $(\widehat{\beta}_{\texttt{WLDP}}, \widehat{\mathrm{V}}_{\texttt{WLDP}})$ respectively.

Throughout the simulation, we set $X_{i,1} = 1$ to represent the intercept and generate the covariates $\{X_{i,-1}\}_{1\leq i\leq n}$ from the multivariate normal distribution with zero mean and covariance matrix Σ. Conditioning on $X_{i\cdot}$, the binary outcome is generated by $y_i \sim \mathrm{Bernoulli}\left(h\left(X_{i\cdot}^\intercal\beta\right)\right)$, for $1 \leq i \leq n$. We generate the following loadings x_*.

- **Loading 1**: We set $x_{\mathrm{basis},1} = 1$ and generate $x_{\mathrm{basis},-1} \in \mathbb{R}^{(p-1)}$ following $N(0, \widetilde{\Sigma})$ with $\widetilde{\Sigma} = \{q \cdot 0.5^{1+|j-l|}\}_{1\leq j,l\leq(p-1)}$ for some $q > 0$; for $r \geq 0$, generate x_* as

$$x_{*,j} = \begin{cases} x_{\mathrm{basis},j} & \text{for } 1 \leq j \leq 11 \\ r \cdot x_{\mathrm{basis},j} & \text{for } 12 \leq j \leq p \end{cases}. \tag{29}$$

- **Loading 2**: $x_{\mathrm{basis},1}$ is set as 1 and $x_{\mathrm{basis},-1} \in \mathbb{R}^{(p-1)}$ is generated as following $N(0, \widetilde{\Sigma})$ with $\widetilde{\Sigma} = \{q \cdot (-0.75)^{|j-l|}/2\}_{1\leq j,l\leq(p-1)}$ for some $q > 0$; generate x_* using (29).

All simulation results are averaged over 500 replications. The loadings are only generated once and kept the same across all 500 replications.

4.2 Varying sample size n and loading norm $\|x_*\|_2$

We investigate the performance of our method across different sample sizes n and loading norms $\|x_*\|_2$. We set $p = 501$, $\Sigma = \{0.5^{1+|j-l|}\}_{1\leq j\leq l\leq(p-1)}$ and vary $n \in \{200, 400, 600\}$. We carry out the simulations for both Loading 1 and Loading 2 with $q = 1$ and $r \in \{1, 1/25\}$. We generate the exactly sparse regression vector β as

(S1) $\beta_1 = 0$, $\beta_j = (j-1)/20$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq p$.

Here, the scale parameter r in (29) 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 $x_*^\intercal\beta$. 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. In Section C.4 in the supplement, we conduct additional simulation studies for models with a non-zero intercept.

In Table 1, we compare the proposed LiVE method with post-selection, `hdi` and `WLDP`, in terms of CI construction and hypothesis testing for the setting (S1) with x_* generated as Loading 1. 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` overcovers and the post-selection method undercovers.

Regarding the testing procedure, we report the empirical rejection rate (ERR), which is defined as the proportion of null hypothesis in (2) being rejected out of the 500 replications.

Setting (S1), Loading 1 with $q = 1$																			
				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	0.732	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	$\frac{1}{25}$	0.732	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

Table 1: **Varying n and $\|x_*\|_2$.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; "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.

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 a 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.

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 algorithms for bias-correction.

The inference results for Loading 2 are similar and reported in Table C.5 in the supplement. We report Root Mean Squared Error (RMSE), bias and standard deviation of the proposed LiVE estimator, plug-in Lasso, post-selection, `hdi` and `WLDP` in Table C.6 in the supplementary material. It is observed that the plug-in Lasso estimator cannot be used for confidence interval construction as its bias component is a dominant component of the RMSE 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. The post-selection method can select either a larger model or a smaller model compared to the true one. In Table 1, post selection undercovers since post-selection tends to select a relatively large set of variables and this results in a perfect separation in the re-fitting step. In Section C.3 in the supplementary material, we show another setting where the post-selection method selects a smaller model and leads to a substantial omitted variable bias.

In practical settings, the regression vector β might not be exactly 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 following generation of β:

(S2) $\beta_1 = 0$ and $\beta_j = (j-1)^{-\text{decay}}$ for $2 \leq j \leq p$, with decay $\in \{1, 2\}$.

We illustrate the method comparison using Loading 1. The inference results are reported in Table 2 for decay = 1. The results for decay = 2 are similar to those for decay = 1 and summarized in Table C.7 in the supplement. The estimation results are reported in Table C.8 in the supplement.

Setting (S2) with decay=1, Loading 1 with $q=1$																			
				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

Table 2: **Varying n and $\|x_*\|_2$.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; "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.

Note that as the regression coefficient is decaying, the shrinking parameter r in (29) plays a role in determining the case probability. The main observations are consistent with those in Table 1: only the proposed LiVE method and `hdi` have proper coverage across different scenarios while the CI by post selection undercovers and the CI by `WLDP` overcovers. 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 3200 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.

4.3 Comparison with the Transformation Method

We compare LiVE with the Transformation method (Zhu and Bradic, 2018; Tripuraneni and Mackey, 2020) based on settings (S3) and (S4), which are variations of setting (S1).

(S3) $p = 501$; $\beta_1 = 0$, $\beta_j = (j-1)/10$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq 501$.

(S4) $p = 1001$; $\beta_1 = 0$; $\beta_j = (j-1)/20$ for $2 \le j \le 11$ but $j \ne 3, 4, 6$; $\beta_j = 1$ for $j = 3, 4, 6$ and $\beta_j = 0$ for $12 \le j \le 1001$.

Set $\Sigma = \{0.5^{1+|j-l|}\}_{1 \le j \le l \le (p-1)}$. Table 3 summarizes the results for (S3) and (S4) with Loading 1 with $q = 1$ and $r \in \{1, 1/2, 1/5, 1/25\}$. We vary $n \in \{200, 400, 600\}$.

Setting (S3), **Loading 1 with** $q = 1$															
				LiVE						Transformation Method					
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE	Cov	ERR	Len	RMSE	Bias	SE
16.1	1	0.881	200	0.99	0.08	0.92	0.28	-0.12	0.25	0.84	0.06	0.82	0.34	-0.26	0.22
			400	1.00	0.12	0.87	0.23	-0.08	0.22	0.98	0.13	0.82	0.28	-0.18	0.22
			600	0.98	0.17	0.81	0.19	-0.05	0.18	0.96	0.14	0.82	0.26	-0.16	0.21
8.18	$\frac{1}{2}$	0.881	200	0.99	0.22	0.72	0.16	-0.06	0.14	0.65	0.11	0.63	0.27	-0.24	0.14
			400	0.98	0.38	0.62	0.13	-0.04	0.12	0.90	0.20	0.64	0.23	-0.17	0.15
			600	0.97	0.58	0.51	0.10	-0.03	0.10	0.92	0.28	0.60	0.20	-0.14	0.14
3.66	$\frac{1}{5}$	0.881	200	0.94	0.78	0.40	0.10	-0.06	0.09	0.60	0.54	0.41	0.19	-0.15	0.12
			400	0.97	0.95	0.32	0.07	-0.03	0.06	0.74	0.85	0.32	0.12	-0.08	0.09
			600	0.95	0.98	0.25	0.06	-0.02	0.06	0.84	0.94	0.26	0.09	-0.06	0.07
1.90	$\frac{1}{25}$	0.881	200	0.88	0.96	0.26	0.09	-0.05	0.07	0.92	0.99	0.20	0.06	-0.03	0.06
			400	0.92	0.99	0.21	0.06	-0.03	0.05	0.94	0.99	0.16	0.04	-0.01	0.04
			600	0.92	0.99	0.17	0.05	-0.02	0.04	0.96	1.00	0.14	0.03	0.00	0.03

Setting (S4), **Loading 1 with** $q = 1$															
				LiVE						Transformation Method					
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE	Cov	ERR	Len	RMSE	Bias	SE
22.2	1	0.814	200	0.99	0.02	0.97	0.33	-0.11	0.31	0.00	0.35	0.04	0.31	-0.30	0.08
			400	0.99	0.04	0.94	0.30	-0.09	0.29	0.04	0.10	0.31	0.28	-0.27	0.09
			600	0.98	0.08	0.91	0.26	-0.06	0.25	0.87	0.08	0.69	0.26	-0.19	0.18
11.2	$\frac{1}{2}$	0.814	200	0.99	0.10	0.85	0.20	-0.05	0.20	0.06	0.17	0.29	0.27	-0.25	0.10
			400	0.99	0.17	0.75	0.17	-0.03	0.17	0.77	0.12	0.68	0.21	-0.13	0.16
			600	0.99	0.22	0.71	0.15	-0.03	0.15	0.87	0.20	0.64	0.19	-0.11	0.15
4.90	$\frac{1}{5}$	0.814	200	0.96	0.50	0.52	0.10	-0.02	0.10	0.89	0.50	0.42	0.14	-0.08	0.11
			400	0.97	0.75	0.41	0.08	-0.01	0.08	0.90	0.73	0.35	0.11	-0.05	0.10
			600	0.97	0.76	0.39	0.08	-0.02	0.08	0.91	0.85	0.30	0.09	-0.03	0.09
2.29	$\frac{1}{25}$	0.814	200	0.96	0.91	0.30	0.08	-0.02	0.07	0.89	0.98	0.22	0.07	0.02	0.07
			400	0.94	0.98	0.24	0.06	-0.02	0.06	0.88	0.99	0.19	0.06	0.02	0.06
			600	0.93	0.98	0.23	0.06	-0.02	0.06	0.90	1.00	0.17	0.05	0.02	0.04

Table 3: **Comparison with Transformation Method.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively. The columns under "LiVE" and "Transformation Method" correspond to LiVE and the transformation method, respectively.

