High-Dimensional Inference for Generalized Linear Models with Hidden Confounding

Jing Ouyang   JINGOY@UMICH.EDU
Kean Ming Tan   KEANMING@UMICH.EDU
Gongjun Xu   GONGJUN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Abstract

Statistical inferences for high-dimensional regression models have been extensively studied for their wide applications ranging from genomics, neuroscience, to economics. However, in practice, there are often potential unmeasured confounders associated with both the response and covariates, which can lead to invalidity of standard debiasing methods. This paper focuses on a generalized linear regression framework with hidden confounding and proposes a debiasing approach to address this high-dimensional problem, by adjusting for the effects induced by the unmeasured confounders. We establish consistency and asymptotic normality for the proposed debiased estimator. The finite sample performance of the proposed method is demonstrated through extensive numerical studies and an application to a genetic data set.

Keywords: High-dimensional inference; Generalized linear model; Latent variable; Unmeasured confounder.

1. Introduction

Statistical inferences for high-dimensional regression models have received a growing interest due to the increasing number of complex data sets with high-dimensional covariates that are collected across many different scientific disciplines ranging from genomics to social science to econometrics (Peng et al., 2010; Belloni et al., 2012; Fan et al., 2014; Bühlmann et al., 2014). One of the most popularly used high-dimensional regression methods is the lasso linear regression, which assumes that the underlying regression coefficients are sparse (Tibshirani, 1996). However, the lasso penalty introduces non-negligible bias that renders high-dimensional statistical inference challenging (Wainwright, 2019). To address this challenge, Zhang and Zhang (2014) and van de Geer et al. (2014) proposed the debiasing method to construct confidence intervals for the lasso regression coefficients: the main idea is to first obtain the lasso estimator, and then correct for the bias of the lasso estimator using a low-dimensional projection method. We refer the reader to Javanmard and Montanari (2014), Belloni et al. (2014), Ning and Liu (2017), Chernozhukov et al. (2018), among others, for detailed discussions of the debiasing approach, and also Wainwright (2019) for an overview of high-dimensional statistical inference.

The aforementioned studies were established under the assumption that there are no unmeasured confounders that are associated with both the response and covariates. However, this assumption is often violated in observational studies. For instance, in genetic studies, the effect of certain segments of DNA on the gene expression may be confounded by population structure and microarray expression artifacts (Listgarten et al., 2010). Another example is in healthcare studies where the effect of nutrients intake on the risk of cancer may be confounded by physical wellness, social class, and behavioral factors (Fewell et al., 2007). Without adjusting for the unmeasured confounders, the resulting inferences from the standard debiasing methods could be biased and consequently, lead to spurious scientific discoveries.

Various methods have been proposed to perform valid statistical inferences for regression parameters in the presence of hidden confounders. One commonly used approach is instrumental variables regression, which typically requires domain knowledge to identify valid instrumental variables, making it challenging for high-dimensional applications (Kang et al., 2016; Burgess et al., 2017; Guo et al., 2018; Windmeijer et al., 2019). Recently, under a relaxed linear structural equation modeling framework, Guo et al. (2022) proposed a deconfounding approach for statistical inference for individual regression coefficients. Fan et al. (2023) developed a factor-adjusted debiasing method and conducted statistical estimation and inference for the coefficient vector. However, these two methods rely on the linearity assumption between the response and covariates, which may not hold when the response is not continuous but categorical (binary, count, etc.). Such categorical data commonly occurs in genetic, biomedical, and social science applications, and thus alternative methods are needed to perform valid statistical inferences in these cases.

To bridge the gap in the existing literature, we propose a novel framework to perform statistical inference in the context of high-dimensional generalized linear models with hidden confounding. The main idea of our proposed method involves estimating the unmeasured confounders using high-dimensional factor analysis techniques and then applying the debiasing method on the regression coefficient of interest, with the estimated unmeasured confounders treated as surrogate variables. This method does not rely on the linear model assumption or any specific model form and it is applicable to more general models beyond the generalized linear models, such as graphical models (Ren et al., 2015; Zhu et al., 2020) and additive hazards models (Lin and Ying, 1994; Lin and Lv, 2013).

Theoretically, we show that under some mild scaling conditions, the estimation errors of proposed estimators achieve comparable rates as that of the ℓ_1-penalized generalized linear model without unmeasured confounders. We further show that the debiased estimator for the coefficient of interest is asymptotically Gaussian after adjusting for the unmeasured confounders, which results in a valid statistical inference for high-dimensional generalized linear models with hidden confounding. It is worth highlighting that when using a factor model to relate covariates and unmeasured confounders, we make more general assumptions on the random noise compared to existing works. Specifically, we allow the random noise to be non-identically distributed. This represents a significant improvement over the majority of previous works, which assume that the random noise follows an independent and identical distribution (Guo et al., 2022; Fan et al., 2023). Furthermore, unlike existing methods under the linear framework (Guo et al., 2022; Fan et al., 2023), generalized linear models pose new challenges as many transformation and decomposition techniques commonly used

for the linear models are not directly applicable. Consequently, more general and refined intermediate results are needed as "building blocks" in the proofs (see Remark 8 in Appendix D.1 for details).

Our paper is organized and structured as follows. In Section 2, we introduce the model setup and provide a comprehensive discussion of related literature. Section 3 presents our two-step approach for estimating the parameter of interest while adjusting for the effect of unmeasured confounders. Section 4 establishes the theoretical properties of the parameter including the estimation consistency as well as the asymptotic normality for the estimator. In Section 5, we demonstrate the performance of our proposed method and the validity of theoretical results via extensive simulation studies. We also provide an application to genetic data containing gene expression quantifications and stimulation statuses in mouse bone marrow-derived dendritic cells, where we identify significant gene expressions under stimulations, which are consistent with the experimental findings in genetic studies (Nemetz et al., 1999; Ather and Poynter, 2018; Jang et al., 2018; Toyoshima et al., 2019). Lastly, in Section 6, we provide concluding remarks and outline potential future directions for this research.

2. Generalized Linear Models with Hidden Confounding

In this section, we first set up a generalized linear model with hidden confounding and introduce a scientific application of our model framework. Then we will discuss related high-dimensional models in the existing literature.

2.1 Problem Setup

Consider a high-dimensional regression problem with a unidimensional response y and a p-dimensional observed covariates $\boldsymbol{X}$. In addition, assume that there is a K-dimensional unmeasured confounders $\boldsymbol{U}$ that are related to both y and $\boldsymbol{X}$. Without loss of generality, we write $\boldsymbol{X} = (D, \boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}$, where $D \in \mathbb{R}$ is the covariate of interest and $\boldsymbol{Q} \in \mathbb{R}^{p-1}$ is a $(p-1)$-dimensional vector of nuisance covariates. Furthermore, we denote $\theta \in \mathbb{R}$ as the univariate parameter of interest, $\boldsymbol{v} \in \mathbb{R}^{p-1}$ as parameters for the nuisance covariates, and $\boldsymbol{\beta} \in \mathbb{R}^K$ as parameters that quantify the effects induced by the unmeasured confounders. The goal is to perform statistical inference on the parameter θ.

We assume that y given D, $\boldsymbol{Q}$, and the unmeasured confounders $\boldsymbol{U}$ follows a generalized linear model with probability density (mass) function:

$$f(y) = \exp\left[\{y(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}) - b(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U})\}/a(\phi) + c(y, \phi)\right], \qquad (1)$$

where ϕ is the scale parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are some known functions. As the distribution of y belongs to the exponential family, we have $E(y) = b'(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U})$ and $\mathrm{var}(y) = b''(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U})a(\phi)$. For simplicity, we take $a(\phi) = 1$. For notational convenience, let $\boldsymbol{Z} = (D, \boldsymbol{Q}^{\mathrm{T}}, \boldsymbol{U}^{\mathrm{T}})^{\mathrm{T}}$ be a vector that includes all observed covariates and the unmeasured confounders, and let $\boldsymbol{\eta} = (\theta, \boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ be the corresponding parameters. We now provide three commonly used examples.

Example 1 (Logistic Regression) *Let $y \in \{0, 1\}$ be a binary variable. Given covariates D, $\boldsymbol{Q}$, and unmeasured confounders $\boldsymbol{U}$, the response y follows the logistic regression model with $\phi = 1$, $a(\phi) = 1$, $b(t) = \log\{1 + \exp(t)\}$, and $c(y, \phi) = 0$.*

Example 2 (Poisson Regression) *Let $y \in \{0, 1, 2, \ldots\}$ be a discrete variable. Given covariates D, $\boldsymbol{Q}$, and unmeasured confounders $\boldsymbol{U}$, the response y follows the Poisson regression model with $\phi = 1$, $a(\phi) = 1$, $b(t) = \exp(t)$, and $c(y, \phi) = -\log(y!)$.*

Example 3 (Linear Regression) *Let $y \in \mathbb{R}$ be a real-valued response variable. Given covariates D, $\boldsymbol{Q}$, and unmeasured confounders $\boldsymbol{U}$, the response y follows the linear regression model $y = \theta D + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{U} + \varepsilon$ with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$. The model parameters are $\phi = \sigma^2$, $a(\sigma^2) = \sigma^2$, $b(t) = t^2/2$ and $c(y, \sigma^2) = -y^2/(2\sigma^2) - \log(2\pi\sigma^2)/2$.*

In addition, we assume that the relationship between covariates $\boldsymbol{X}$ and unmeasured confounders $\boldsymbol{U}$ is captured by the following model:

$$\boldsymbol{X} = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{U} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{K \times p}$ is the loading matrix that describes the linear effect of unmeasured confounders $\boldsymbol{U}$ on covariates $\boldsymbol{X}$, and $\boldsymbol{E}$ is the random noise independent of $\boldsymbol{U}$. While similar model as in (2) is considered in Guo et al. (2022) and Fan et al. (2023), here we assume model (2) is an approximate factor model which allows for weak correlation and non-identical distribution of the random noise. This is a general setting compared to many existing works assuming $\boldsymbol{E}$ to be identically and independently distributed (Guo et al., 2022; Fan et al., 2023). We will elaborate on this model setup and expound the assumptions for weakly correlated random noise in Sections 3 and 4.

The aforementioned structural equation modeling framework can be applied to many scientific applications. In the following, we provide one motivating example in genetic studies. Various authors have found that the effect of gene expression in response to the environmental conditions, e.g., viral or bacterial stimulation, might be confounded by unmeasured factors (Price et al., 2006; Lazar et al., 2013). One interesting scientific problem is to assess the effect of gene expression responding to the viral stimulation to cells, while adjusting for the confounding effects from the unmeasured variables. The viral stimulation status, a binary variable, is considered as the response variable y. Therefore, a generalized linear model for modeling y is preferred. In this example, y is the viral stimulation status, $\boldsymbol{X}$ is a vector of high-dimensional gene expression, and $\boldsymbol{U}$ is a vector of possible unmeasured confounders. More details of the data application will be provided in Section 5.2.

2.2 Related Models in Existing Literature

In high-dimensional regression, the adjustment for the hidden confounding effects is a challenging and intriguing problem. There are various related models in the literature. For instance, to address the difficulty of model selection when $p \gg n$, Paul et al. (2008) assumed that y and $\boldsymbol{X}$ are connected via a low-dimensional latent variable model: $y = \boldsymbol{\beta}^T \boldsymbol{U} + \varepsilon$ and $\boldsymbol{X} = \boldsymbol{W}^T \boldsymbol{U} + \boldsymbol{E}$, where the latent factors are associated with response and covariates are only used to infer the latent factors. However, the response and covariates are not directly associated. Different from the latent variable model in Paul et al. (2008), Bai and Ng

(2006) considered an additional low-dimensional covariate $\boldsymbol{G}$ to the latent variable model, that is expressed as $y = \boldsymbol{\beta}^T \boldsymbol{U} + \boldsymbol{\varrho}^T \boldsymbol{G} + \varepsilon$ and $\boldsymbol{X} = \boldsymbol{W}^T \boldsymbol{U} + \boldsymbol{E}$. Besides, there are other factor-adjusted models that can be extended to adjust hidden confounding. For example, Fan et al. (2020) studied the high-dimensional model selection problem when covariates are highly correlated. As most commonly used model selection methods may fail with highly-correlated covariates, they used a factor model to reduce the dependency among covariates and proposed a factor-adjusted regularized model section method. Fan et al. (2020) considered a generalized linear model between response and covariate, which together with the factor model forms a similar model framework as ours. However, the problem they studied is fundamentally different than our problem. They did not assume hidden confounding and the factor model is only used to identify a low rank part of highly-correlated covariates whereas in our problem, we focus on the regression problem with unmeasured confounders associated with response and covariates. Similar factor-adjusted methods have also been studied in other settings (e.g., Gagnon-Bartsch and Speed, 2011; Wang and Blei, 2019); however, as noted in Ćevid et al. (2020), related theoretical justifications are still underdeveloped in the literature.