We observe that the performance of our proposed LiVE estimator and the transformation method are similar for a sparse loading (e.g. $r = 1/25$) while the performance of these two methods can be quite different for a dense loading (e.g. $r = 1, 1/2, 1/5$). Specifically, for a dense loading, the bias of the transformation method is typically larger than that of our proposed LiVE estimator; our proposed confidence intervals in general have coverage while the transformation method undercovers. For (S3), when $r = 1/5$, our proposed confidence intervals have better coverage than those by the transformation method while their lengths are similar; with a relatively dense loading (e.g. $r = 1$ or $r = 1/2$) and a smaller sample size ($n = 200$), the transformation method undercovers while the confidence intervals by the LiVE method provide coverage at the expense of longer lengths. For (S4), the observation is similar to that for (S3).

4.4 Increasing dimension p and coefficient magnitudes

We vary p across $\{1001, 2001, 5001\}$ and generate β as a mixture of large and small signals.

(S5) $\beta_1 = 0$; $\beta_j = (j-1)/20$ for $7 \leq j \leq 11$; $\beta_j = 1$ for $j = 2, 3, 4$; $\beta_j = -1$ for $j = 5, 6$ and $\beta_j = 0$ for $12 \leq j \leq p$

(S6) $\beta_1 = 0$; $\beta_j = (j-1)^{-\text{decay}}$ for $7 \leq j \leq p$; $\beta_j = 1$ for $j = 2, 3, 4$ and $\beta_j = -1$ for $j = 5, 6$ with decay $\in \{1, 2\}$

Set $\Sigma = \{0.5^{1+|j-l|}/2\}_{1 \leq j \leq l \leq (p-1)}$. Settings (S5) and (S6) are variations of Settings (S1) and (S2), respectively. The results for Setting (S5) with respect to Loading 1 with $q = 1/2$ and $r = 1/5$ are summarized in Table 4. We vary n across $\{400, 600, 1000\}$. In Section C.6 in the supplement, we report the results for Setting (S5) with respect to Loading 2 with $q = 1/2$ and $r = 1/5$ in Table C.9 and the results for (S6) in Tables C.10 and C.11. The results are similar to that reported in Table 4.

Setting (S5)**, Loading 1 with** $q = 1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
1001	3.21	$\frac{1}{5}$	0.263	400	0.91	0.00	0.42	0.13	0.07	0.11
				600	0.94	0.00	0.38	0.11	0.04	0.10
				1000	0.93	0.00	0.30	0.08	0.04	0.07
2001	4.60	$\frac{1}{5}$	0.531	400	0.96	0.12	0.55	0.15	0.07	0.14
				600	0.95	0.12	0.49	0.13	0.05	0.12
				1000	0.97	0.15	0.42	0.11	0.03	0.11
5001	7.07	$\frac{1}{5}$	0.385	400	0.97	0.00	0.72	0.19	0.02	0.19
				600	0.98	0.00	0.65	0.15	0.02	0.15
				1000	0.97	0.00	0.57	0.13	0.01	0.13

Table 4: **Inference properties of LiVE with increasing p and coefficient magnitudes.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

The observations are persistent with those in Section 4.2. The proposed LiVE method has coverage across different scenarios. In Table 4, when $p \in \{1001, 5001\}$ the case probabilities (< 0.5) correspond to the null hypothesis and the testing procedure has type I error controlled. However when $p = 2001$, the case probability (0.531) corresponds to an alternative in the indistinguishable region and consequently the testing procedure does not have power.

4.5 Varying sparsity of β

We test the sensitivity of our method to the sparsity assumption $k \lesssim \sqrt{n}/\log p$ by varying the sparsity over a range of values. To mimic the configuration of real data in Section 5, we set $n = 318$ and $p = 199$. Set $\Sigma = \{0.5^{1+|j-l|}\}_{1\leq j\leq l\leq (p-1)}$ and generate β as

(S7) $\beta_1 = 0$, $\beta_j = (j-1)/d$ for $2 \leq j \leq l+1$ and $\beta_j = 0$ for $l+2 \leq j \leq 199$.

We vary $d \in \{5, 10, 20\}$ and $l \in \{5, 10, 12, 15, 20\}$. With d increasing or l decreasing, the effective sparsity level decreases. Table 5 summarizes the results for the setting (S7) with x_* generated as Loading 1 with $q = 1/2$ and $r = 1/5$. For $d = 5$, our proposed LiVE method is more reliable for $l = 5$ or 10. Even though the coverage is guaranteed for a larger l, the proposed confidence intervals overcover. When d is increased to 10, the method works well for $l \leq 12$; for $l = 15, 20$, the confidence intervals are conservative. For $d = 20$, our proposed method is reliable for l being as large as 20. In general, our proposed confidence intervals are reliable for the relatively sparse signals; for the setting with a dense signal, the proposed confidence intervals are conservative and hence less informative.

Setting (S7)**, Loading 1 with** $q = 1/2$ **and** $r = 1/5$									
$\|x_*\|_2$	d	l	Prob	Cov	ERR	Len	RMSE	Bias	SE
		5	0.627	0.94	0.35	0.32	0.08	-0.01	0.08
		10	0.939	0.95	0.38	0.66	0.08	-0.06	0.05
2.22	5	12	0.940	0.99	0.07	0.89	0.09	-0.07	0.06
		15	0.951	0.99	0.01	0.98	0.13	-0.11	0.07
		20	0.891	0.99	0.00	0.99	0.23	-0.20	0.11
		5	0.564	0.93	0.23	0.29	0.08	0.00	0.08
		10	0.798	0.94	0.88	0.32	0.08	-0.04	0.07
2.22	10	12	0.798	0.95	0.62	0.43	0.08	-0.04	0.08
		15	0.815	0.97	0.20	0.67	0.10	-0.06	0.08
		20	0.741	0.99	0.01	0.93	0.14	-0.10	0.10
		5	0.532	0.94	0.14	0.29	0.08	0.00	0.08
		10	0.665	0.92	0.62	0.28	0.08	-0.01	0.08
2.22	20	12	0.665	0.94	0.60	0.30	0.08	-0.01	0.08
		15	0.677	0.95	0.49	0.35	0.08	-0.01	0.08
		20	0.629	0.97	0.08	0.53	0.10	-0.04	0.10

Table 5: **Varying sparsity of β.** The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents the empirical rejection rate of the testing procedure; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

4.6 Robustness to Violation of (A2)

We now test the robustness of our proposed method to the violation of Condition (A2). We generate β as

(S8) $\beta_1 = 0$, $\beta_j = 1$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq 501$.

We construct different covariance matrices Σ such that a certain proportion of the conditional case probability $\left\{h\left(X_i^\top \beta\right)\right\}_{i=1}^n$ are near 0 or 1.

(i) **Toeplitz Covariance :** $\Sigma \in \mathbb{R}^{500\times 500}$ is constructed as a block diagonal matrix. Each block, Σ_0 is a matrix of dimension 50×50 constructed as: $(\Sigma_0)_{i,i} = 0.5$ for $1 \le i \le 50$ and $(\Sigma_0)_{i,j} = \frac{0.03}{2}(1 - |i-j|/49)$ for $1 \le i \neq j \le 50$.

(ii) **Decaying Covariance :** $\Sigma = \{0.5^{1+|j-l|}\}_{1\le j\le l\le 500}$.

In Figure 1, we plot the case probability for $n = 600$. The left panel of Figure 1 corresponds to the setting with the Toeplitz covariance matrix, where 37 out of 600 conditional case probabilities lie below 0.1 while 35 lie above 0.9; The right panel of Figure 1 corresponds to the setting with the decaying covariance matrix, where 79 out of 600 conditional case probabilities lie below 0.1 while 77 lie above 0.9, which suggests stronger violation of assumption (A2). Due to the deeper U-shape on the right of Figure 1, Condition (A2) is more violated for the setting with a decaying covariance matrix. The inference results are summarized in Table 6.

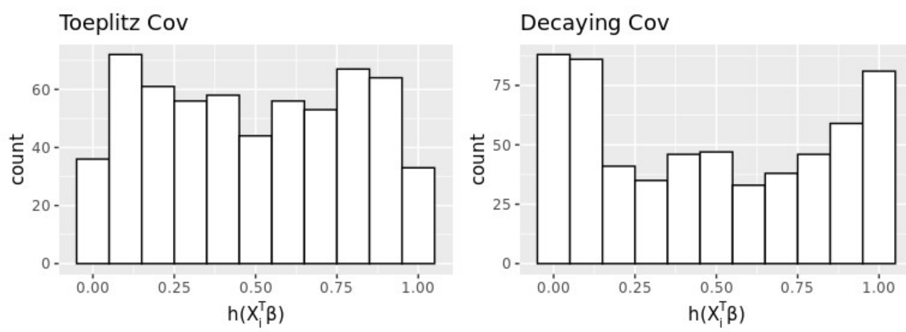

Figure 1: Histogram of $\{h(X_{i\cdot}^\intercal\beta)\}_{i=1}^n$ for a sample of $n = 600$ observations with respect to setting (S8) with Toeplitz Covariance (left) and Decaying Covariance (right).

In table 6, we observe that the stronger violation of (A2) results in our constructed CIs overcovering. The wider CIs are expected since the weights, $[h(X_{i\cdot}^\intercal\beta)(1 - h(X_{i\cdot}^\intercal\beta))]^{-1}$, can be quite large when a large proportion of $\{h(X_{i\cdot}^\intercal\beta)\}_{i=1}^n$ are close to 0 or 1. To summarize, the less U-shaped the histogram of the conditional case probabilities is, the better is the inference produced by the LiVE method.

We plot the histogram of the conditional case probability $\{h(X_{i\cdot}^\intercal\beta)\}_{i=1}^n$ and Condition (A2) is strongly violated if a large proportion of $\{h(X_{i\cdot}^\intercal\beta)\}_{i=1}^n$ concentrate around 0 or 1. We have plotted the histogram for simulation settings (S1), (S2), (S5) and (S6) in Figure C.3 in the supplement.

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

Setting (S8), Loading 2 with $q = 1$									
Toeplitz Covariance									
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
6.21	$\frac{1}{5}$	0.200	200	0.97	0.00	0.75	0.28	0.09	0.27
			400	0.96	0.00	0.61	0.21	0.06	0.20
			600	0.93	0.00	0.54	0.19	0.04	0.18
Decaying Covariance									
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
6.21	$\frac{1}{5}$	0.200	200	0.98	0.00	0.95	0.30	0.11	0.28
			400	0.99	0.00	0.90	0.26	0.10	0.24
			600	0.98	0.00	0.86	0.19	0.06	0.18

Table 6: **Robustness to violation of** (A2)**.** "r" and"Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

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 2 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 3 presented confidence intervals constructed using our method for the case probabilities shown in the top two panels in the right column in Figure 2, 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,

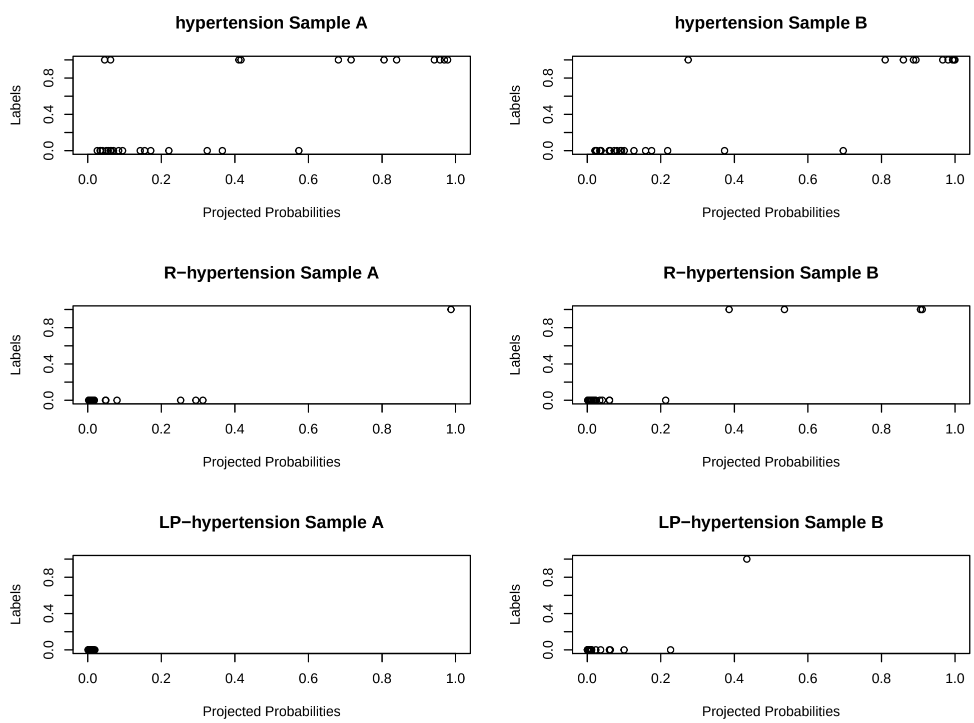

Figure 2: Performance for predicting three phenotypes in two random sub-samples.

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%.

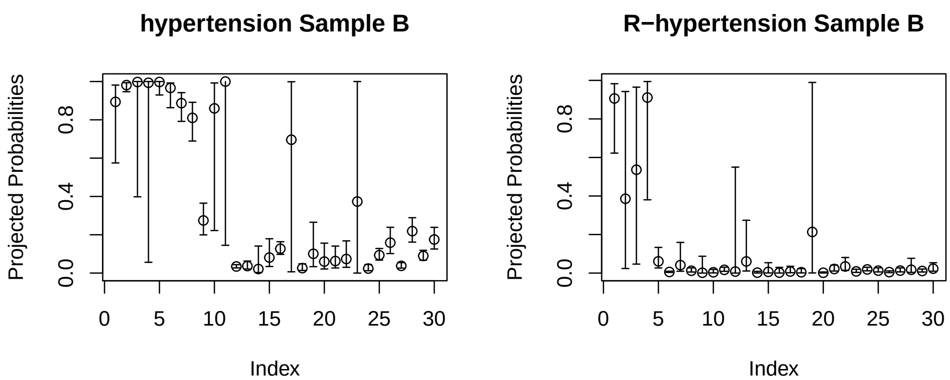

Figure 3: Confidence interval construction: on the left panel, indexes 1 to 11 correspond to observations with hypertension; indices 12 to 30 correspond to those without hypertension. On the right panel, indices 1 to 4 correspond to observations with R-hypertension; indices 5 to 30 correspond to those without R-hypertension.

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 being controls. On the left panel of Figure 3, 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 3, 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.

Additional results corresponding to the remaining four panels are presented in Figure D.1 in the supplementary materials. The observation is similar to that in Figure 3.

Acknowledgments

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 and NIH R01GM140463-01. 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 acknowledge one reviewer for suggesting the comparison with the Transformation Method. 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.

6. Proof

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

We introduce the following events

$$\mathcal{A}_1=\left\{\max_{1\leq i\leq n,\,1\leq j\leq p}|X_{ij}|\leq C\sqrt{\log n+\log p}\right\},\;\; \mathcal{A}_2=\left\{\min_{\|\eta\|_2=1,\|\eta_{S^c}\|_1\leq C\|\eta_S\|_1}\frac{1}{n}\sum_{i=1}^{n}(X_{i\cdot}^{\intercal}\eta)^2\geq c\lambda_{\min}(\Sigma)\right\}$$

$$\mathcal{A}_3=\left\{\min_{1\leq i\leq n}\frac{\exp(X_{i\cdot}^{\intercal}\beta)}{[1+\exp(X_{i\cdot}^{\intercal}\beta)]^2}\geq c_{\min}^2\right\},\;\; \mathcal{A}_4=\left\{\lambda_0=\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i\right\|_{\infty}\leq C\sqrt{\frac{\log p}{n}}\right\}$$

$$\mathcal{A}_5=\left\{\|\widehat{\beta}-\beta\|_2\leq C\sqrt{\frac{k\log p}{n}}\right\},\;\; \mathcal{A}_6=\left\{\|(\widehat{\beta}-\beta)_{S^c}\|_1\leq C_0\|(\widehat{\beta}-\beta)_S\|_1\right\}$$

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 Cai and Guo (2017).

Lemma 2 *Suppose Conditions* (A1) *and* (A2) *hold, then* $\mathbb{P}\left(\cap_{i=1}^{4}\mathcal{A}_i\right)\geq 1-\exp(-cn)-p^{-c}$ *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 B.5 in the supplementary material.

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

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

where $h'(x) = \frac{\exp(x)}{(1+\exp(x))^2}$ *and* $h''(x) = \frac{2\exp(2x)}{(1+\exp(x))^3}$. *We further have*

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

and

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

6.1 Proof of Theorem 1

Proof of (20). On the event $\mathcal{A}_3$, we have $\widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u} \leq \mathrm{V} \leq \frac{1}{c_{\min}^2}\widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u}$. To control the upper bound part $\sqrt{\mathrm{V}} \leq \frac{C_0\|x_*\|_2}{n}$, we define the following events

$$\begin{aligned} &\mathcal{B}_1 = \left\{\left\|\widehat{\Sigma}\Sigma^{-1}x_* - x_*\right\|_\infty \leq \|x_*\|_2\lambda_n\right\}; \quad \mathcal{B}_2 = \left\{\left|x_*^\intercal\widehat{\Sigma}\Sigma^{-1}x_* - \|x_*\|_2^2\right| \leq \|x_*\|_2^2\lambda_n\right\} \\ &\mathcal{B}_3 = \left\{\|X\Sigma^{-1}x_*\|_\infty \leq \|x_*\|_2\tau_n\right\} \end{aligned} \tag{33}$$

The following lemma controls the probability of $\cap_{i=1}^3\mathcal{B}_i$ and its proof is presented in section B.2.

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

$$\mathbb{P}\left(\cap_{i=1}^3\mathcal{B}_i\right) \geq 1 - n^{-c} - p^{-c}. \tag{34}$$

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, $\mathrm{V} \leq x_*^\intercal\Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}x_*/n$.

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,

$$\widetilde{u} = \arg\min_{u\in\mathbb{R}^p} u^\intercal\left(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right)u \quad \text{subject to} \quad |x_*^\intercal\widehat{\Sigma}u - \|x_*\|_2^2| \leq \|x_*\|_2^2\lambda_n \tag{35}$$

Note that $\widehat{u}$ satisfies the feasible set of (35) and hence

$$\begin{aligned} \widehat{u}^\intercal\left(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right)\widehat{u} &\geq \widetilde{u}^\intercal\left(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right)\widetilde{u} \\ &\geq \widetilde{u}^\intercal\left(\frac{1}{n}\sum_{i=1}^n X_{i\cdot}X_{i\cdot}^\intercal\right)\widetilde{u} + t\left((1-\lambda_n)\|x_*\|_2^2 - x_*^\intercal\widehat{\Sigma}\widetilde{u}\right) \text{ for any } t \geq 0, \end{aligned} \tag{36}$$

where the last inequality follows from the constraint of (35). For a given $t \geq 0$, we have

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

By solving the minimization problem of the right hand side of (37), 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 (37) is $-\frac{t^2}{4}x_*^{\intercal}\widehat{\Sigma}x_* + t(1-\lambda_n)\|x_*\|_2^2$. Combined with (36) and (37), we have

$$\widehat{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widehat{u} \geq \max_{t\geq 0}\left[-\frac{t^2}{4}x_*^{\intercal}\widehat{\Sigma}x_* + t(1-\lambda_n)\|x_*\|_2^2\right]. \tag{38}$$