Recently, many researchers studied the following linear hidden confounding model, which can be viewed as a special case of our framework,

$$y = \theta D + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{U} + \varepsilon, \quad \boldsymbol{X} = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{U} + \boldsymbol{E}. \tag{3}$$

Under this model, Kneip and Sarda (2011) used the principal component method to estimate the unmeasured confounders and then applied a selection procedure on a projected model. Ćevid et al. (2020) proposed a method that first performs the spectral transformation pre-processing step and then applies the lasso regression on the transformed response and covariates. However, these two works focused on estimation consistency and did not address inference issues. Several works have investigated statistical inference on the covariate coefficients. For example, Guo et al. (2022) proposed a doubly debiased lasso method to perform statistical inference for θ. Different from the approach of Guo et al. (2022) that implicitly adjusts for the hidden confounding effects, Fan et al. (2023) proposed to first use the principal component method to estimate unmeasured confounders and then construct the bias-corrected estimator for $\|(\theta, \boldsymbol{v}^{\mathrm{T}})\|_\infty$, which involves the decomposition of the estimation error relying on the linear form of the response and uses the projection of the response onto the factor space. The motivation of Fan et al. (2023) originated from Fan et al. (2020). When covariates are highly correlated, the leading factors are likely to have extra impacts on the response. So they augmented the factor into the sparse linear regression model between response and covariates, which is written as (3). However, these techniques are designed for linear models and may not be directly applicable to generalized linear model settings.

3. Estimation Method

In this section, we propose a novel framework to perform statistical inference for a parameter of interest in the context of high-dimensional generalized linear models with hidden confounding. In the proposed framework, we first estimate the unmeasured confounders using a factor analysis approach. Subsequently, the estimated unmeasured confounders are

treated as surrogate variables for fitting a high-dimensional generalized linear model, and a debiased estimator is constructed to perform statistical inference.

Throughout this section, we assume that the observed data $\{y_i, \boldsymbol{X}_i\}_{i=1,\dots,n}$ and the unmeasured confounders $\{\boldsymbol{U}_i\}_{i=1,\dots,n}$ are realizations of (1) and (2). Moreover, the random noise $\boldsymbol{E}_i = (E_{i1}, \dots, E_{ip})^{\mathrm{T}}$ has mean zero and variance $\boldsymbol{\Omega}_i = \mathbb{E}(\boldsymbol{E}_i \boldsymbol{E}_i^{\mathrm{T}})$. Let $\boldsymbol{\Sigma}_e = \mathrm{diag}(n^{-1} \sum_{i=1}^{n} \boldsymbol{\Omega}_i)$, where $\mathrm{diag}(\boldsymbol{A})$ denotes a diagonal matrix by setting off-diagonal entries in $\boldsymbol{A}$ to zero. In the $p \times p$ diagonal matrix $\boldsymbol{\Sigma}_e$, we denote the j-th diagonal element to be $\sigma_j^2 = n^{-1} \sum_{i=1}^{n} \tau_{i,jj}$, where $\tau_{i,jj}$ is the (j,j) element of $\boldsymbol{\Omega}_i$. The model assumption on the random noise is general as it does not assume that the random noise $\boldsymbol{E}_i$ is identical nor does it require the covariance matrix $\boldsymbol{\Omega}_i$ to be diagonal. The detailed theoretical assumptions regarding the random noise will be presented in Section 4.

Estimation of the Unmeasured Confounders: In this work, we consider the dimension K as pre-specified. In practice, there are various methods to estimate the dimension of unmeasured confounders such as scree plot (Cattell, 1966), cross-validation method (Owen and Wang, 2016), information criteria method (Bai and Ng, 2002), the eigenvalue ratio method (Lam and Yao, 2012; Ahn and Horenstein, 2013), among others. In the implementation of our proposed method, we recommend the parallel analysis approach because of its good finite sample performance, easy implementation, and popularity in scientific applications (Hayton et al., 2004; Costello and Osborne, 2005; Brown, 2015) – it has shown superior performances compared to many other methods in various empirical studies (Zwick and Velicer, 1986; Peres-Neto et al., 2005). Detailed discussions on the estimation of the dimension of unmeasured confounders are presented in Appendix C.

We employ the maximum likelihood estimation to estimate the unmeasured confounders under (2). Without loss of generality, let $\bar{\boldsymbol{U}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i = 0$ and let $\boldsymbol{S}_u = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{U}_i^{\mathrm{T}}$ be the sample variance of $\boldsymbol{U}$. Similarly, let $\boldsymbol{S}_x = n^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\mathrm{T}}$ be the sample variance of $\boldsymbol{X}$, where $\bar{\boldsymbol{X}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$. Given unmeasured confounders $\boldsymbol{U}_1, \dots, \boldsymbol{U}_n$, define an approximation of population variance of $\boldsymbol{X}$ to be $\boldsymbol{\Sigma}_x = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_u \boldsymbol{W} + \boldsymbol{\Sigma}_e$. Note that $\boldsymbol{\Sigma}_x$ is not exactly the covariance matrix of $\boldsymbol{X}$ because we do not restrict $\boldsymbol{\Omega}_i$ to be diagonal and define $\boldsymbol{\Sigma}_e$ to be diagonal by setting the off-diagonal of $n^{-1} \sum_{i=1}^{n} \boldsymbol{\Omega}_i$ to be zero. Based on the factor model in (2), the maximum likelihood estimators of $\boldsymbol{W}$, $\boldsymbol{S}_u$ and $\boldsymbol{\Sigma}_e$ are obtained as follows:

$$(\hat{\boldsymbol{W}}, \hat{\boldsymbol{S}}_u, \hat{\boldsymbol{\Sigma}}_e) = \operatorname*{argmax}_{\boldsymbol{W}, \boldsymbol{S}_u, \boldsymbol{\Sigma}_e} \left\{ -(2p)^{-1} \log |\boldsymbol{\Sigma}_x| - (2p)^{-1} \operatorname{tr}(\boldsymbol{S}_x \boldsymbol{\Sigma}_x^{-1}) \right\}.$$

Computationally, we employ the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimators as suggested in Bai and Li (2012) and Bai and Li (2016), where the authors proved the EM solutions are the stationary points for the likelihood function. Specifically, in this iterated EM algorithm, we use the principal components estimator as the initial estimator. Because principal component estimators are shown to be consistent estimators under similar model assumptions as ours (Fan et al., 2013; Wang and Fan, 2017), using the principal component estimators in initialization instead of using random initialization helps to improve algorithm efficiency and find more refined estimation results. At the t-th iteration, denote the estimators at this step to be $\boldsymbol{W}^{(t)}$ and $\boldsymbol{\Sigma}_e^{(t)}$. The

EM algorithm updates the estimators to be

$$\{\boldsymbol{W}^{(t+1)}\}^{\mathrm{T}} = \left[n^{-1} \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{X}_i \boldsymbol{U}_i^{\mathrm{T}} \mid \boldsymbol{X}, \boldsymbol{W}^{(t)}, \boldsymbol{\Sigma}_e^{(t)}) \right] \left[n^{-1} \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{U}_i \boldsymbol{U}_i^{\mathrm{T}} \mid \boldsymbol{X}, \boldsymbol{W}^{(t)}, \boldsymbol{\Sigma}_e^{(t)}) \right]^{-1}$$

$$\boldsymbol{\Sigma}_e^{(t+1)} = \mathrm{diag}(\boldsymbol{S}_x - \boldsymbol{W}^{(t+1)}\{\boldsymbol{W}^{(t)}\}^{\mathrm{T}}\{\boldsymbol{\Sigma}_x^{(t)}\}^{-1}\boldsymbol{S}_x)$$

where $\boldsymbol{\Sigma}_x^t = \{\boldsymbol{W}^{(t)}\}^{\mathrm{T}}\boldsymbol{W}^{(t)} + \boldsymbol{\Sigma}_e^{(t)}$,

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{X}_i \boldsymbol{U}_i^{\mathrm{T}} \mid \boldsymbol{X}, \boldsymbol{W}^{(t)}, \boldsymbol{\Sigma}_e^{(t)}) = \boldsymbol{S}_x\{\boldsymbol{\Sigma}_e^{(t)}\}^{-1}\{\boldsymbol{W}^{(t)}\}^{\mathrm{T}},$$

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{U}_i \boldsymbol{U}_i^{\mathrm{T}} \mid \boldsymbol{X}, \boldsymbol{W}^{(t)}, \boldsymbol{\Sigma}_e^{(t)}) = \boldsymbol{W}^{(t)}\{\boldsymbol{\Sigma}_e^{(t)}\}^{-1}\boldsymbol{S}_x\{\boldsymbol{\Sigma}_e^{(t)}\}^{-1}\{\boldsymbol{W}^{(t)}\}^{\mathrm{T}}$$

$$+ \mathbb{I}_K - \boldsymbol{W}^{(t)}\{\boldsymbol{\Sigma}_e^{(t)}\}^{-1}\{\boldsymbol{W}^{(t)}\}^{\mathrm{T}}.$$

The iterative steps stop when $\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^{(t)}\|_F$ and $\|\boldsymbol{\Sigma}_e^{(t+1)} - \boldsymbol{\Sigma}_e^{(t)}\|_F$ are less than certain tolerance value. Let $\boldsymbol{W}^{\dagger}$ and $\boldsymbol{\Sigma}_e^{\dagger}$ to be the estimators at the last step, and $\mathcal{V}$ to be the matrix containing eigenvectors of $p^{-1}\boldsymbol{W}^{\dagger}(\boldsymbol{\Sigma}_e^{\dagger})^{-1}(\boldsymbol{W}^{\dagger})^{\mathrm{T}}$ corresponding to the descending eigenvalues. The maximum likelihood estimators are $\hat{\boldsymbol{W}} = \mathcal{V}^{\mathrm{T}}\boldsymbol{W}^{\dagger}$ and $\hat{\boldsymbol{\Sigma}}_e = \boldsymbol{\Sigma}_e^{\dagger}$. The estimator $\hat{\boldsymbol{S}}_u$ is obtained after estimating $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{\Sigma}}_e$. As $\hat{\boldsymbol{S}}_u$ is not our focus in this method, we omit its derivation details in this paper.

With $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{\Sigma}}_e$, we then estimate $\boldsymbol{U}_i$ using the generalized least squares estimator:

$$\hat{\boldsymbol{U}}_i = (\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\hat{\boldsymbol{W}}^{\mathrm{T}})^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}). \tag{4}$$

Next, we treat the estimators $\hat{\boldsymbol{U}}_1, \ldots, \hat{\boldsymbol{U}}_n$ as surrogate variables for the underlying unmeasured confounders to fit a high-dimensional generalized linear model. We then construct a debiased estimator for the parameter of interest by generalizing the decorrelated score method proposed by Ning and Liu (2017). Other debiasing approaches such as that of van de Geer et al. (2014) could also be similarly developed.

Initial ℓ_1-Penalized Estimator: Recall that $\boldsymbol{\eta} = (\theta, \boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. For vector $\boldsymbol{r} = (r_1, \ldots, r_l)^{\mathrm{T}}$, define $\|\boldsymbol{r}\|_1 = \sum_{j=1}^{l} |r_j|$. To fit a high-dimensional generalized linear model, we solve the ℓ_1-penalized optimization problem

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta} \in \mathbb{R}^{(p+K)}}{\mathrm{argmin}} \left[-\frac{1}{n}\sum_{i=1}^{n}\{y_i(\theta D_i + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}_i + \boldsymbol{\beta}^{\mathrm{T}}\hat{U}_i) - b(\theta D_i + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}_i + \boldsymbol{\beta}^{\mathrm{T}}\hat{U}_i)\} + \lambda(|\theta| + \|\boldsymbol{v}\|_1) \right],$$

where $\hat{\boldsymbol{\eta}} = (\hat{\theta}, \hat{\boldsymbol{v}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ and $b(t)$ is a known function, given a specific generalized linear model. Throughout the manuscript, for notational convenience, let $\boldsymbol{\zeta} = (\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ be regression coefficients for the nuisance covariates and unmeasured confounders. Accordingly, we have $\boldsymbol{\eta} = (\theta, \boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}$, with the goal of performing statistical inference on θ. In the following, we construct a debiased estimator for θ, generalizing the approach in Ning and Liu (2017) to situations involving unmeasured confounders.

A Debiased Estimator: Before we unfold the details of constructing a debiased estimator, we start with introducing some notations. Let $\boldsymbol{I} = E\{b''(\theta D_i + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}_i + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}_i)\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}\}$ be the

Fisher information matrix, and let $\boldsymbol{w}^{\mathrm{T}} = \boldsymbol{I}_{\theta\zeta}\boldsymbol{I}_{\zeta\zeta}^{-1}$, where $\boldsymbol{I}_{\theta\zeta}$ and $\boldsymbol{I}_{\zeta\zeta}$ are corresponding block matrices of $\boldsymbol{I}$. In addition, let $I_{\theta|\zeta} = E\{b''(\theta D_i + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}_i + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}_i)D_i(D_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}_i)\}$ be the partial Fisher information matrix, where $\boldsymbol{M}_i = (\boldsymbol{Q}_i^{\mathrm{T}}, \boldsymbol{U}_i^{\mathrm{T}})^{\mathrm{T}}$ is a vector of nuisance covariates and unmeasured confounders. Finally, let $l(\theta, \boldsymbol{\zeta}) = -n^{-1}\sum_{i=1}^{n}\{y_i(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i) - b(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i)\}$ with $\dot{\boldsymbol{M}}_i = (\boldsymbol{Q}_i^{\mathrm{T}}, \hat{\boldsymbol{U}}_i^{\mathrm{T}})^{\mathrm{T}}$ be the loss function.

We define $S(\theta, \boldsymbol{\zeta}) = \nabla_\theta l(\theta, \boldsymbol{\zeta}) - \boldsymbol{w}^{\mathrm{T}}\nabla_\zeta l(\theta, \boldsymbol{\zeta})$ as the generalized decorrelated score function, where $\nabla_\theta l(\theta, \boldsymbol{\zeta})$ and $\nabla_\zeta l(\theta, \boldsymbol{\zeta})$ are the partial derivatives of the loss function with respect to θ and $\boldsymbol{\zeta}$, respectively. Different from the existing definition of the decorrelated score function in Ning and Liu (2017), the generalized decorrelated score function takes into account the effects induced by the unmeasured confounders. Specifically, in the presence of unmeasured confounders, the generalized decorrelated score function is uncorrelated with the score function corresponding to the nuisance covariates as well as the unmeasured confounders, i.e., $E\{S(\theta, \boldsymbol{\zeta})^{\mathrm{T}}\nabla_\zeta l(\theta, \boldsymbol{\zeta})\} = \boldsymbol{I}_{\theta\zeta} - \boldsymbol{I}_{\theta\zeta}\boldsymbol{I}_{\zeta\zeta}^{-1}\boldsymbol{I}_{\zeta\zeta} = \boldsymbol{0}$. The debiased estimator of θ is constructed by solving for $\tilde{\theta}$ from the first-order approximation of the generalized decorrelated score function $\hat{S}(\hat{\theta}, \hat{\boldsymbol{\zeta}}) + \hat{I}_{\theta|\zeta}(\tilde{\theta} - \hat{\theta}) = 0$. From the first-order approximation equation, we see that to establish the debiased estimator $\tilde{\theta}$, we need to construct two estimators $\hat{S}(\hat{\theta}, \hat{\boldsymbol{\zeta}})$ and $\hat{I}_{\theta|\zeta}$, and the key is to estimate $\boldsymbol{w}$.

We estimate $\boldsymbol{w}$ by solving the following convex optimization problem:

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{(p+K-1)}} \frac{1}{2n}\sum_{i=1}^{n}\{\boldsymbol{w}^{\mathrm{T}}\nabla_{\zeta\zeta}l_i(\hat{\theta}, \hat{\boldsymbol{\zeta}})\boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}}\nabla_{\zeta\theta}l_i(\hat{\theta}, \hat{\boldsymbol{\zeta}})\} + \lambda'\|\boldsymbol{w}\|_1,$$

where $l_i(\theta, \boldsymbol{\zeta}) = -y_i(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i) + b(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i)$ is the ith component of the loss function and $\lambda' > 0$ is a sparsity tuning parameter for $\boldsymbol{w}$. Equivalently, the estimator $\hat{\boldsymbol{w}}$ is obtained by

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{(p+K-1)}} \frac{1}{2n}\sum_{i=1}^{n}b''(\hat{\theta}D_i + \hat{\boldsymbol{\zeta}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i)(D_i - \boldsymbol{w}^{\mathrm{T}}\dot{\boldsymbol{M}}_i)^2 + \lambda'\|\boldsymbol{w}\|_1. \tag{5}$$

The estimator $\hat{\boldsymbol{w}}$ is constructed with the intuition of finding a sparse vector such that the generalized decorrelated score function is approximately zero. This coincides with the intuition to solve for $\tilde{\theta}$ from the first-order approximation of the generalized decorrelated score function. Under the null hypothesis $H_0 : \theta^* = \theta^0$, we estimate the generalized decorrelated score function and the partial Fisher information matrix by

$$\hat{S}(\theta^0, \hat{\boldsymbol{\zeta}}) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i - b'(\theta^0 D_i + \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{Q}_i + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{U}}_i)\}(D_i - \hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i);$$

$$\hat{I}_{\theta|\zeta} = \frac{1}{n}\sum_{i=1}^{n}b''(\hat{\theta}D_i + \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{Q}_i + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{U}}_i)D_i(D_i - \hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i).$$

The debiased estimator can then be constructed as $\tilde{\theta} = \hat{\theta} - (\hat{I}_{\theta|\zeta})^{-1}\hat{S}(\theta^0, \hat{\boldsymbol{\zeta}})$. We will show in Section 4 that the debiased estimator $\tilde{\theta}$ is asymptotically normal. Subsequently, the $(1 - \alpha) \times 100\%$ confidence interval for θ can be constructed as

$$\{\tilde{\theta} - (n\hat{I}_{\theta|\zeta})^{-1/2}\Phi^{-1}(1 - \alpha/2), \ \tilde{\theta} + (n\hat{I}_{\theta|\zeta})^{-1/2}\Phi^{-1}(1 - \alpha/2)\}, \tag{6}$$

where $\Phi(t)$ is the cumulative distribution function for the standard normal random variable.

4. Theoretical Results

Recall that our proposed method yields estimators $\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{w}}$, and the debiased estimator $\tilde{\theta}$. In this section, we first establish upper bounds for the estimation errors of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$ under the ℓ_1 norm. Subsequently, we show that the debiased estimator $\tilde{\theta}$ is asymptotically normal.

For a vector $\boldsymbol{r} = (r_1, \ldots, r_l)^{\mathrm{T}}$, let $\|\boldsymbol{r}\|_q = (\sum_{j=1}^{l} |r_j|^q)^{1/q}$ for $q \geq 1$ and let $\|\boldsymbol{r}\|_{\infty} = \max_{j=1,\ldots,l} |r_j|$. For any matrix $\boldsymbol{A}$, let $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ represent the largest and smallest eigenvalues of $\boldsymbol{A}$, respectively. Moreover, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a sub-exponential random variable Y_1, we write $\|Y_1\|_{\varphi_1} = \inf[s > 0 : E\{\exp(Y_1/s)\} \leq 2]$ as the sub-exponential norm. For a sub-Gaussian random variable Y_2, we write $\|Y_2\|_{\varphi_2} = \inf[s > 0 : E\{\exp(Y_2^2/s^2)\} \leq 2]$ as the sub-Gaussian norm. Throughout the manuscript, we will use an asterisk on the upper subscript to indicate the population parameters. In addition, we define $s_{\eta} = \mathrm{card}\{(j : \eta_j^* \neq 0)\}$ and $s_w = \mathrm{card}\{(j : \boldsymbol{w}_j^* \neq 0)\}$ as the cardinalities of $\boldsymbol{\eta}^*$ and $\boldsymbol{w}^*$, respectively. All of our theoretical analysis are performed under the regime in which n, p, s_{η}, and s_w are allowed to increase, and the number of unmeasured confounders K is fixed.

We start with some conditions on the factor model in (2). Similar conditions were also considered in Bai and Li (2016) in the context of high-dimensional approximate factor model.

Assumption 1 *For some large constant $C > 0$,*

(a) $\mathbb{E}(E_{ij}) = 0$, $\mathbb{E}(E_{ij}^8) \leq C$.

(b) $\mathbb{E}(E_{ih}E_{ij}) = \tau_{i,hj}$ with $|\tau_{i,hj}| \leq \tau_{hj}$ for some $\tau_{hj} > 0$ and all $i = 1, \ldots, n$, and $\sum_{h=1}^{p} \tau_{hj} \leq C$ for all $j = 1, \ldots, p$.

(c) $\mathbb{E}(E_{ij}E_{sj}) = \rho_{is,j}$ with $|\rho_{is,j}| \leq \rho_{is}$ for some $\rho_{is} > 0$ and all $j = 1, \ldots, p$, and $n^{-1} \sum_{i=1}^{n} \sum_{s=1}^{n} \rho_{is} \leq C$.

(d) For all $j, q = 1, \ldots, p$,

$$\mathbb{E}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[E_{ij}E_{iq} - \mathbb{E}(E_{ij}E_{iq})]\right|^4\right\} \leq C.$$

(e) $\|\boldsymbol{W}_j^*\|_2 \leq C$ and σ_j^2 are estimated within the set $[C^{-2}, C^2]$ for all j. For positive definite matrices $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\Upsilon}^*$, $\lim_{p\to\infty} p^{-1}\boldsymbol{W}^*(\boldsymbol{\Sigma}_e^*)^{-1}(\boldsymbol{W}^*)^{\mathrm{T}} = \boldsymbol{\Gamma}^*$ and $\lim_{p\to\infty} p^{-1}\sum_{j=1}^{p}(\sigma_j^*)^{-4}\{(\boldsymbol{W}_j^*)^{\mathrm{T}} \otimes (\boldsymbol{W}_j^*)^{\mathrm{T}}\}(\boldsymbol{W}_j^* \otimes \boldsymbol{W}_j^*) = \boldsymbol{\Upsilon}^*$, where $\otimes$ is the Kronecker product.

Assumption 1 is more general than assumptions in classical factor analysis (Anderson and Amemiya, 1988; Bai, 2003; Fan et al., 2013; Bai and Li, 2016). Instead of constraining all the $\boldsymbol{E}_i$ to have a diagonal covariance matrix, we now only require the higher-order moment of $\boldsymbol{E}_i$ to be bounded, the diagonal entries σ_j^*'s to be bounded, as well as the magnitudes of the correlations among entries of $\boldsymbol{E}_i$ to be controlled in Assumption 1. The conditions are mild as we only control the magnitude of correlations rather than assuming zero correlations as in classical factor analysis. Assumption 1(e) is a regularity condition for restricting the parameters. Overall, Assumption 1 follows standard conditions for the approximate factor model in Bai and Li (2016) and is required for the estimation consistency of the unmeasured confounders.