For $t^* = 2\frac{(1-\lambda_n)\|x_*\|_2^2}{x_*^{\intercal}\widehat{\Sigma}x_*} > 0$, the minimum of the right hand side of (38) is achieved and hence establish

$$\widehat{u}^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\intercal}\right)\widehat{u} \geq \frac{(1-\lambda_n)^2\|x_*\|_2^4}{x_*^{\intercal}\widehat{\Sigma}x_*}. \tag{39}$$

Then $\mathbb{P}\left[\mathrm{V}^{-1/2}\left(\widehat{x_*^{\intercal}\beta} - x_*^{\intercal}\beta\right) \geq z_\alpha\right] \to \alpha$ follows from the decomposition (8), the variance control in (20), 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* $\frac{1}{V^{1/2}}\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\beta)(1-h(X_{i\cdot}^{\intercal}\beta))]^{-1}X_{i\cdot}\epsilon_i \to N(0,1)$ *where* V *is defined in* (19)*.*

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 $1 \leq i \leq n$

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

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

$$\mathcal{C} = \left\{\delta : \|\delta_{S^c}\|_1 \leq c\|\delta_S\|_1,\ \|\delta\|_2 \leq C^*\sqrt{\frac{k\log p}{n}}\right\}. \tag{40}$$

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

$$\mathcal{T} = \left\{t = (t_1, \cdots, t_n) : t_i = X_{i\cdot}^{\intercal}\delta \ \text{ where } \ \delta \in \mathcal{C}\right\}, \tag{41}$$

We can rewrite the main component of the left hand side of (26) as

$$\begin{aligned}
&\left|\widehat{u}^{\intercal}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left(\frac{\exp(X_{i\cdot}^{\intercal}\widehat{\beta})}{\left(1+\exp(X_{i\cdot}^{\intercal}\widehat{\beta})\right)^2}\right)^{-1} - \left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2}\right)^{-1}\right)X_{i\cdot}\epsilon_i\right| \cdot \mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3\cap\mathcal{A}_5\cap\mathcal{A}_6} \\
&\leq \sup_{\delta\in\mathcal{C}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i(X_{i\cdot}^{\intercal}\delta)\cdot\mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right| = \sup_{t\in\mathcal{T}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i(t_i)\cdot\mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right|
\end{aligned} \tag{42}$$

where $\mathcal{C}$ is defined in (40) and $\mathcal{T}$ is defined in (41). In the following, we control the last part of (42) via applying the symmetrization technique van de Geer (2006) stated in Lemma 6 and the contraction principle in Lemma 7. The proofs of Lemma 6 and Lemma 7 are presented in Sections B.6 and B.7 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*

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} g_i(t_i)\epsilon_i\right|\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} g_i(t_i)\xi_i\right|\right), \tag{43}$$

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

The following lemma is a modification of Theorem 2.2 in Koltchinskii (2011), where the result in Koltchinskii (2011) 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 (45). The following lemma can also be derived by extending the proof of Theorem 4.12 in Ledoux and Talagrand (1991). To be self-contained, we give a proof of Lemma 7 in the supplementary section B.7.

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 $|\phi_i(u) - \phi_i(v)| \leq |u - v|, u, v \in \mathbb{R}$. For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, then*

$$\mathbf{E}\Phi\left(\frac{1}{2}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} \phi_i(t_i)\xi_i\right|\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n} t_i\xi_i\right|\right), \tag{44}$$

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

$$\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) \in [0, \frac{1}{2}],\ \mathbb{P}(\xi_i = 0) = 1 - 2\mathbb{P}(\xi_i = 1). \tag{45}$$

We will apply Lemmas 6 and 7 and control $\sup_{t\in\mathcal{T}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(t_i)\cdot \mathbf{1}_{\mathcal{A}_1\cap\mathcal{A}_3}\cdot\epsilon_i\right|$ in (42). 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 $t_i - s_i = X_{i\cdot}^{\intercal}\left(\delta^t - \delta^s\right)$ and $t_i = X_{i\cdot}^{\intercal}\delta^t$ for $1 \leq i \leq n$. Hence on the event $\mathcal{A}_1$,

$$\max\left\{\max_{1\leq i\leq n}|t_i - s_i|, \max_{1\leq i\leq n}|t_i|\right\} \leq Ck\sqrt{\frac{\log p}{n}}\sqrt{\log p + \log n} \leq 1. \tag{46}$$

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

$$g_i(s_i) - g_i(t_i) = \left(\frac{q\left(X_{i\cdot}^{\intercal}\beta + s_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)} - 1\right)\frac{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta\right)} q\left(X_{i\cdot}^{\intercal}\beta\right)\widehat{u}^{\intercal}X_{i\cdot}. \tag{47}$$

By (31), we have

$$\left|\frac{q\left(X_{i\cdot}^{\intercal}\beta + s_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta + t_i\right)} - 1\right| \leq |\exp(|s_i - t_i|) - 1| \leq e|s_i - t_i|, \tag{48}$$

where the last inequality holds as long as $|s_i - t_i|$ is sufficiently small, as verified in (46). Similarly, we establish that $\frac{q\left(X_{i\cdot}^{\intercal}\beta+t_i\right)}{q\left(X_{i\cdot}^{\intercal}\beta\right)} \leq e$. Combined with (47) and (48), we obtain

$$|g_i(s_i) - g_i(t_i)| \leq \frac{1}{c_{\min}^2} e^2 \, |s_i - t_i| \, |\widehat{u}^{\intercal} X_{i\cdot}| \leq \frac{1}{c_{\min}^2} e^2 \, |s_i - t_i| \, \|x_*\|_2 \tau_n, \tag{49}$$

where the last inequality follows from the constraint (11). By applying (49), we have

$$\frac{1}{L_n} |g_i(t_i) - g_i(s_i)| \cdot \mathbf{1}_{\mathcal{A}_1 \cap \mathcal{A}_3} \leq |t_i - s_i| \quad \text{where} \quad L_n = \frac{e^2}{c_{\min}^2} \|x_*\|_2 \tau_n. \tag{50}$$

Define $\phi_i(t_i) = \frac{1}{L_n} g_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1}$. We then apply (43) and (44) 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 $\sup_{\delta \in \mathcal{C}} \|\delta\|_1 \leq Ck\sqrt{\frac{\log p}{n}}$, 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 $\mathbf{E} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \phi_i(t_i) \cdot \mathbf{1}_{\mathcal{A}_1} \xi_i \right| \leq C \frac{k \log p}{n} \|X_{i\cdot}\|_{\psi_2}$. By Chebyshev's inequality,

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

By (42), we establish that (26) holds with probability larger than $1 - (\frac{1}{t} + p^{-c} + \exp(-cn))$.

References

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.

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.

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

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

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

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

Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

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

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and ℓ_1+ ℓ_2 penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

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.

T Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for genetic relatedness based on high-dimensional logistic regression. *Journal of the American Statistical Association*, to appear, 2021.

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.

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.

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.

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.

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.

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.

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

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

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

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.

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.

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.

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.

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.

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.

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.

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.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer, 1991.

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.

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.

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.

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.

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.

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. O'Connor. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Network Open*, 2019.

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

Prabrisha Rakshit, T Tony Cai, and Zijian Guo. Sihr: An r package for statistical inference in high-dimensional linear and logistic regression models. *arXiv preprint arXiv:2109.03365*, 2021.

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 Kröner-Fresenius-Hyperaldosteronism Registry. Observational study mortality in treated primary aldosteronism: the german conn's registry. *Hypertension*, 60:618–624, 2012.

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.

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.

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.

N. Tripuraneni and L. Mackey. Single point transductive prediction. *arXiv preprint arXiv:1908.02341*, 2020.

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

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.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* Cambridge University Press, 2011.

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.

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.

Appendix A. Additional Discussion

A.1 Technical Difficulty of the Plug-in Debiased Estimator

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 van de Geer et al. (2014); Ning and Liu (2017); Ma et al. (2018). To see this, we can apply the results in van de Geer et al. (2014); Ning and Liu (2017); Ma et al. (2018) to show that for $1 \leq j \leq p$,

$$\widetilde{\beta}_j = \beta_j + M(\widetilde{\beta}_j) + \mathrm{Bias}(\widetilde{\beta}_j)$$

where $\sqrt{n}M(\widetilde{\beta}_j)$ is asymptotically normal and $\mathrm{Bias}(\widetilde{\beta}_j)$ is a small bias component. Then we have the following error decomposition

$$x_*^{\intercal}\widetilde{\beta} - x_*^{\intercal}\beta = \sum_{j=1}^{p} x_{*,j} M(\widetilde{\beta}_j) + \sum_{j=1}^{p} x_{*,j} \mathrm{Bias}(\widetilde{\beta}_j).$$

The component $\sqrt{n}\sum_{j=1}^{p} x_{*,j} M(\widetilde{\beta}_j)$ is asymptotically normal with its standard error of the 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 $\sum_{j=1}^{p} x_{*,j} M(\widetilde{\beta}_j)$, 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_*.