Comparing Assumption 1 to the dense confounding assumption (A2) imposed on the linear framework in Guo et al. (2022), our assumption is mild as it holds for a broad regime of n and p. Specifically, the dense confounding assumption (A2) in Guo et al. (2022) is related to our assumption 1(e) that $\lim_{p \to \infty} p^{-1} \boldsymbol{W}^* (\boldsymbol{\Sigma}_e^*)^{-1} (\boldsymbol{W}^*)^{\mathrm{T}} = \boldsymbol{\Gamma}^*$ where $\boldsymbol{\Gamma}^*$ is positive definite. Our assumption follows from classical factor analysis literature and as a result, we have $\lambda_q(\boldsymbol{W}^*) \asymp \sqrt{p}$, where $\lambda_q(\boldsymbol{W}^*)$ is the q-th singular value of the factor loading matrix $\boldsymbol{W}^*$. With $\boldsymbol{\gamma}^{\mathrm{T}} = (\theta, \boldsymbol{v}^{\mathrm{T}})$ being the coefficient for all covariates and γ_j being the individual coefficient of interest, the dense confounding assumption mainly requires that

$$\lambda_q(\boldsymbol{W}_{-j}^*) \gg \max \left\{ M\sqrt{Kpn^{-1}}(\log p)^{3/4}, \sqrt{MK}p^{1/4}(\log p)^{3/8}, \sqrt{Kn \log p} \right\},$$

where $\boldsymbol{W}_{-j}^*$ denotes the factor loading matrix $\boldsymbol{W}$ with j-th column removed. Guo et al. (2022) focuses on the setting $p \gg n$ and they point out that in the high-dimensional regime and under certain settings, the dense confounding assumption holds with high probability. Specifically, when the entries of $\boldsymbol{W}$ are i.i.d. Sub-Gaussian with zero mean and variance σ_w^2, it holds that $\lambda_q(\boldsymbol{W}_{-j}) \asymp \sqrt{p}\sigma_w$ and the dense confounding assumption requires $p \gg Kn \log p$ and $\min\{n, p\} \gg K^3 (\log p)^{3/2} M^2$ to make σ_w diminish to zero. However, this condition is restricted to the high-dimensional regime and may not hold when p is of relatively lower order.

Moreover, without loss of generality, we assume a working identifiability condition: $\boldsymbol{S}_u = \boldsymbol{I}_K$ and $p^{-1} \boldsymbol{W}^* (\boldsymbol{\Sigma}_e^*)^{-1} (\boldsymbol{W}^*)^{\mathrm{T}}$ is a diagonal matrix with distinct entries. Note that the aforementioned working identifiability condition is for presentation purpose only and not an assumption on the model structure. As presented in Appendix B, when such a working identifiability condition is not satisfied, our theoretical results in Theorems 2 and 3 are still valid. We also illustrate this via simulation in Section 5.1. Next, we impose some assumptions on the generalized linear model with unmeasured confounders in (1).

Assumption 2 *(a) The Fisher information $\boldsymbol{I}^* = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\} \boldsymbol{Z}_i \boldsymbol{Z}_i^{\mathrm{T}}]$ satisfies $\lambda_{\min}(\boldsymbol{I}^*) \geq \kappa$, where $\kappa > 0$ is some constant.*

(b) For some constant $M > 0$, $\|\boldsymbol{X}_i\|_\infty \leq M$, $\|\boldsymbol{U}_i\|_\infty \leq M$, $\|\boldsymbol{\eta}^*\|_\infty \leq M$ and $|(\boldsymbol{w}_q^*)^{\mathrm{T}} \boldsymbol{Q}_i| \leq M$, where $\boldsymbol{w}_q^* = (w_2^*, \ldots, w_p^*)^{\mathrm{T}}$.

(c) The term $|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}|$ is sub-exponential with $\|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}\|_{\varphi_1} \leq M$.

(d) Assume that $a_1 \leq (\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i \leq a_2$ and $0 \leq |b'(t)| \leq B$ with $|b'(t_1) - b'(t)| \leq B|(t_1 - t)b'(t)|$ and $0 \leq b''(t) \leq B$ with $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ for constants a_1, a_2 and B, where $t \in [a_1 - \epsilon, a_2 + \epsilon]$ for $\epsilon > 0$ and sequence t_1 satisfies $|t_1 - t| = o(1)$.

In the absence of unmeasured confounders, similar conditions in Assumption 2 can be implied from the conditions in Theorem 3.3 in van de Geer et al. (2014) and Assumption E.1 in Ning and Liu (2017). When $\boldsymbol{U}_i$ and $\boldsymbol{X}_i$ are binary or categorical, Assumption 2(b) holds with a constant $M > 0$. When $\boldsymbol{U}_i$ and $\boldsymbol{X}_i$ are sub-exponential random vectors, Assumption 2(b) holds with $M = cn^{-1/2}(\log p)^{1/2}$ for some constant $c > 0$, with probability at least $1 - p^{-1}$. Assumption 2(d) imposes mild regularity conditions on the function $b(t)$, and is commonly used in analyzing high-dimensional generalized linear models without unmeasured confounders. We require the function $b(t)$ to be at least twice differentiable and $b'(t)$ and $b''(t)$ to be bounded. Specifically, $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ can be implied by $|b'''(t)| \leq Bb''(t)$ when $b'''(t)$ exists, which is a weaker condition than the self-concordance

property (Bach, 2010). This boundary assumption is important for the concentration of the Hessian matrix of the loss function. Assumption 2(d) holds for commonly used generalized linear models. For example, for logistic regression where $b(t) = \log\{1 + \exp(t)\}$, Assumption 2(d) holds with $B = 1$, and a_1, a_2 extended to infinity and it can be similarly verified to hold at $B = \max\{1, \exp(a_2 + \epsilon)\}$ for Poisson regression and $B = \max\{1/(a_2 + \epsilon)^2, 2/|a_2 + \epsilon|\}$ for exponential regression with $a_2 < 0$. For linear model, Assumption 2 can be relaxed as stated in Remark 4.

The theoretical analysis on the unmeasured confounders estimator is important in establishing the theoretical guarantee for our debiased method. As the decomposition techniques commonly used in linear models may not be applicable in generalized linear model settings, it is necessary to establish more general and refined intermediate results as the foundation of our theoretical analysis (see Remark 8 in Appendix D.1 for more details). We first present a uniform convergence result for the estimators of the unmeasured confounders.

Proposition 1 *Under Assumptions 1–2, if $n, p \to \infty$, we have*

$$\max_{i \in \{1,\ldots,n\}} \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i^*\|_\infty = O_p\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right).$$

From Proposition 1, the estimator of unmeasured confounders $\hat{\boldsymbol{U}}_i$ uniformly converges to $\boldsymbol{U}_i$ at $p \gg \log n$, which holds naturally under the high-dimensional regime $p \gg n$. Moreover, the convergence rate $O_p(n^{-1/2} + p^{-1/2} (\log n)^{1/2})$ is of a similar order to the convergence rates of principal component estimators (Fan et al., 2013; Wang and Fan, 2017). The estimation results of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$ are dependent on the accuracy of the estimator of unmeasured confounders and we next establish upper bounds on the estimation errors for $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$.

Theorem 2 (Estimation consistency) *Suppose that Assumptions 1–2 hold. With $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$, if the scaling condition $n, p \to \infty$ and $(s_w \vee s_\eta)\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\} = o_p(1)$ hold, then we have*

$$\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 = O_p\left\{s_\eta\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right)\right\};$$

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 = O_p\left\{(s_w \vee s_\eta)\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right)\right\}.$$

The effect of unmeasured confounders enters the rate of the estimation error in our results through the term $p^{-1/2}(\log n)^{1/2}$. Under the high-dimensional setting with $p \gg n$, the unmeasured confounders effect is dominated by $n^{-1/2}(\log p)^{1/2}$ and as a result, the convergence rates in Theorem 2 are of the same order as in the oracle case when the unmeasured confounders are assumed to be known (van de Geer et al., 2014; Ning and Liu, 2017). With the estimation consistency established, we show that the debiased estimator from the proposed estimation method is asymptotically normal, and thus valid confidence intervals can be constructed.

Theorem 3 (Asymptotic normality) *Suppose that Assumptions 1–2 hold. With $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$, if the scaling condition $n, p \to \infty$ and $(s_w \vee s_\eta)$ $(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$ hold, then*

$$n^{1/2}(I^*_{\theta|\zeta})^{1/2}(\tilde{\theta} - \theta^*) \to_d N(0, 1). \tag{7}$$

The result in Theorem 3 is similar to existing inference results for high-dimensional generalized linear models without unmeasured confounders (van de Geer et al., 2014; Ning and Liu, 2017; Ma et al., 2021; Shi et al., 2021; Cai et al., 2023). Theorem 3 is established under the condition that the estimation errors of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$ in Theorem 2 are $o_p(1)$. As a consequence of the asymptotic normality result in Theorem 3 and that $\hat{I}_{\theta|\zeta}$ is consistent estimator for $I^*_{\theta|\zeta}$, the construction of the confidence interval in (6) is valid.

Remark 4 *For the linear model, estimation consistency and asymptotic normality results in Theorems 2 and 3 hold with less stringent conditions than that in Assumption 2. Suppose Assumption 1 and Assumption 2(a) hold and assume that the random noise ε_i is independent sub-Gaussian random variable and $\boldsymbol{Z}_i$ is sub-Gaussian vector such that $\|\varepsilon_i\|_{\varphi_2} \leq M$ and $\|Z_{ij}\|_{\varphi_2} \leq M$ for some constant $M > 0$. By choosing $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$, if $n, p \to \infty$ and $(s_w \vee s_\eta)(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, the estimation error bounds in Theorem 2 hold and the debiased estimator $\tilde{\theta}$ is asymptotically normal with limiting distribution (7). Similar assumptions are commonly imposed in many existing inference methods for high-dimensional linear models without unmeasured confounders (Zhang and Zhang, 2014; Javanmard and Montanari, 2014) as well as in the existence of confounders (Ćevid et al., 2020; Guo et al., 2022).*

Remark 5 *The sparsity assumption on $\boldsymbol{w}$ is a standard assumption in high-dimensional regression models without unmeasured confounders. It may be more suitable and even weaker in our proposed framework with unmeasured confounders. In high-dimensional regression models without unmeasured confounders, the sparsity assumption on $\boldsymbol{w}$ is implied by the sparse inverse population Hessian condition (van de Geer et al., 2014; Belloni et al., 2016; Ning and Liu, 2017; Jankova and van de Geer, 2018). Specifically, the sparse inverse population Hessian condition implies the sparsity of the coefficient parameter $\boldsymbol{w}$ in a weighted node-wise lasso regression $D_i \sim \boldsymbol{Q}_i$, which is similar to (5) but regresses the covariate of interest D_i on the rest of the covariates $\boldsymbol{Q}_i$.*

In our proposed setting with unmeasured confounders, as shown from (5), we assume that in the weighted node-wise lasso regression $D_i \sim \boldsymbol{Q}_i + \boldsymbol{U}_i$, the coefficient of $\boldsymbol{Q}_i$, denoted as $\boldsymbol{w}_q$, to be sparse. That is, $\boldsymbol{w}_q$ is sparse, conditional on the unmeasured confounders $\boldsymbol{U}_i$. The sparsity assumption is mild and can be satisfied in many settings. For example, under the assumptions in classical factor analysis in which the covariance matrix of the random error $\boldsymbol{E}_i$ is diagonal, D_i is uncorrelated with the other covariates $\boldsymbol{Q}_i$, conditioned on the unmeasured confounders $\boldsymbol{U}_i$. In this case, we have $\boldsymbol{w}_q = \boldsymbol{0}$, conditioned on $\boldsymbol{U}_i$ and thus the sparsity assumption holds naturally. We also want to point out that the imposed sparsity assumption may be weaker than that of existing work on high-dimensional inference without unmeasured confounders, where the coefficients of regression $D_i \sim \boldsymbol{Q}_i$ are assumed to be sparse. For instance, we allow D_i to be densely correlated with $\boldsymbol{Q}_i$ marginally. In other words, we only require $\boldsymbol{w}_q$, the coefficient for $\boldsymbol{Q}_i$ in $D_i \sim \boldsymbol{Q}_i + \boldsymbol{U}_i$, to be sparse, conditional on the confounders $\boldsymbol{U}_i$ while marginally the coefficients of $D_i \sim \boldsymbol{Q}_i$ could be dense.

Remark 6 *For the simplicity of theoretical analysis, we assume the dimension of unmeasured confounders K to be fixed and known, which is also assumed in Wang and Fan (2017), Fan et al. (2023), Wang (2022) and many other works. Nevertheless, our theoretical results hold as long as K is consistently estimated. As introduced in Section 3, we use parallel analysis to estimate the dimension of unmeasured confounders in practice. Theoretically, it has been shown that parallel analysis consistently selects the dimension of unmeasured confounders in factor analysis (Dobriban, 2020). Specifically, when the dimension p is relatively large compared to the sample size n, each factor loads on not too few variables, and the signal size of unmeasured confounders is not too large, parallel analysis selects the number of factors with probability tending to one (Dobriban, 2020). These conditions can be satisfied under our framework, implying that the dimension of unmeasured confounders is consistently determined. Moreover, empirically we find that the proposed method still provides satisfying inference results under some overestimation of K. Intuitively, as long as the corresponding linear combinations of the true underlying factors $\boldsymbol{U}$ in the considered models can be well approximated by those of the estimated $\hat{\boldsymbol{U}}$, the developed inference results for θ^* would still hold. To further illustrate this, we perform simulation studies in Appendix C to show that the overestimation of K may not affect the asymptotic properties of the debiased estimator.*

5. Numerical Studies

To illustrate the performance of our proposed method, we conduct numerical experiments including simulation studies and an application of our method to a genetic data set.

5.1 Simulations

We consider two different models in (1), a linear regression model and a logistic regression model. We compare our method with two alternative approaches for performing high-dimensional inference: the oracle method where we perform the debiasing approach assuming that the true values of unmeasured confounders are known; and the naive method in which the unmeasured confounders are neglected in the estimation and we perform the debiasing approach with the observed covariates only. In addition, for the linear regression setting, we compare the proposed method with the doubly debiased lasso method proposed by Guo et al. (2022). The sparsity tuning parameters for all of the aforementioned methods are selected using 10-fold cross-validation. Our proposed method involves estimating the dimension of the unmeasured confounders, which we estimate using the parallel analysis (Horn, 1965; Dinno, 2009). To evaluate the performance across the different methods, we construct the 95% confidence intervals for the parameter of interest, and compute the average confidence length and the coverage probability of the true parameter over 300 independent replications.

First, we generate each entry of the unmeasured confounders $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n \in \mathbb{R}^3$ and each entry of the random error $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_n \in \mathbb{R}^p$ from a standard normal distribution. We set

$$(\boldsymbol{W}^*)^T = \begin{pmatrix} \mathbf{0.5}_{p/3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{p/3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1.5}_{p/3} \end{pmatrix}_{p \times 3},$$

13

where $\mathbf{0.5}_{p/3}$, $\mathbf{1}_{p/3}$, and $\mathbf{1.5}_{p/3}$ are vectors of dimension $p/3$ with all entries equal 0.5, 1, and 1.5, respectively (in our simulation, $p/3$ is set as integer values). The $\mathbf{0}$ in the matrix denotes a vector of dimension $p/3$ with all entries equal 0. We then generate the covariates $\boldsymbol{X}_i = (\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{U}_i + \boldsymbol{E}_i$ with $D_i = \boldsymbol{X}_{i,1}$ and $\boldsymbol{Q}_i = \boldsymbol{X}_{i,-1}$. It can be verified that the above setting satisfies the working identifiability condition described in Section 4. Later in this section, we also perform additional simulation studies as to illustrate the validity of the theory when the working identifiability condition does not hold.

We consider both the linear regression model and the logistic regression model: for linear regression, we generate the response y_i according to $y_i = \theta^* D_i + (\boldsymbol{v}^*)^{\mathrm{T}}\boldsymbol{Q}_i + (\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{U}_i + \varepsilon_i$, where the random noise ε_i is generated from a standard normal distribution; for logistic regression, the response y_i is generated from a Bernoulli distribution with probability $p_i(\theta^*, \boldsymbol{v}^*, \boldsymbol{\beta}^*) = 1/(1 + \exp[-\{\theta^* D_i + (\boldsymbol{v}^*)^{\mathrm{T}}\boldsymbol{Q}_i + (\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{U}_i\}])$. The regression coefficients for the unmeasured confounders, parameter of interest, and nuisance parameters are set to be $\boldsymbol{\beta}^* = (1,1,1)^{\mathrm{T}}$, $\theta^* = 0$, and $\boldsymbol{v}^* = (1, \mathbf{0}_{p-2})^{\mathrm{T}}$, respectively.

Results for the cases when $n = 500$ and p varies from 100 to 1500 and when $p = 600$ and n varies from 100 to 1500 under both linear regression model and logistic regression model are presented in Fig. 1 and Fig. 2, respectively. From Fig. 1, we see that the naive method suffers from undercoverage since the effects induced by the unmeasured confounders are not taken into account. Our proposed method has coverage at approximately 0.95 and is similar to that of the oracle method. The proposed method also has similar length of the confidence intervals as that of the oracle method. As a comparison, the doubly debiased lasso method suffers from undercoverage when p is relatively small ($p < 100$ and $n = 500$): our finding is consistent with that of Guo et al. (2022), where they indicate that when p is relatively small to n, bias from the hidden confounding effects can be relatively large, which leads to undercoverage of their method. As p increases, we see that coverage for the doubly debiased lasso method approaches to 0.95, but its confidence interval lengths are larger than that of our proposed method, indicating our method is preferred in this case.

Under logistic regression model, our proposed method performs similarly to the oracle method in terms of both coverage and length of the confidence intervals. Results for doubly debiased lasso are not reported as it is not directly applicable to generalized linear model.

In the remaining subsection, we show that when the working identifiability condition fails, our proposed method can still perform well. More discussions about the working identifiability condition are in Appendix B. Specifically, we set $W_{jk}^* \sim \mathrm{Unif}[0,1]$ for $j = 1, \dots, p$ and $k = 1, 2$, and 3. This loading matrix implies that each covariate vector $\boldsymbol{X}_i$ is related to all the confounders. The identifiability condition fails in such case as $p^{-1}\boldsymbol{W}^*(\boldsymbol{\Sigma}_e^*)^{-1}(\boldsymbol{W}^*)^{\mathrm{T}}$ is not a diagonal matrix. We continue with the parameter setting at the start of this section and the data generation process is unchanged except for replacing the original loading matrix with the new loading matrix where elements are uniformly distributed.

At varying dimensions n and p, we construct the confidence intervals using each method over 300 simulations, and then compute the coverage probabilities of the 95% confidence intervals on the true parameter and the average confidence interval lengths. Under the new loading matrix, the results for linear regression model are shown in Fig. 3 and the results for logistic regression model are shown in Fig. 4. We observe the coverage rates of our proposed method can still achieve the desirable 0.95 level and are close to the oracle case. The naive method performs worse than the proposed method, with coverage rates much less than the
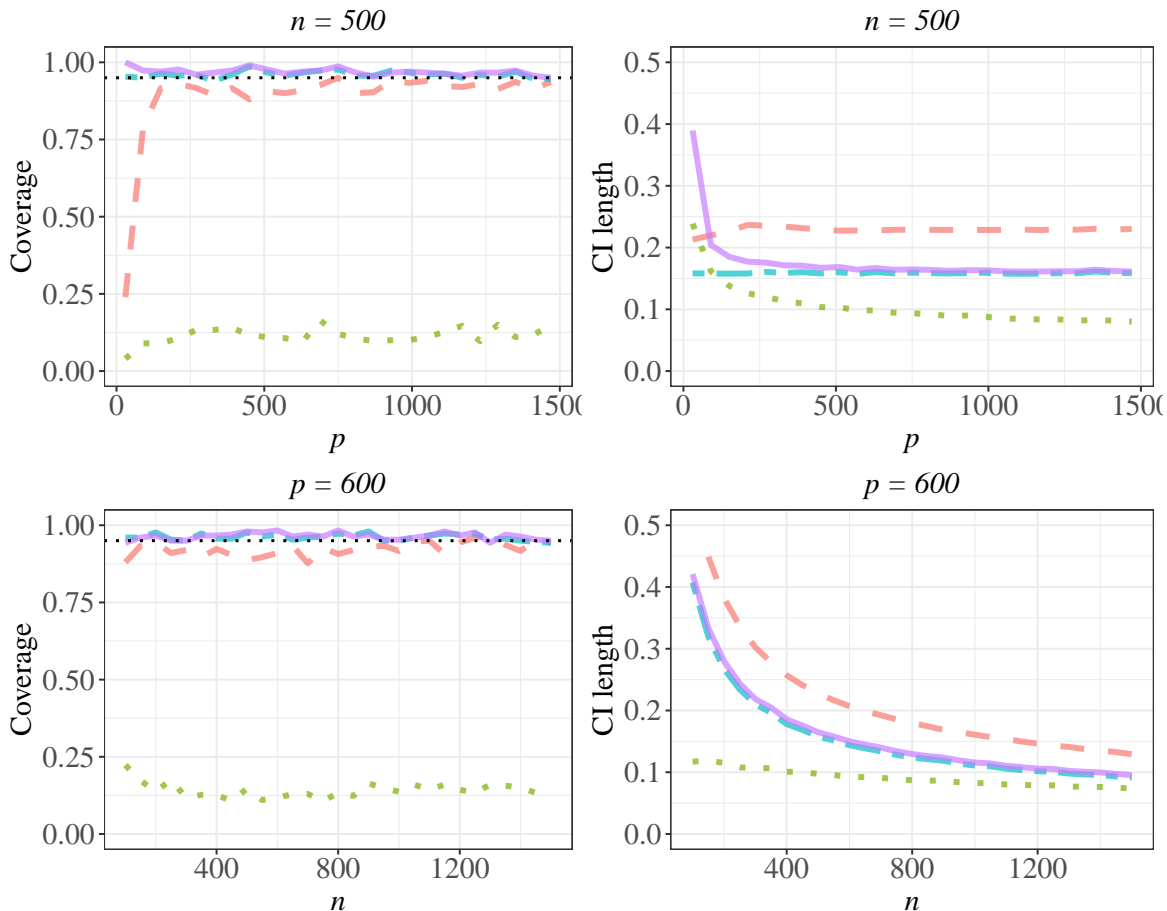
Figure 1: Coverage and length of the confidence interval under linear regression, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dashlines indicate the 0.95 level. Purple solid lines (——) are our proposed method. Blue two-dashed lines (- —) represent the oracle case. Green dotted lines (· · ·) indicate the naive method. Orange dashed lines (— —) represent the doubly debiased lasso method.

0.95 level for most cases. For the linear regression results, although a little under coverage in some cases, the doubly debiased lasso method exhibits good performance in this setting with coverage rates approximating to the oracle case, while its average confidence interval length is larger than that of the proposed method.

5.2 Data Application

In this section, we apply the proposed method to a genetic data containing gene expression quantifications and stimulation statuses in mouse bone marrow derived dendritic cells. The data were also previously analyzed in Shalek et al. (2014) and Cai et al. (2023). Specifically,
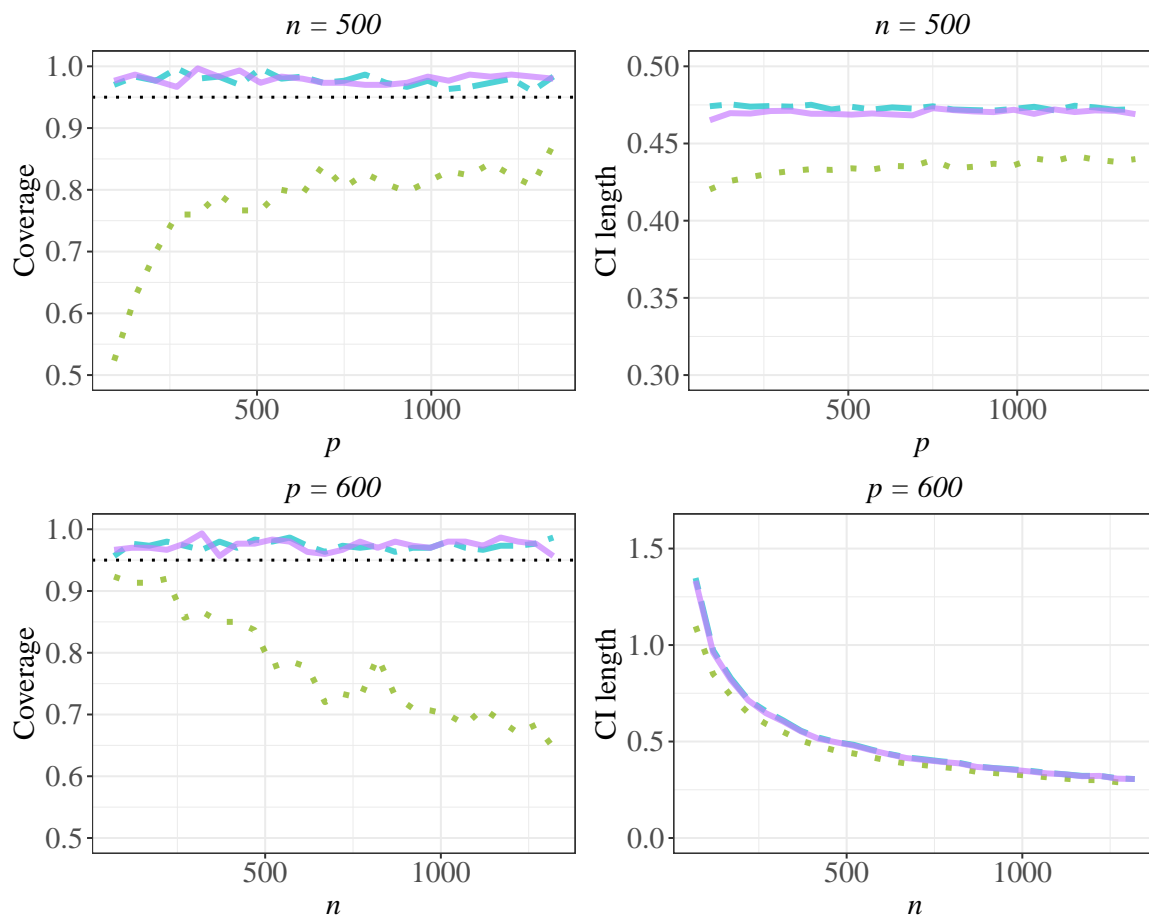
Figure 2: Coverage and length of the confidence interval under logistic regression, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dashlines indicate the 0.95 level. Purple solid lines (——) are our proposed method. Blue two-dashed lines (- —) represent the oracle case. Green dotted lines (· · ·) indicate the naive method.