A.2 A Brief Review of Ma et al. (2018)

The bias corrected estimator proposed in Ma et al. (2018) is

$$\widetilde{\beta}_j = \widehat{\beta}_j + \frac{\sum_{i=1}^{n} v_{ij}\left(y_i - h\left(\widehat{\beta}^{\intercal} X_{i\cdot}\right)\right)}{\sum_{i=1}^{n} v_{ij} h'\left(\widehat{\beta}^{\intercal} X_{i\cdot}\right) X_{ij}}, \quad j = 1, \ldots, p \tag{51}$$

where $\widehat{\beta}$ is the penalized logistic estimator of β and $v_j = (v_{j1}, \cdots, v_{jn})^{\intercal}$ is the score vector to be constructed. Let X_j and X_{-j} denote the j-th column of X and the submatrix of X excluding the j-th column, respectively. Ma et al. (2018) construct the score v_j as follows,

$$v_j = \widehat{W}^{-1}(X_j - X_{-j}\widehat{\gamma})$$

where

$$\widehat{\gamma} = \arg\min_{b} \left\{ \frac{\|X_j - X_{-j} b\|_2^2}{2n} + \lambda \|b\|_1 \right\}, \quad \widehat{W} = \mathrm{diag}\left(h'(\widehat{\beta}^{\intercal} X_{1\cdot}), \cdots, h'(\widehat{\beta}^{\intercal} X_{n\cdot}) \right).$$

Then the estimator in (51) can be written as

$$\widetilde{\beta}_j = \widehat{\beta}_j + \frac{\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}(X_{i,j}-X_{i,-j}^{\intercal}\widehat{\gamma})\left(y_i - h\left(\widehat{\beta}^{\intercal}X_{i\cdot}\right)\right)}{\sum_{i=1}^{n}(X_{i,j}-X_{i,-j}^{\intercal}\widehat{\gamma})X_{ij}}. \tag{52}$$

The results in Ma et al. (2018) are about inference for β_j instead of an arbitrary linear combination $x_*^{\intercal}\beta$. This bias-corrected estimator in (51) is shown to be effective under a sparsity condition on $\Sigma^{-1}e_j$ where e_j is the $j-$th Euclidean basis. However, it is not straightforward to extend this to deal with arbitrary x_* and non-sparse Σ^{-1}.

A.3 Additional discussion about Tripuraneni and Mackey (2020)

The focus of the paper by Tripuraneni and Mackey (2020) is on estimation of linear functional or the related prediction problem in high-dimensional linear models. However, the high-dimensional estimation and confidence interval construction can be very different for a dense loading x_*. With respect to the method proposed in Section 3.1 in Tripuraneni and Mackey (2020), this difference has been established in Cai et al. (2019) in the high-dimensional linear model.

Firstly, Proposition 3 in Cai et al. (2019) established that if the loading x_* is of certain dense structure, then the projection direction introduced in Section 3.1 of Tripuraneni and Mackey (2020) is zero and hence the "bias-corrected" estimator is reduced to the plug-in estimator. We believe that this fact is also true in case of high-dimensional logistic regression. We have further shown that the plug-in estimator has a large bias and is not suitable for confidence interval construction.

Secondly, the confidence interval construction in Proposition 4 of Tripuraneni and Mackey (2020) requires the sparsity of β, which may be hard to obtain in practical applications.

Appendix B. Additional Proofs

B.1 Proof of Proposition 4

Proof of (22) The first inequality of (22) follows from Holder's inequality and the second inequality follows from Condition (B).
Proof of (23) By Cauchy inequality, we have

$$\sqrt{n}\left|\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\Delta_i\right| \leq \max_{1\leq i\leq n}|\widehat{u}^{\intercal}X_{i\cdot}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\Delta_i| \leq \tau_n\|x_*\|_2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\Delta_i| \tag{53}$$

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

$$\begin{aligned}
\sum_{i=1}^{n}|\Delta_i| &\leq \sum_{i=1}^{n}\exp\left(|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)|\right)\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \\
&\leq \exp\left(\max|X_{ij}|\cdot\|\widehat{\beta}-\beta\|_1\right)\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \leq C\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2.
\end{aligned} \tag{54}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 \leq C\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{k\log p}{n}. \tag{55}$$

Together with (53) and (54), we establish that, on the event $\mathcal{A}$,

$$\sqrt{n}\left|\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}\Delta_i\right| \leq C\tau_n\|x_*\|_2\frac{k\log p}{\sqrt{n}}. \tag{56}$$

B.2 Proof of Lemma 4

Define $u^* = \Sigma^{-1}x_*$. Let $D \in \mathbb{R}^n$ be defined as $D := \frac{\widehat{\Sigma}u^* - x_*}{\|u^*\|_2}$ so that the j^{th} element of D is given by $D_j = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{i\cdot}^{\intercal}u^* - x_{*,j}\right)}{\|u^*\|_2}$ where $x_{*,j}$ denotes the j^{th} component of x_*. Due to sub-gaussianity of the design $\{X_{i\cdot}\}_{i=1}^n$, D_j is a sub-exponential random variable. Since $\mathbb{E}(D_j) = 0$ for all $1 \leq j \leq p$, we apply Corollary 5.17 in Vershynin (2011) and establish,

$$\begin{aligned} &\mathbb{P}\left(|D_j| \geq C\sqrt{\frac{\log p}{n}}\right) \leq p^{-c_0} \quad \text{for some} \quad C, c_0 > 0 \quad \forall j \\ \implies &\mathbb{P}\left(\|D\|_\infty \geq C\sqrt{\frac{\log p}{n}}\right) \leq p^{1-c_0} \end{aligned} \tag{57}$$

Let $\widetilde{D} := \frac{x_*^{\intercal}\widehat{\Sigma}u^* - \|x_*\|_2^2}{\|u^*\|_2^2}$. Note that $\widetilde{D}$ is centered and sub-exponential random variable. Hence, by Corollary 5.17 in Vershynin (2011),

$$\mathbb{P}\left(|\widetilde{D}| \geq C_1\sqrt{\frac{\log p}{n}}\right) \leq p^{-c_0} \quad \text{for some} \quad C_1, c_1 > 0. \tag{58}$$

By Condition (A1), we have $\|\Sigma^{-1}\| \leq C_2$ where $C_2 > 0$ is a constant. This implies $\|u^*\|_2 = \|\Sigma^{-1}x_*\|_2 \leq C_2\|x_*\|_2$. By sub-gaussianity of X, using Proposition 5.10 in Vershynin (2011),

$$\|Xu^*\|_\infty = \max_{1\leq i\leq n}|X_{i\cdot}^{\intercal}u^*| \leq \|x_*\|_2\tau_n \tag{59}$$

holds with probability of at least $n^{(1-c_1)}$ where $c_1 > 0$ is a constant. Combining (57), (58) and (59) we establish (34).

B.3 Proof of Lemma 5

We want to establish

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2}\right)^{-1}\widehat{u}^{\intercal}X_{i\cdot}\epsilon_i \to N(0, \mathrm{V}) \tag{60}$$

Define

$$W_i = \frac{1}{\sqrt{\mathrm{V}}} \left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2} \right)^{-1} \widehat{u}^{\intercal} X_{i\cdot} \epsilon_i \tag{61}$$

Conditioning on X, then $\{W_i\}_{1\leq i\leq n}$ 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 (60), it is sufficient to check the Lindeberg's condition, that is, for any constant $\bar{\epsilon} > 0$,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E}\left(W_i^2 \mathbf{1}_{\{|W_i| \geq \bar{\epsilon}\sqrt{n}\}} \right) = 0. \tag{62}$$

Note that

$$\max_{1\leq i\leq n} \left| \frac{1}{\sqrt{\mathrm{V}}} \left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2} \right)^{-1} \widehat{u}^{\intercal} X_{i\cdot}\epsilon_i \right| \leq 2\frac{\max_{1\leq i\leq n} |\widehat{u}^{\intercal} X_{i\cdot}|}{\sqrt{V}} \leq \frac{2\tau_n \|x_*\|_2}{\sqrt{V}} \leq \bar{\epsilon}\sqrt{n} \tag{63}$$

where the first inequality follows from the fact that $\left(\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left(1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right)^2} \right)^{-1} \epsilon_i \leq \frac{2}{c_{\min}}$, the second inequality follows from $|\widehat{u}^{\intercal} X_{i\cdot}| \leq \tau_n \|x_*\|_2$ and the last inequality follows from (20) and the condition $\tau_n \ll \sqrt{n}$. Then (62) follows from (63) and by Lindeberg's central limit theorem, we establish (60).

B.4 Proof of Proposition 1

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

$$\ell(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_1 \leq \ell(\widehat{\beta}+t(\beta-\widehat{\beta})) + \lambda\|\widehat{\beta}+t(\beta-\widehat{\beta})\|_1 \leq \ell(\widehat{\beta}+t(\beta-\widehat{\beta})) + (1-t)\lambda\|\widehat{\beta}\|_1 + t\lambda\|\beta\|_1 \tag{64}$$

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

$$\frac{\ell(\widehat{\beta}) - \ell(\widehat{\beta}+t(\beta-\widehat{\beta}))}{t} + \lambda\|\widehat{\beta}\|_1 \leq \lambda\|\beta\|_1 \quad \text{for any } t \in (0,1) \tag{65}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\exp(X_{i\cdot}^{\intercal}\widehat{\beta})}{1+\exp(X_{i\cdot}^{\intercal}\widehat{\beta})} - y_i \right) X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta) + \lambda\|\widehat{\beta}\|_1 \leq \lambda\|\beta\|_1 \tag{66}$$

Note that

$$\begin{aligned} \left(\frac{\exp(X_{i\cdot}^{\intercal}\widehat{\beta})}{1+\exp(X_{i\cdot}^{\intercal}\widehat{\beta})} - y_i \right) X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta) &= \left(-\epsilon_i + \left(\frac{\exp(X_{i\cdot}^{\intercal}\widehat{\beta})}{1+\exp(X_{i\cdot}^{\intercal}\widehat{\beta})} - \frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)} \right) \right) X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta) \\ = -\epsilon_i X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta) + \int_0^1 \frac{\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta + tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)\right]^2} \left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^2 dt \end{aligned} \tag{67}$$

By (31), we have

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

Combined with (67), we have

$$
\begin{aligned}
&\int_{0}^{1}\frac{\exp\left(X_{i\cdot}^{\intercal}\beta+tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta+tX_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}dt\\
&\geq\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}\int_{0}^{1}\exp\left(-t\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)dt\\
&=\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}
\end{aligned}
\tag{69}
$$