Cai et al. (2023) aimed to find significant gene expressions in response to three stimulations, which are (a) PAM, a synthetic mimic of bacterial lipopeptides; (b) PIC, a viralike ribonucleic acid; and (c) LPS, a component of bacteria. However, there may be potential unmeasured confounders that may lead to spurious discoveries. Our goal is to investigate the relationship between gene expressions levels and the different stimulations while accounting for possible unmeasured confounders.

We start with pre-processing the data: we consider expression profiles after six hours of stimulations and perform genes filtering, expression level transformation, and normalization. The details can be found in Cai et al. (2023). After the pre-processing steps, we have three groups of cells including 64 PAM stimulated cells, 96 PIC stimulated cells, and 96 LPS stimulated cells, respectively. Moreover, each of the three groups contains 96 control cells
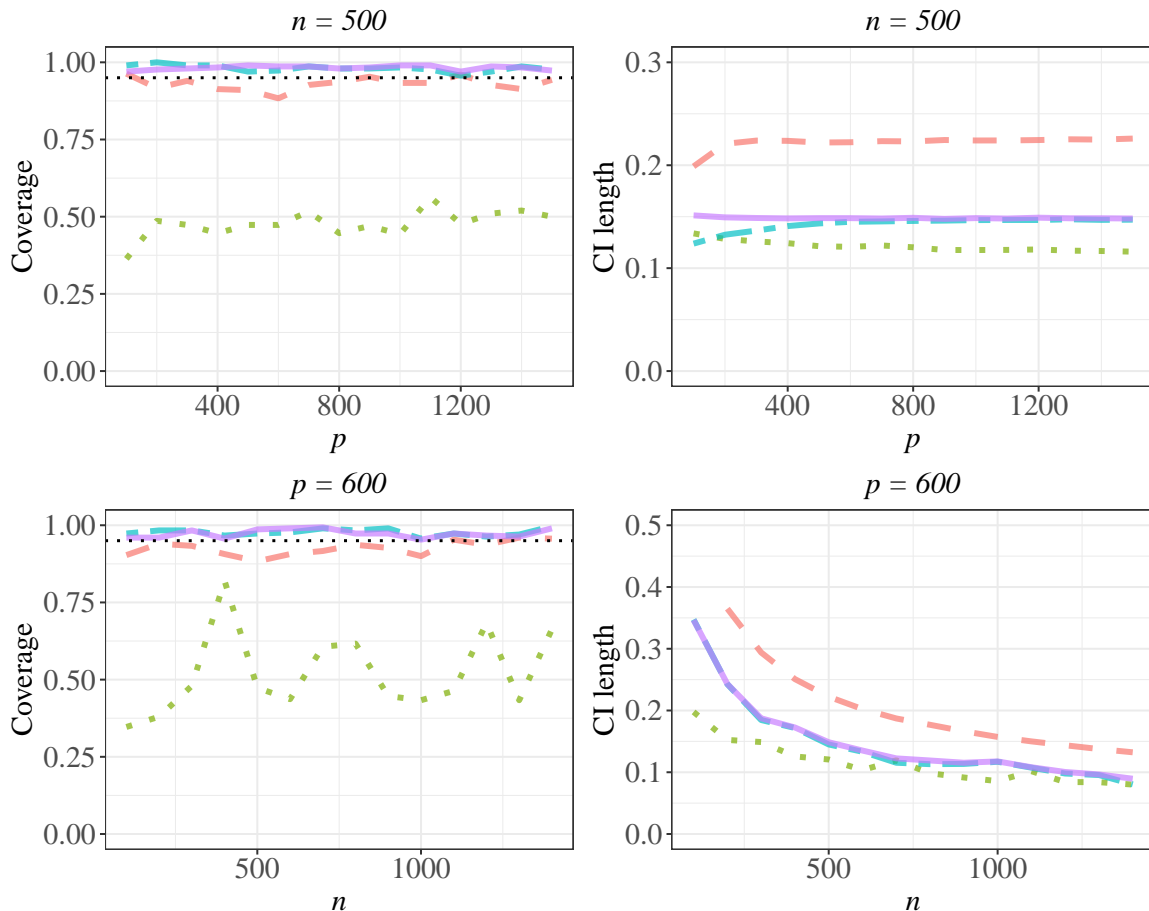
Figure 3: Coverage and length of the confidence interval under linear regression with new loading matrix, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dashlines indicate the 0.95 level. Purple solid lines (——) are our proposed method. Blue two-dashed lines (-—·) represent the oracle case. Green dotted lines (· · ·) indicate the naive method. Orange dashed lines(- -) represent the doubly debiased lasso method.

without any stimulation. The gene expression quantifications are computed on 768 genes for PAM stimulation group, 697 genes on PIC stimulation group, and 798 genes for LPS stimulation group. For each group of cells, the high-dimensional covariates are the gene expression levels, the stimulation status is a binary response variable recording whether a cell is stimulated.

We fit the data using our proposed method and the naive method in which we perform the debiasing approach without adjusting for unmeasured confounders. Applying each method, we construct 95% confidence intervals for the regression coefficient for every gene. In addition, we compute the p-value and effect size for each gene using each of two methods, respectively. Our proposed method found that there are 7 possible confounders for the PAM
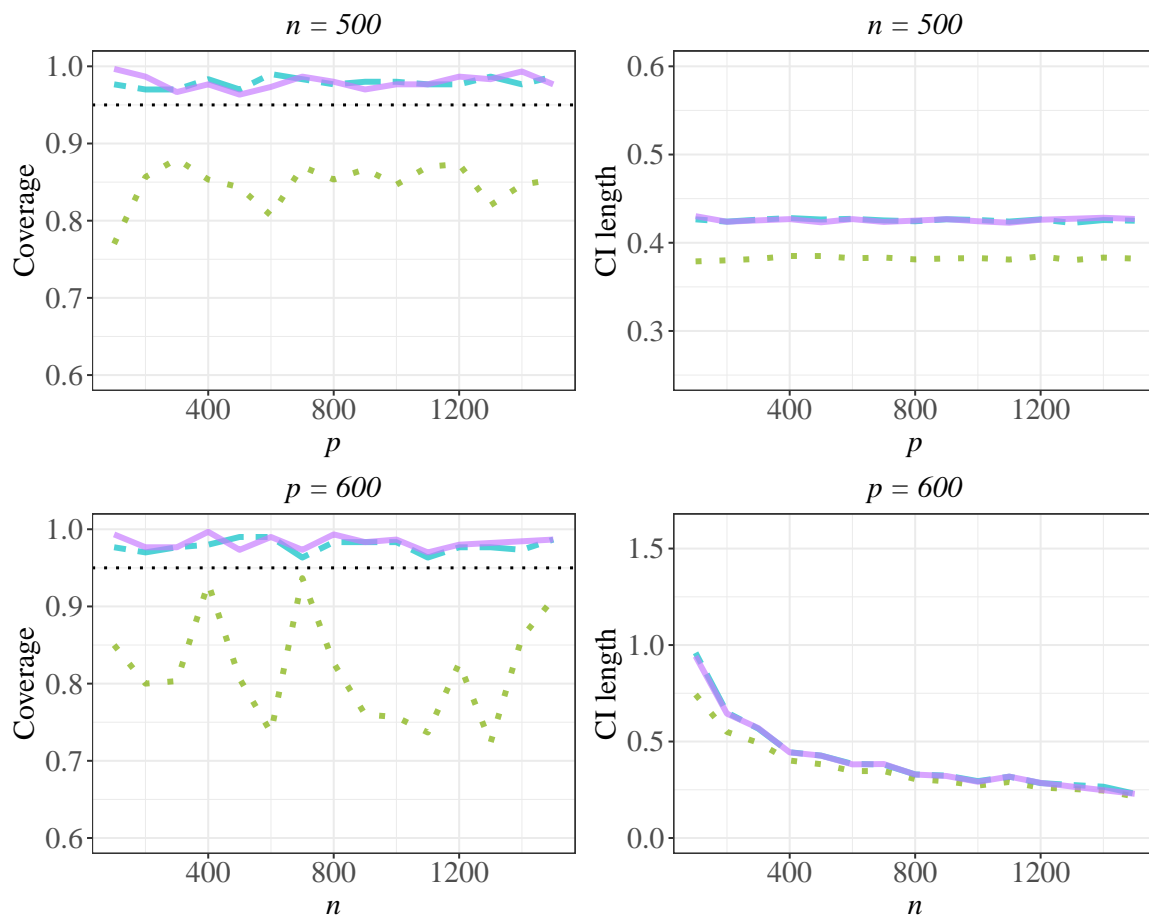
Figure 4: Coverage and length of confidence interval under logistic regression new loading matrix, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dashlines indicate the 0.95 level. Purple solid lines (——) are our proposed method. Blue two-dashed lines (‑ —) represent the oracle case. Green dotted lines (· · ·) indicate the naive method.

and PIC stimulation groups, and 9 possible confounders for the LPS stimulation group. The 95% confidence intervals constructed with our proposed method for each of the three groups are shown in Fig. 5.