Together with (66), we have

$$
\begin{aligned}
&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}\right)+\lambda\|\widehat{\beta}\|_{1}\\
&\leq\lambda\|\beta\|_{1}+\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i}X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\leq\lambda\|\beta\|_{1}+\lambda_{0}\|\widehat{\beta}-\beta\|_{1}.
\end{aligned}
\tag{70}
$$

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

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

Then we deduce (18) and

$$
\begin{aligned}
&\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}\right)\\
&\leq(2+\delta_0)\,\lambda_0\|\widehat{\beta}_S-\beta_S\|_1.
\end{aligned}
\tag{72}
$$

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

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(X_{i\cdot}^{\intercal}\beta\right)}{\left[1+\exp\left(X_{i\cdot}^{\intercal}\beta\right)\right]^{2}}\left(X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right)^{2}\geq c\lambda_{\min}\left(\Sigma\right)\|\widehat{\beta}-\beta\|_2^2
\tag{73}
$$

Then (72) is further simplified as

$$\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}c\lambda_{\min}\left(\Sigma\right)\|\widehat{\beta}-\beta\|_2^2\leq\left(2+\delta_0\right)\lambda_0\|\widehat{\beta}_S-\beta_S\|_1 \tag{74}$$

Case 1: Assume that

$$\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\leq c_1 \quad \text{for some } c_1>0 \tag{75}$$

then we have

$$\begin{aligned}\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}&=\int_0^1\exp\left(-t\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)dt\\ \geq\int_0^1\exp\left(-tc_1\right)dt=\frac{1-\exp\left(-c_1\right)}{c_1}\end{aligned} \tag{76}$$

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

$$c_2\|\widehat{\beta}-\beta\|_2^2\leq\lambda_0\|\widehat{\beta}_S-\beta_S\|_1\leq\sqrt{k}\lambda_0\|\widehat{\beta}_S-\beta_S\|_2 \tag{77}$$

and hence

$$\|\widehat{\beta}-\beta\|_2\lesssim\frac{1}{\lambda_{\min}}\sqrt{k}\lambda_0 \quad \text{and} \quad \|\widehat{\beta}-\beta\|_1\leq k\lambda_0 \tag{78}$$

Case 2: Assume that (75) does not hold, then

$$\frac{1-\exp\left(-\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\right)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|}\geq\frac{1-\exp(-c_1)}{\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|} \tag{79}$$

Together with (74), we have

$$c_2c_1\|\widehat{\beta}-\beta\|_2^2\leq\lambda_0\|\widehat{\beta}_S-\beta_S\|_1\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right| \tag{80}$$

By $\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|\leq\max\left|X_{ij}\right|\|\widehat{\beta}-\beta\|_1$ and (18), we further have

$$\begin{aligned}\lambda_0\|\widehat{\beta}_S-\beta_S\|_1\max_{1\leq i\leq n}\left|X_{i\cdot}^{\intercal}(\widehat{\beta}-\beta)\right|&\leq\frac{2+2\delta_0}{\delta_0}\max\left|X_{ij}\right|\lambda_0\|\widehat{\beta}_S-\beta_S\|_1^2\\&\leq\frac{2+2\delta_0}{\delta_0}\max\left|X_{ij}\right|k\lambda_0\|\widehat{\beta}_S-\beta_S\|_2^2,\end{aligned} \tag{81}$$

where the last inequality follows from Cauchy inequality. Combining (80) and (81), we have shown that if (75) does not hold, then

$$\max\left|X_{ij}\right|\frac{2+2\delta_0}{\delta_0}k\lambda_0\geq c_2c_1, \tag{82}$$

Since this contradicts the assumption that $\max\left|X_{ij}\right|k\lambda_0<\frac{c_2c_1\delta_0}{2+2\delta_0}$, we establish (78) and hence (18).

B.5 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*

$$f(x) - f(a) = f'(a)(x-a) + \int_0^1 (1-t)(x-a)^2 f''(a+t(x-a))dt \tag{83}$$

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,

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

B.6 Proof of Lemma 6

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

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

Since Φ is a convex function, we have

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

Integration of both sides of the above inequality leads to (43).

B.7 Proof of Lemma 7

The proof follows from that of Theorem 2.2 in Koltchinskii (2011) 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 (45).

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

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[A(t) + \sum_{i=1}^{n} \phi_i(t_i)\xi_i]\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[A(t) + \sum_{i=1}^{n} t_i\xi_i]\right), \tag{84}$$

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

$$\mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + \phi(t_2)\xi_0]\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + t_2\xi_0]\right), \tag{85}$$

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 (85) by establishing the following inequality,

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

This is equivalent to show

$$\Phi\left(\sup_{t\in\mathcal{T}}[t_1 + \phi(t_2)]\right) + \Phi\left(\sup_{t\in\mathcal{T}}[t_1 - \phi(t_2)]\right) \leq \Phi\left(\sup_{t\in\mathcal{T}}[t_1 + t_2]\right) + \Phi\left(\sup_{t\in\mathcal{T}}[t_1 - t_2]\right) \tag{86}$$

The above inequality follows from the same line of proof as that in Koltchinskii (2011). It remains to prove the lemma by applying induction and (85), that is,

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

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 (84). In the following, we will apply (84) to establish (44). Note that

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

By applying (84) to the function $u \to \Phi(u_+)$, which is convex and non-decreasing, we have $\mathbf{E}\Phi\left(\left(\sup_{t\in\mathcal{T}} \sum_{i=1}^n \phi_i(t_i)\xi_i\right)_+\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}} \sum_{i=1}^n t_i\xi_i\right) \leq \mathbf{E}\Phi\left(\sup_{t\in\mathcal{T}} \left|\sum_{i=1}^n t_i\xi_i\right|\right)$. Then we establish (44).

Appendix C. Additional Simulation Studies

C.1 Scale of λ_n

For our implemented algorithm, we identify the tuning parameter λ_n using the following steps. Recall $H = [b, \mathbb{I}_{p\times p}]$ and $b = \frac{1}{\|x_*\|_2} x_*$. For $t = 0$, we initialise $\lambda^0 = \sqrt{2.01 \cdot \log p/n}$ and calculate

$$\widehat{v} = \arg\min_{v\in\mathbb{R}^{p+1}} \frac{1}{4} v^\intercal H^\intercal \widehat{\Sigma} H v + b^\intercal H v + \lambda^t \|v\|_1.$$

1. If $\|\widehat{v}\|_2 < \infty$, then, for $t \geq 0$, we set $\lambda^{t+1} = \lambda^t/1.5$ and calculate

$$\widehat{v} = \arg\min_{v\in\mathbb{R}^{p+1}} \frac{1}{4} v^\intercal H^\intercal \widehat{\Sigma} H v + b^\intercal H v + \lambda^{t+1} \|v\|_1.$$

   Repeat until $\widehat{v}$ cannot be solved or $t = 5$.

2. If $\|\widehat{v}\|_2 = \infty$, then, for $t \geq 0$, we set $\lambda^{t+1} = \lambda^t \cdot 1.5$ and calculate

$$\widehat{v} = \arg\min_{v\in\mathbb{R}^{p+1}} \frac{1}{4} v^\intercal H^\intercal \widehat{\Sigma} H v + b^\intercal H v + \lambda^{t+1} \|v\|_1.$$

   Repeat until $\widehat{v}$ can be solved.

By the above algorithm, we choose the smallest $\lambda_n > 0$ such that the dual problem has a finite minimum value. Through the above algorithm, λ_n, starting from $\sqrt{2.01 \cdot \log p/n}$, can be at most reduced to $\frac{\sqrt{2.01\cdot\log p/n}}{(1.5)^6}$. In theory, we can also increase λ but our observation is that the decreasing of λ happens for almost all settings if we start with $\sqrt{2.01 \cdot \log p/n}$. We provide results of the numerical experiment for our simulation setting (S1) in Table C.1. The table shows $\lambda_n \approx 0.4\sqrt{\log p/n}$ in this specific simulation setting.

Setting (S1)**, Loading 1 with** $q = 1$		
n	λ_n	$\sqrt{\log p/n}$
200	0.07	0.18
400	0.05	0.12
600	0.04	0.10

Table C.1: Report of λ_n and $\sqrt{\log p/n}$ for the Setting (S1).

C.2 Constraint (11)

We now investigate whether our constructed projection direction $\widehat{u}$ in (27) satisfies the constraint (11) through computing the ratio $\|X\widehat{u}\|_\infty/\|x_*\|_2$. Note that as long as $\|X\widehat{u}\|_\infty/\|x_*\|_2 \leq C\sqrt{\log n}$ for some positive constant $C > 0$, then the constraint (11) is satisfied, which is sufficient for us to establish the central limit theorem.

For $n = 200$, the boxplot C.1 verifies Constraint (11) for Settings (S1) and (S2)(decay = 1) and Loading 1 with $q = 1$ and $r \in \{1, 1/25\}$. The boxplot C.2 summarizes the same for $n = 600$. These boxplots show that $\|X\widehat{u}\|_\infty/\|x_*\|_2$ is bounded above by $2.35 \cdot \sqrt{\log n}$. They demonstrate that the constraint (11) is satisfied by our constructed $\widehat{u}$.

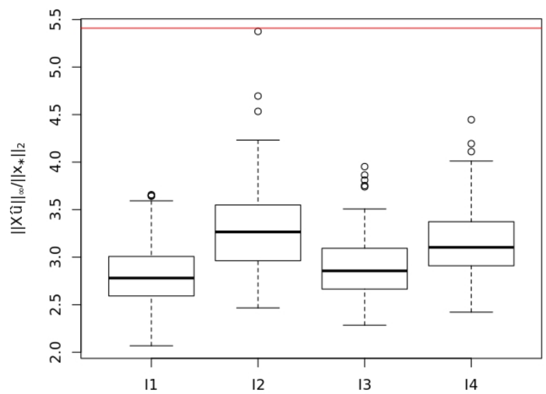

Figure C.1: Boxplot showing the distribution of $\|X\widehat{u}\|_\infty/\|x_*\|_2$ for $n = 200$, summarized over 500 simulations. Setting indices I1 and I2 denote setting (S1) with $r = 1$ and $r = \frac{1}{25}$ respectively while setting indices I3 and I4 denote setting (S2)(decay = 1) with $r = 1$ and $r = \frac{1}{25}$ respectively. The red line is corresponding to $y = 2.35 \cdot \sqrt{\log n}$

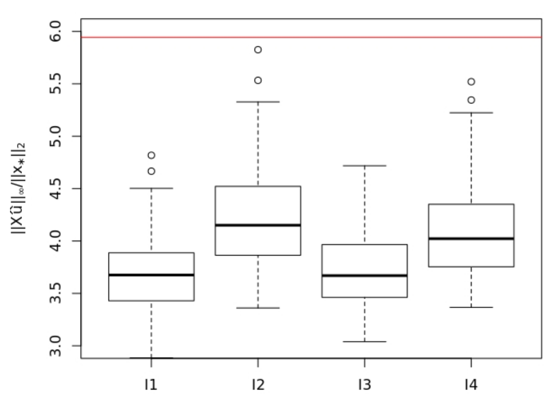

Figure C.2: Boxplot showing the distribution of $\|X\widehat{u}\|_\infty/\|x_*\|_2$ for $n = 600$, summarized over 500 simulations. Setting indices I1 and I2 denote setting (S1) with $r = 1$ and $r = \frac{1}{25}$ respectively while setting indices I3 and I4 denote setting (S2)(decay = 1) with $r = 1$ and $r = \frac{1}{25}$ respectively. The red line is corresponding to $y = 2.35 \cdot \sqrt{\log n}$

C.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.

(S9) $\beta_1 = 0$; $\beta_j = (j-1)/20$ for $2 \leq j \leq 11$ but $j \neq 9, 10$; $\beta_j = 0.01$ for $j = 9, 10$ and $\beta_j = 0$ for $12 \leq j \leq 501$.

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 $\widetilde{\Sigma} = \{0.5^{1+|j-l|}\}_{1 \leq j,l \leq 500}$ 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 comprise 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 C.2 that the post selection estimator has a large omitted variable bias and also produces a under-covered confidence interval.

Setting (S9), **Loading 3**															
				LiVE						Post Selection					
$\|x_{\text{new}}\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE	Cov	ERR	Len	RMSE	Bias	SE
14.19	$\frac{1}{25}$	0.578	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 C.2: **Comparison of the proposed method and the post selection method.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents 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.

C.4 Exactly Sparse with Intercept

Here we explore the performance of the inference procedures in presence of an intercept. We generate β as in (S1) and instead of having null intercepts we 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 with $q = 1$.

We report the simulation results based on 500 replications in Tables C.3 and C.4. Table C.3 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` undercovers 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.

Setting (S1) **with** $\beta_1 = -1$, **Loading 1 with** $q = 1$															
				LiVE			Post Selection			`hdi`			`WLDP`		
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len
16.1	1	0.501	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	$\frac{1}{25}$	0.501	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
Setting (S1) **with** $\beta_1 = 1$, **Loading 1 with** $q = 1$															
				LiVE			Post Selection			`hdi`			`WLDP`		
$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len	Cov	ERR	Len
16.1	1	0.881	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	$\frac{1}{25}$	0.881	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 C.3: **Exactly sparse regression with intercept.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents 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.

Setting (S1) **with** $\beta_1 = -1$**, Loading 1 with** $q = 1$																		
				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
			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
16.1	1	0.501	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
			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
1.09	$\frac{1}{25}$	0.501	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
Setting (S1) **with** $\beta_1 = 1$**, Loading 1 with** $q = 1$																		
				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
			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
16.1	1	0.881	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
			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
1.09	$\frac{1}{25}$	0.881	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 C.4: **Exactly sparse regression with intercept**. "r" and "Prob" represent the shrinkage parameter 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.

C.5 Additional Simulation Results for Section 4.2

We consider the exactly sparse regression setup (S1) and report the inference results for Loading 2 with $q = 1$ in Table C.5. The CIs constructed by LiVE and `hdi` have coverage over different scenarios while `WLDP` and the post-selection suffer from the issue of over-coverage and under-coverage respectively. The case probability being less than 0.5 (0.293) 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 $n = 200$.

In Table C.6, we compare the proposed estimator, the post selection estimator, the plug-in `hdi`, `WLDP` and Lasso estimator in terms of Root Mean Squared Error (RMSE), bias and standard error with respect to regression setting (S1). 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. This matches with the theoretical results presented in Theorem 1.

The inference results in the approximately sparse regression setup (S2) with decay = 2 are summarized in Table C.7. The main observations are similar to that for decay = 1. However for decay = 2, the testing procedures based on the proposed LiVE, `hdi` and `WLDP`

Setting $(S1)$, Loading 2 with $q=1$																				
				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
16.6	1	0.293	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	$\frac{1}{25}$	0.293	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 C.5: **Varying n and $\|x_*\|_2$.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the CIs; the column indexed with "ERR" represents the empirical rejection rate of the test; "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.

Setting (S1), Loading 1 with $q=1$																		
				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	0.732	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	$\frac{1}{25}$	0.732	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
Setting (S1), Loading 2 with $q=1$																		
				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
16.6	1	0.293	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	$\frac{1}{25}$	0.293	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 C.6: **Varying n and $\|x_*\|_2$.** "r" and "Prob" represent the shrinkage parameter 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.

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 $r = 1$.

Here we also report the estimation accuracy results for the approximately sparse regression setup (S2) with decay $\in \{1, 2\}$. Table C.8 summarizes the estimation accuracy results for Loading 1 and the results are similar to the exactly sparse setting in Table C.6. Table C.8 shows again the plug-in Lasso estimator cannot be used for confidence interval construction owing to its large bias.

Setting (S2) **with** decay $= 2$**, Loading 1 with** $q = 1$																			
				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	0.488	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 C.7: **Varying n and $\|x_*\|_2$.** "r" and"Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents 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.

C.6 Additional Simulation Results for Section 4.4

The inference results for the exactly sparse regression setup (S5) with respect to Loading 2 with $q = 1/2$ are reported in Table C.9.

We summarize the results for the approximately sparse regression setup (S6) with decay $= 1$ and decay $= 2$ in Tables C.10 and C.11 respectively. Tables C.9, C.10 and C.11 further support the validity of the constructed confidence intervals. The proposed test has type I error controlled when the case probabilities correspond to the null hypothesis. However for $p = 2001$ the testing procedure does not have power since the case probabilities $(0.508, 0.530, 0.533)$ correspond to alternatives in the indistinguishable region.

C.7 Additional Simulation Results for Section 4.6

We plot the histograms of conditional case probabilities for settings (S1), (S2), (S5) and (S6) in the figure C.3. The inverted U-shape of the histograms indicates that assumption (A2) is plausible or weakly violated. Consequently the LiVE method performs well with respect to CI construction and hypothesis testing as indicated earlier in sections 4.2 and 4.4.

Setting (S2) **with** decay $=1$, **Loading 1 with** $q=1$																		
				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
			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
16.1	1	0.645	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
			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
1.09	$\frac{1}{25}$	0.523	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
Setting (S2) **with** decay $=2$, **Loading 1 with** $q=1$																		
				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
			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
16.1	1	0.488	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
			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
1.09	$\frac{1}{25}$	0.481	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 C.8: **Varying n and $\|x_*\|_2$.** "r" and "Prob" represent the shrinkage parameter 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.

Setting (S5), **Loading 2 with** $q=1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
				400	0.97	0.00	0.56	0.14	-0.03	0.14
1001	3.35	$\frac{1}{5}$	0.278	600	0.97	0.00	0.53	0.13	-0.03	0.13
				1000	0.97	0.00	0.47	0.12	-0.02	0.12
				400	0.95	0.00	0.78	0.26	-0.14	0.22
2001	4.89	$\frac{1}{5}$	0.508	600	0.93	0.01	0.73	0.24	-0.12	0.20
				1000	0.95	0.00	0.67	0.19	-0.10	0.17
				400	0.98	0.02	0.88	0.29	0.08	0.28
5001	7.10	$\frac{1}{5}$	0.363	600	0.99	0.00	0.85	0.23	0.07	0.21
				1000	0.97	0.00	0.80	0.20	0.02	0.20

Table C.9: **Inference properties of LiVE with increasing p and coefficient magnitudes.** "r" and"Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents the empirical rejection rate of the testing procedure; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

Setting (S6) with decay = 1, Loading 1 with $q = 1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
1001	3.21	$\frac{1}{5}$	0.252	400	0.93	0.00	0.40	0.11	0.05	0.10
				600	0.95	0.00	0.37	0.10	0.04	0.09
				1000	0.95	0.00	0.28	0.07	0.03	0.07
2001	4.60	$\frac{1}{5}$	0.412	400	0.95	0.03	0.56	0.16	0.07	0.14
				600	0.97	0.01	0.49	0.14	0.06	0.13
				1000	0.96	0.01	0.45	0.11	0.04	0.11
5001	7.07	$\frac{1}{5}$	0.373	400	0.99	0.00	0.71	0.19	0.03	0.19
				600	0.98	0.00	0.63	0.19	0.03	0.19
				1000	0.97	0.00	0.55	0.13	0.00	0.13
Setting (S6) **with** decay = 1, **Loading 2 with** $q = 1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
1001	3.35	$\frac{1}{5}$	0.302	400	0.95	0.00	0.55	0.15	-0.06	0.14
				600	0.96	0.00	0.53	0.14	-0.04	0.13
				1000	0.96	0.00	0.47	0.13	-0.03	0.12
2001	4.89	$\frac{1}{5}$	0.530	400	0.97	0.02	0.80	0.25	-0.10	0.23
				600	0.96	0.01	0.74	0.22	-0.08	0.20
				1000	0.97	0.01	0.67	0.19	-0.07	0.17
5001	7.10	$\frac{1}{5}$	0.414	400	0.99	0.01	0.89	0.28	0.08	0.27
				600	0.97	0.00	0.86	0.25	0.04	0.25
				1000	0.98	0.00	0.80	0.20	0.01	0.20