We see from Fig. 5 that the confidence intervals for some genes do not contain zero, suggesting that they are possibly associated with their respective stimulations. For valid statistical inference, we perform a Bonferroni correction to adjust for multiple hypothesis testing. As a comparison, we perform similar procedures using the confidence intervals constructed with the naive method to identify the significant genes. Genes that are significantly associated with the stimulation after Bonferroni correction from the proposed method and naive method respectively are reported in Table 1. The small p-values and large effect sizes
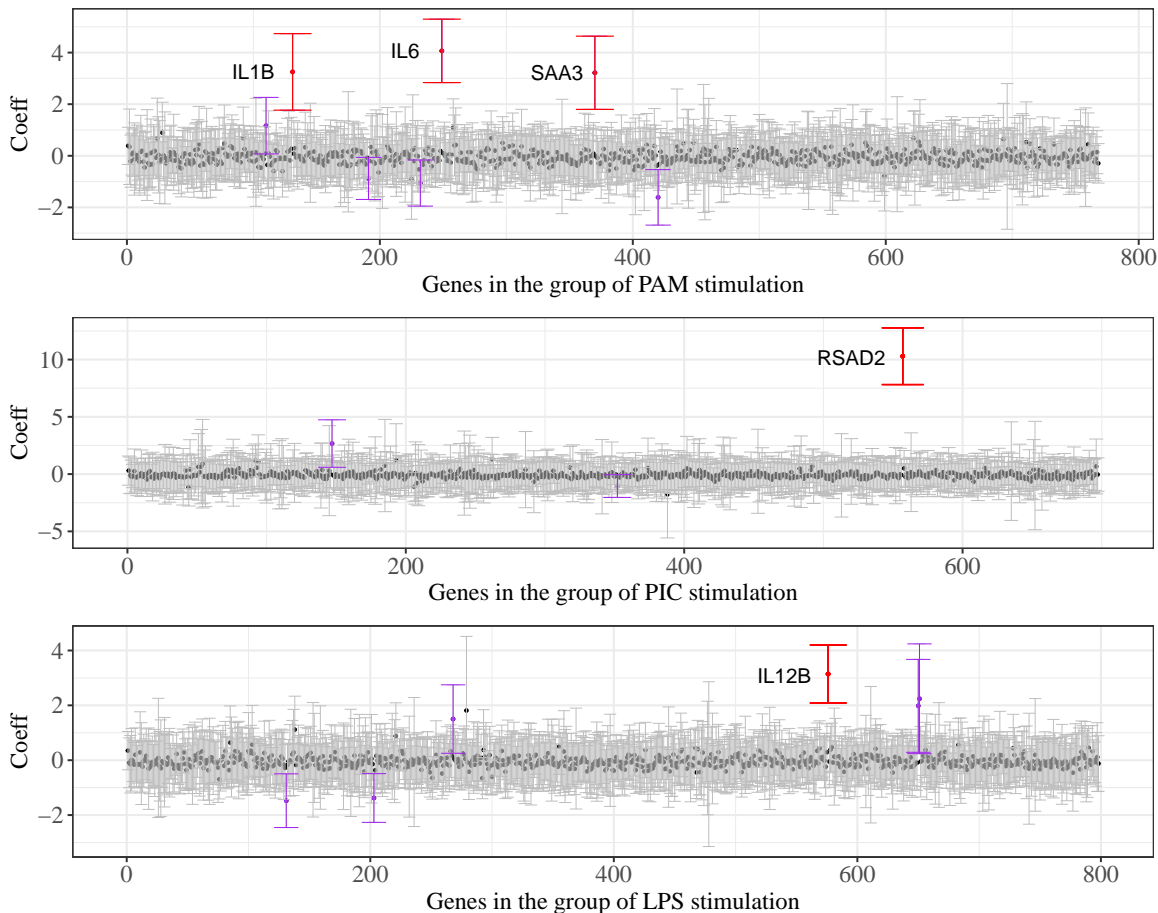
Figure 5: Confidence interval estimation using proposed method under logistic regression for the stimulation groups of PAM (top), PIC (middle) and LPS (bottom). Purple intervals indicate confidence intervals that do not cover zero. Red intervals represent confidence intervals that are among purple intervals and significant after Bonferroni Correction.

indicate that the corresponding genes are strongly associated with their respective stimulations. Some gene codes in Table 1 coincide with the findings in Cai et al. (2023), and their functional consequences to the stimulation are also supported in existing literature. Specifically, Cai et al. (2023) suggested that there exist significant association for "IL6" with PAM stimulation, "RSAD2" with PIC stimulation, and "CXCL10" and "IL12B" with LPS stimulation. Applying our proposed method in which the effects of unmeasured confounders are considered, we identify that "IL1B", "IL6", and "SAA3" are significant genes for PAM stimulated cells, "RSAD2" is significant for PIC stimulated cells, and "IL12B" is significant for LPS stimulated cells. On the other hand, when we apply the naive method which does not adjust for unmeasured confounders, we identify significant genes "IL1B", "IL6" for PAM stimulated cells, "RSAD2" for PIC stimulated cells and "IL12B" for LPS

stimulated cells. Comparing the proposed method with the naive method, our method that adjusts for unmeasured confounders identifies an additional gene "SAA3". In the genetics literature, experimental studies support the finding that "SAA3" plays important roles in immune reactions (Ather and Poynter, 2018). For instance, mice lacking the gene "SAA3" develop metabolic dysfunction along with defects in innate immune development (Ather and Poynter, 2018). This comparison suggests that our proposed method can identify significant genes/variables that are not captured by existing debiasing procedures without for adjusting unmeasured confounders.

Method	Stimulation	Gene code	p-value	Effect size
Proposed method	PAM	IL1B	$1\cdot743 \times10^{-5}$	$4\cdot295$
		IL6	$8\cdot940\times10^{-11}$	$6\cdot484$
		SAA3	$8\cdot800\times10^{-6}$	$4\cdot445$
	PIC	RSAD2	$4\cdot441\times10^{-16}$	$8\cdot144$
	LPS	IL12B	$5\cdot180\times10^{-9}$	$5\cdot841$
Naive method	PAM	IL1B	$6\cdot908\times10^{-7}$	$4\cdot963$
		IL6	$2\cdot745\times10^{-12}$	$6\cdot990$
	PIC	RSAD2	$< 1 \times 10^{-16}$	$8\cdot604$
	LPS	IL12B	$4\cdot187\times10^{-8}$	$5\cdot482$

Table 1: Significant gene expression associated with stimulations after Bonferroni correction estimated using the proposed method and naive method, respectively.

Moreover, many genetic studies support that in addition to the "SAA3", genes discovered by our proposed method all play important roles in the immune response. The gene "IL1B" encodes protein in the family of interleukin 1 cytokine, which is an important mediator of the inflammatory response and its induction can contribute to inflammatory pain hypersensitivity (Nemetz et al., 1999). The genes "IL6" and "IL12B" are also known to encode cytokines that play key roles in hosting defense through stimulation and immune reactions (Müller-Berghaus et al., 2004; Toyoshima et al., 2019). The expression of the gene "RSAD2" is important in antiviral innate immune responses and "RSAD2" is also a powerful stimulator of adaptive immune response mediated via mature dendritic cells (Jang et al., 2018).

6. Discussion

This manuscript studies statistical inference problem for the high-dimensional generalized linear model under the presence of hidden confounding bias. We propose a debiasing approach to construct a consistent estimator of the individual coefficient of interest and its corresponding confidence intervals, which generalizes the existing debiasing approach to account for the effects induced by the unmeasured confounders. Theoretical properties were also established for the proposed procedure.

Our goal of this paper is to conduct statistical inference for individual coefficients θ. The purpose of adjusting for the unmeasured confounders in the proposed method is to use the latent information for the better inference of covariates coefficient θ, while the

inference for the coefficient of unmeasured confounders $\boldsymbol{\beta}$ is not the focus of the current work. This problem setting has wide scientific applications. For instance, as introduced in our data application, biologists are interested in the significance of the genes under the simulations with the unmeasured confounders adjusted. Moreover, in many practices with finite samples, consistent estimation of the number of factors may be problematic; our method only requires that the estimated factors can well approximate the confounders and the interpretation of the factors is of less interest in this setting. Here to investigate how the accuracy of the estimation of the dimension of unmeasured confounders K affects the asymptotic distribution of debiased estimator, we conducted a simulation study in Appendix C and the results suggest that in practice, the overestimation of K may not affect the asymptotic normality results, which may be because an abundant amount of unmeasured confounders can still fully capture the covariate information and would not incur bias in the point and interval estimation.

Nevertheless, the inference of unmeasured confounders is also of great interest in many econometrics and psychometrics applications. For instance, Fan et al. (2023) recently proposed an ANOVA-type testing procedure for testing the existence of unmeasured confounders or not. It would be interesting to develop similar inference testing procedures under the generalized linear models in applications when the inference of the unmeasured confounders of scientific interest.

Another interesting related problem is regarding the inference on group-wise covariate coefficients. We next briefly describe how to generalize our method to obtain the group-wise asymptotic distribution of $\max_{j \in G} |\gamma_j|$ via a bootstrap-assisted procedure for $\boldsymbol{\gamma} = (\theta, \boldsymbol{v}^{\mathrm{T}}) \in \mathbb{R}^p$ and any subset $G \subseteq \{1, \ldots, p\}$. The procedures follow simultaneous inference results from Zhang and Cheng (2017) and are shown as follows.

- Step 1: For each γ_j, we construct a debiased estimator $\tilde{\gamma}_j$ using our proposed method for $j \in G$. We generate a sequence of i.i.d. standard normal random variables and denote them as $\{\varpi_i\}_{i=1,\ldots,n}$.

- Step 2: Then, under the null hypothesis that $H_{0,G} : \gamma_j^* = \gamma_j^0$ for $j \in G$, let $W_G = \max_{j \in G} |\sum_{i=1}^n \hat{I}_{\gamma_j|\boldsymbol{\eta}_{-j}}^{-1} \hat{S}(\gamma_j^0, \hat{\boldsymbol{\eta}}_{-j}) \varpi_i / \sqrt{n}|$, where $\hat{S}(\gamma_j^0, \hat{\boldsymbol{\eta}}_{-j})$ can be similarly calculated as $\hat{S}(\theta^0, \hat{\boldsymbol{\zeta}})$ and $\hat{I}_{\gamma_j|\boldsymbol{\eta}_{-j}}$ can be similarly calculated as $\hat{I}_{\theta|\boldsymbol{\zeta}}$ in Section 3 by treating γ_j as the coefficient of interest and $\boldsymbol{\eta}_{-j}$ as the other coefficient for nuisance covariates and unmeasured confounders.

- Step 3: Next, we calculate the critical value $C_G(\alpha)$ at $(1 - \alpha)$ significance level by $C_G(\alpha) = \inf[t \in \mathbb{R} : P\{W_G \leq t \mid (y_i, \boldsymbol{X}_i)_{i=1,\ldots,n}\} \geqslant 1 - \alpha]$.

With the debiased estimators and the critical value, we expect a similar result as Theorem 4.1 in Zhang and Cheng (2017) that the asymptotic distribution of $\max_{j \in G} \sqrt{n}|\tilde{\gamma}_j - \gamma_j^*|$ satisfy $\sup_{\alpha \in (0,1)} |P\{\max_{j \in G} \sqrt{n}|\tilde{\gamma}_j - \gamma_j^*| > C_G(\alpha)\} - \alpha| = o(1)$. While the inference on group-wise maximum coefficients is an interesting problem, it is not the focus of this work and we leave the related theoretical proof for future research.

Besides the aforementioned extensions, there are several related problems worth investigating in the future. For instance, as discussed in Section 2.2, there are many models in existing literature related to our problem, in particular, the factor-adjusted method can be

extended to the situation involving hidden confounding (Fan et al., 2020). In addition, in our theoretical analysis, we assume that the dimension of unmeasured confounders, K, is fixed, which has also been imposed in the existing literature (e.g., Wang and Fan, 2017; Fan et al., 2023). One possible extension is to allow K to grow as n and p increase, and it involves generalizations of the theoretical results on the maximum likelihood estimation for the unmeasured confounders. It would also be interesting to generalize the factor model to a nonlinear structure and investigate the theoretical properties of the debiased estimator under generalized factor model (Chen et al., 2020; Liu et al., 2023). Besides generalized linear models, the high-dimensional debiasing technique is also popularly studied in a variety of models such as Gaussian graphical models (Ren et al., 2015; Zhu et al., 2020) and additive hazards models (Lin and Ying, 1994; Lin and Lv, 2013). For instance, consider additive hazards models, which are popularly used in survival analysis and assume the conditional hazard function at time t as $\lambda(t \mid D, \boldsymbol{Q}, \boldsymbol{U}) = \lambda_0(t) + \theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}$ with D and $\boldsymbol{Q}$ covariates and $\boldsymbol{U}$ unmeasured confounders. The relationship between covariates $\{D, \boldsymbol{Q}\}$ and unmeasured confounders $\boldsymbol{U}$ are also modeled by a linear factor model. Our method can be used to perform inference on θ under the aforementioned model setup based on the quadratic loss function. Generalizing our proposed approach to these models to adjust for possible unmeasured confounders is also an interesting direction to investigate.

Appendix A. Preliminaries

We start with introducing some notations and definitions. For a vector $\boldsymbol{r} = (r_1, \ldots, r_l)^{\mathrm{T}}$, we let $\mathcal{S}_r = \{j : r_j \neq 0\}$, $\|\boldsymbol{r}\|_\infty = \max_{j=1,\ldots,l} |r_j|$, $\|\boldsymbol{r}\|_q = (\sum_{j=1}^l |r_j|^q)^{1/q}$ for $q \geq 1$ and $s_r = \mathrm{card}(\mathcal{S}_r)$. For a matrix $\boldsymbol{A} = (a_{ij})_{n \times l}$, let $\|\boldsymbol{A}\|_{\infty,1} = \max_{j=1,\ldots,l} \sum_{i=1}^n |a_{ij}|$ to be the maximum absolute column sum, $\|\boldsymbol{A}\|_{1,\infty} = \max_{i=1,\ldots,n} \sum_{j=1}^l |a_{ij}|$ to be the maximum of the absolute row sum, $\|\boldsymbol{A}\|_{\max} = \max_{i,j} |a_{ij}|$ to be the maximum of the matrix entry, $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ to be the smallest and largest eigenvalues of $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_F = (\sum_{i=1}^n \sum_{j=1}^l |a_{ij}|^2)^{1/2}$ to be the Frobenius norm of $\boldsymbol{A}$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For any sub-exponential random variable Y_1, we define the sub-exponential norm as $\|Y_1\|_{\varphi_1} = \inf[s > 0 : E\{\exp(Y_1/s)\} \leq 2]$. For any sub-Gaussian random variable Y_2, we define the sub-Gaussian norm as $\|Y_2\|_{\varphi_2} = \inf[s > 0 : E\{\exp(Y_2^2/s^2)\} \leq 2]$.