Table C.10: **Inference properties of LiVE with increasing p and coefficient magnitudes.** "r" and "Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents the empirical rejection rate of the testing procedure; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

Setting (S6) **with** decay = 2, **Loading 1 with** $q = 1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
1001	3.21	$\frac{1}{5}$	0.257	400	0.95	0.00	0.40	0.11	0.05	0.10
				600	0.95	0.00	0.37	0.10	0.04	0.09
				1000	0.96	0.00	0.28	0.07	0.02	0.07
2001	4.60	$\frac{1}{5}$	0.365	400	0.97	0.01	0.55	0.15	0.05	0.14
				600	0.95	0.01	0.48	0.13	0.03	0.13
				1000	0.96	0.00	0.43	0.10	0.02	0.10
5001	7.07	$\frac{1}{5}$	0.396	400	0.96	0.01	0.71	0.20	0.03	0.20
				600	0.96	0.00	0.66	0.16	0.02	0.16
				1000	0.95	0.00	0.55	0.14	0.00	0.14
Setting (S6) **with** decay = 2, **Loading 2 with** $q = 1/2$										
p	$\|x_*\|_2$	r	Prob	n	Cov	ERR	Len	RMSE	Bias	SE
1001	3.35	$\frac{1}{5}$	0.322	400	0.97	0.00	0.58	0.15	-0.05	0.14
				600	0.96	0.00	0.55	0.14	-0.04	0.14
				1000	0.97	0.00	0.49	0.13	-0.03	0.12
2001	4.89	$\frac{1}{5}$	0.533	400	0.98	0.02	0.80	0.24	-0.06	0.23
				600	0.96	0.02	0.75	0.20	-0.04	0.20
				1000	0.97	0.03	0.67	0.19	-0.04	0.18
5001	7.10	$\frac{1}{5}$	0.431	400	0.98	0.03	0.89	0.32	0.11	0.30
				600	0.99	0.06	0.81	0.28	0.06	0.27
				1000	0.97	0.00	0.80	0.24	0.06	0.23

Table C.11: **Inference properties of LiVE with increasing p and coefficient magnitudes.** "r" and"Prob" represent the shrinkage parameter and Case Probability respectively. The columns indexed with "Cov" and "Len" represent the empirical coverage and length of the constructed CIs respectively; the column indexed with "ERR" represents the empirical rejection rate of the testing procedure; The columns indexed with "RMSE", "Bias" and "SE" represent the RMSE, bias and standard error, respectively.

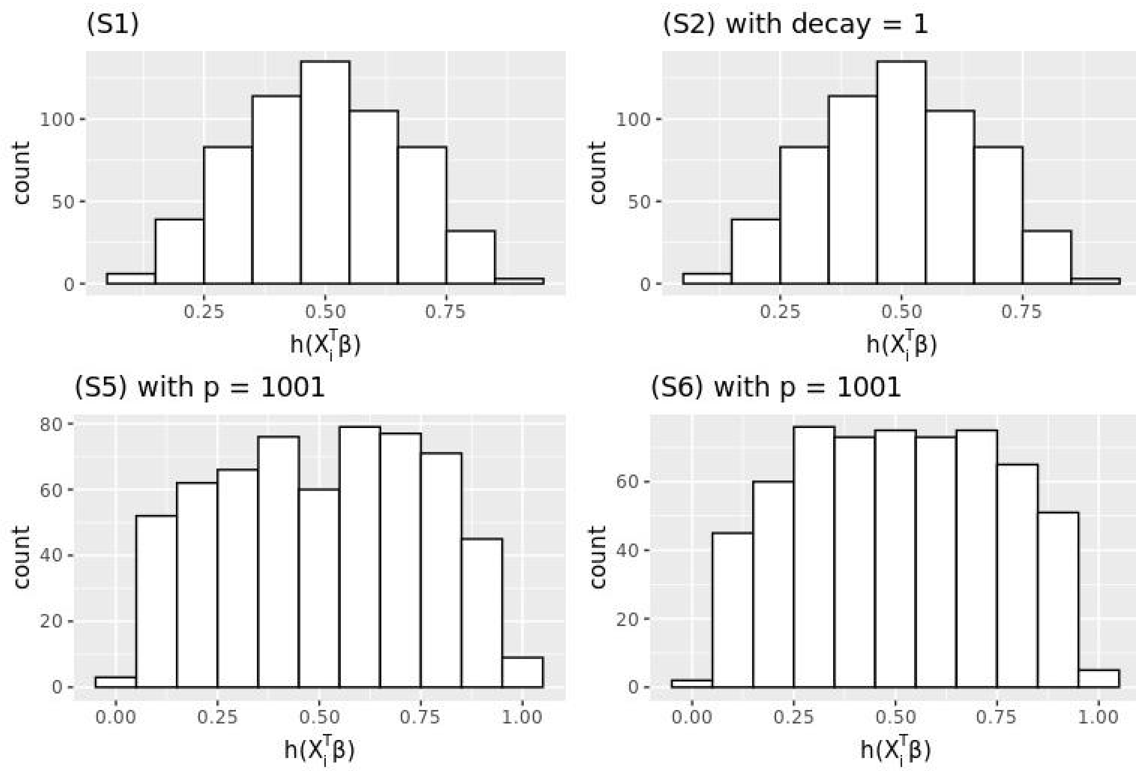

Figure C.3: Histogram showing the distribution of $\{h(X_{i\cdot}^{\intercal}\beta)\}_{i=1}^{n}$ for sample 1 with respect to regression settings (S1) (top left), (S2) with decay = 1 (top right), (S5) with $p = 1001$ (bottom left) and (S6) with decay = 1 and $p = 1001$ (bottom right). Here sample size $n = 600$

Appendix D. Additional Real Data Analysis

Figure D.1 presents confidence intervals constructed using our method for the predicted probabilities shown in all six panels in Figure 2, corresponding to prediction of hypertension, resistant hypertension and high blood pressure with unexplained low blood potassium across two random subsamples.

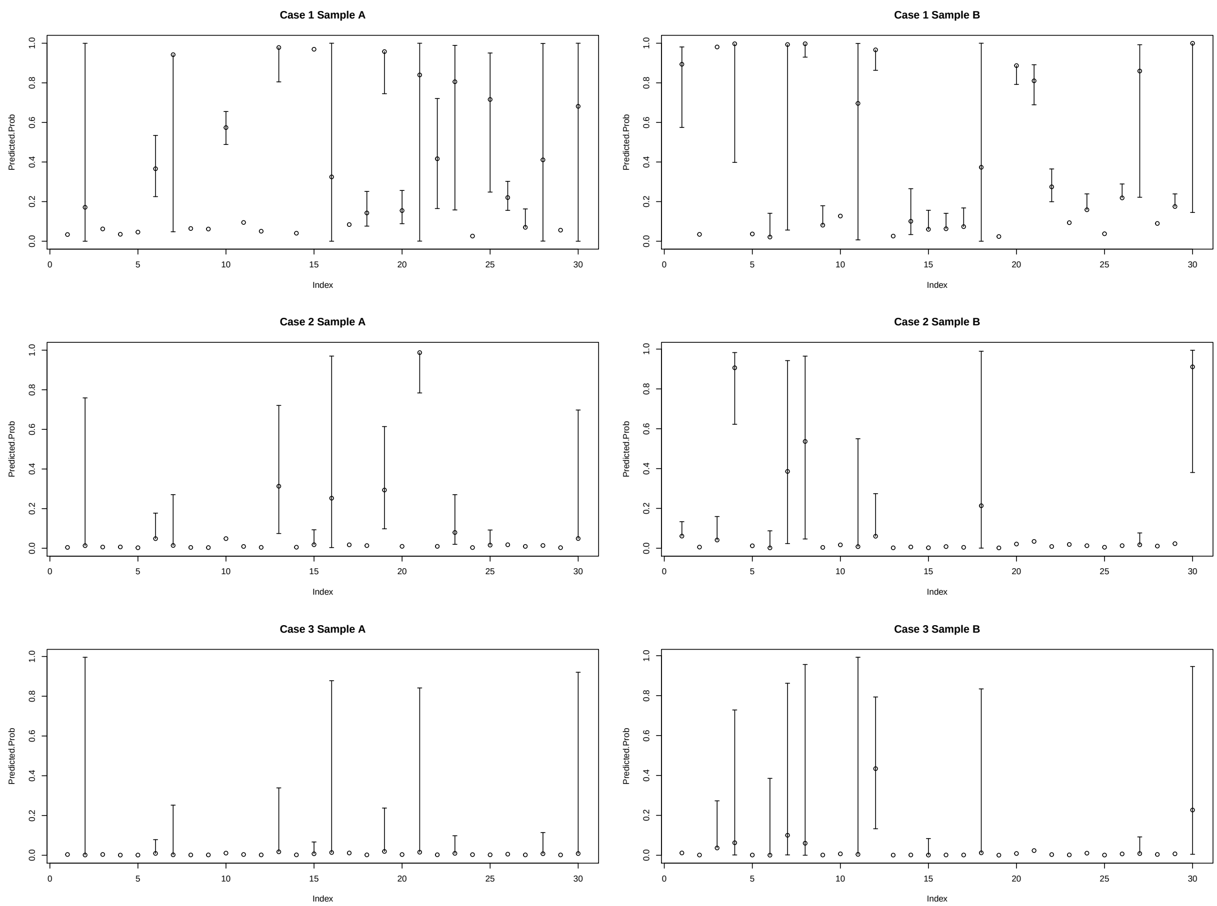

Figure D.1: Confidence interval construction for the random subsamples