Next, we give a review of our model framework. We assume the response y given the covariate of interest D, the nuisance covariates $\boldsymbol{Q}$ and the unmeasured confounders $\boldsymbol{U}$ follows the generalized linear model with the probability density (mass) function to be

$$f(y) = \exp\left[\{y(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U}) - b(\theta D + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{U})\}/a(\phi) + c(y, \phi)\right], \qquad \text{(A8)}$$

and the relationship between $\boldsymbol{X}$ and $\boldsymbol{U}$ is

$$\boldsymbol{X} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{U} + \boldsymbol{E}. \qquad \text{(A9)}$$

We let $\boldsymbol{Z} = (D, \boldsymbol{Q}^{\mathrm{T}}, \boldsymbol{U}^{\mathrm{T}})^{\mathrm{T}}$ to be a vector that includes all of the covariates and the unmeasured confounders, and let $\boldsymbol{\eta}^{\mathrm{T}} = (\theta, \boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})$ to be the corresponding parameters. For notational convenience, we also let $\boldsymbol{M} = (\boldsymbol{Q}^{\mathrm{T}}, \boldsymbol{U}^{\mathrm{T}})^{\mathrm{T}}$ and its coefficient $\boldsymbol{\zeta} = (\boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, so $\boldsymbol{\eta} = (\theta, \boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}$. We denote $\boldsymbol{\gamma} = (\theta, \boldsymbol{v}^{\mathrm{T}})^{\mathrm{T}}$ as the coefficient for covariates $\boldsymbol{X}$, so $\boldsymbol{\eta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$.

Throughout the appendix, we use an asterisk on the upper subscript to indicate the population parameters. We assume that the observed data $\{y_i, \boldsymbol{X}_i\}_{i=1,\ldots,n}$ and the unmeasured confounders $\{\boldsymbol{U}_i\}_{i=1,\ldots,n}$ are realizations of (A8) and (A9). The noise for the factor model are denoted as $\{\boldsymbol{E}_i\}_{i=1,\ldots,n}$ with $\boldsymbol{E}_i = (E_{ij} : j = 1, \ldots, p)$.

In the proposed inferential procedure, we first obtain the maximum likelihood estimator for unmeasured confounders $\hat{\boldsymbol{U}}_i$. With the unmeasured confounder estimators, we let $\dot{\boldsymbol{Z}}_i = (D_i, \boldsymbol{Q}_i^{\mathrm{T}}, \hat{\boldsymbol{U}}_i^{\mathrm{T}})^{\mathrm{T}}$, $\dot{\boldsymbol{M}}_i = (\boldsymbol{X}_i^{\mathrm{T}}, \hat{\boldsymbol{U}}_i^{\mathrm{T}})^{\mathrm{T}}$ and get an initial lasso estimator

$$\hat{\boldsymbol{\eta}} = \underset{\eta \in \mathbb{R}^{(p+K)}}{\operatorname{argmin}}\ l(\boldsymbol{\eta}) + \lambda\|\boldsymbol{\gamma}\|_1, \qquad \text{(A10)}$$

where the loss function is the negative log-likelihood function defined as

$$l(\boldsymbol{\eta}) = -\frac{1}{n}\sum_{i=1}^n \{y_i \boldsymbol{\eta}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i - b(\boldsymbol{\eta}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\}.$$

The gradient $\nabla l(\boldsymbol{\eta})$ and Hessian $\nabla^2 l(\boldsymbol{\eta})$ of the loss function are commonly used in our subsequent proofs, which are expressed as

$$\nabla l(\boldsymbol{\eta}) = -\frac{1}{n}\sum_{i=1}^n \{y_i - b'(\boldsymbol{\eta}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\}\dot{\boldsymbol{Z}}_i; \quad \nabla^2 l(\boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^n b''(\boldsymbol{\eta}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}}.$$

In constructing the debiased estimator, we define $\boldsymbol{w}^* = \boldsymbol{I}_{\theta\zeta}^*(\boldsymbol{I}_{\zeta\zeta}^*)^{-1}$ and $\boldsymbol{\tau}^* = (1, -\boldsymbol{w}^*)^{\mathrm{T}}$, where $\boldsymbol{I}_{\theta\zeta}^* = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i\boldsymbol{M}_i^{\mathrm{T}}]$ and $\boldsymbol{I}_{\zeta\zeta}^* = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}]$. We denote two sub-vectors of $\boldsymbol{w}^*$ as $\boldsymbol{w}_q^* = (w_2^*, \ldots, w_q^*)^{\mathrm{T}}$ and $\boldsymbol{w}_u^* = (w_{p+1}^*, \ldots, w_{p+K}^*)^{\mathrm{T}}$. We define the estimator for the sparse vector $\boldsymbol{w}$ as

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{(p+K-1)}} \frac{1}{2n}\sum_{i=1}^n \{\boldsymbol{w}^{\mathrm{T}}\nabla_{\zeta\zeta}l_i(\hat{\theta}, \hat{\boldsymbol{\zeta}})\boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}}\nabla_{\zeta\theta}l_i(\hat{\theta}, \hat{\boldsymbol{\zeta}})\} + \lambda'\|\boldsymbol{w}\|_1, \qquad (A11)$$

where $l_i(\theta, \boldsymbol{\zeta}) = -y_i(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i) + b(\theta D_i + \boldsymbol{\zeta}^T\dot{\boldsymbol{M}}_i)$ is equivalent to $l_i(\boldsymbol{\eta})$, the ith component of the loss function. Equivalently, the estimator $\hat{\boldsymbol{w}}$ is obtained by

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^{(p+K-1)}} \frac{1}{2n}\sum_{i=1}^n b''(\hat{\theta}D_i + \hat{\boldsymbol{\zeta}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i)(D_i - \boldsymbol{w}^{\mathrm{T}}\dot{\boldsymbol{M}}_i)^2 + \lambda'\|\boldsymbol{w}\|_1.$$

The estimator $\hat{\boldsymbol{w}}$ is used in estimating the generalized decorrelated score function and the partial Fisher information matrix. Under null hypothesis $H_0 : \theta^* = \theta^0$, the two estimators are

$$\hat{S}(\theta^0, \hat{\boldsymbol{\zeta}}) = -\frac{1}{n}\sum_{i=1}^n \{y_i - b'(\theta^0 D_i + \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{Q}_i + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{U}}_i)\}(D_i - \hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i);$$

$$\hat{I}_{\theta|\zeta} = \frac{1}{n}\sum_{i=1}^n b''(\hat{\theta}D_i + \hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{Q}_i + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{U}}_i)D_i(D_i - \hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i).$$

Our theoretical results are established under the asymptotic regime with $n, p \to \infty$. Regarding the factor model, as stated in Assumption 1 in Section 4 of the main text, we do not assume the random errors $\boldsymbol{E}_i$ to be identically distributed nor does the model assumption require the covariance of $\boldsymbol{E}_i$ to be diagonal. Specifically, we assume for some large constant $C > 0$: (a) $\mathbb{E}(E_{ij}) = 0$, $\mathbb{E}(E_{ij}^8) \leq C$; (b) $\mathbb{E}(E_{ih}E_{ij}) = \tau_{i,hj}$ with $|\tau_{i,hj}| \leq \tau_{hj}$ for some $\tau_{hj} > 0$ and all $i = 1, \ldots, n$, and $\sum_{h=1}^p \tau_{hj} \leq C$ for all $j = 1, \ldots, p$. (c) $\mathbb{E}(E_{ij}E_{sj}) = \rho_{is,j}$ with $|\rho_{is,j}| \leq \rho_{is}$ for some $\rho_{is} > 0$ and all $j = 1, \ldots, p$, and $n^{-1}\sum_{i=1}^n\sum_{s=1}^n \rho_{is} \leq C$. (d) For all $j, q = 1, \ldots, p$,

$$\mathbb{E}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n [E_{ij}E_{iq} - \mathbb{E}(E_{ij}E_{iq})]\right|^4\right\} \leq C.$$

Regarding the loading matrix, we assume $\|\boldsymbol{W}_j^*\|_2 \leq C$. There exist positive definite matrices $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\Upsilon}^*$ such that $\lim_{p\to\infty} p^{-1}\boldsymbol{W}^*(\boldsymbol{\Sigma}_e^*)^{-1}(\boldsymbol{W}^*)^{\mathrm{T}} = \boldsymbol{\Gamma}^*$ and $\lim_{p\to\infty} p^{-1}\sum_{j=1}^p(\sigma_j^*)^{-4}\{(\boldsymbol{W}_j^*)^{\mathrm{T}} \otimes (\boldsymbol{W}_j^*)^{\mathrm{T}}\}(\boldsymbol{W}_j^* \otimes \boldsymbol{W}_j^*) = \boldsymbol{\Upsilon}^*$. In addition, we also assume a working identifiability condition that $\boldsymbol{S}_u = \boldsymbol{I}_K$ and $p^{-1}\boldsymbol{W}^*(\boldsymbol{\Sigma}_e^*)^{-1}(\boldsymbol{W}^*)^{\mathrm{T}}$ is a diagonal matrix with distinct entries.

For the assumptions regarding the generalized linear model relating y and $(\boldsymbol{X}, \boldsymbol{U})$, as mentioned in Assumption 2 in Section 4 of the main text, we let $\lambda_{\min}(\boldsymbol{I}^*) \geq \kappa$, where $\boldsymbol{I}^* = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{Z}_i\boldsymbol{Z}_i^{\mathrm{T}}]$ and κ is some constant. The unmeasured confounders, the covariates and coefficients parameters are assumed to be bounded, that is, $\|\boldsymbol{U}_i\|_\infty \leq M$, $\|\boldsymbol{X}_i\|_\infty \leq M$, $\|\boldsymbol{\eta}^*\|_\infty \leq M$, and $|(\boldsymbol{w}_q^*)^{\mathrm{T}}\boldsymbol{Q}_i| \leq M$ for some constant $M > 0$. We let $|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}|$ to be sub-exponential with $\|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\|_{\varphi_1} \leq M$. In addition, we assume $a_1 \leq (\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i \leq a_2$, $0 \leq |b'(t)| \leq B$ with $|b'(t_1) - b'(t)| \leq B|(t_1 - t)b'(t)|$ and

$0 \leq b''(t) \leq B$ with $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ for constants a_1, a_2 and B, where $t \in [a_1 - \epsilon, a_2 + \epsilon]$ for $\epsilon > 0$ and sequence t_1 satisfying $|t_1 - t| = o(1)$.

With the notations and assumptions revisited, we next present the discussion for the working identifiability condition and then the proofs of Theorems 2 and 3.

Appendix B. Discussion on Working Identifiability Condition

In this section, we make a detailed clarification on the working identifiability condition in the main text. Specifically, we show that when the working identifiability condition does not hold, we can transform the model into an identifiable model and the transformed identifiable model has identical parameter of interest θ compared to that of the pre-transformed model.

When the identifiability condition is not satisfied, the estimators are not fully identifiable in the sense that, for any invertible matrix $\boldsymbol{O}$, we have $\tilde{\boldsymbol{W}} = \boldsymbol{O}^{\mathrm{T}}\hat{\boldsymbol{W}}$ and $\tilde{\boldsymbol{S}}_u = (\boldsymbol{O}^{-1})^{\mathrm{T}}\hat{\boldsymbol{S}}_u\boldsymbol{O}^{-1}$ to be valid maximum likelihood estimators. At $\tilde{\boldsymbol{W}} = \boldsymbol{O}^{\mathrm{T}}\hat{\boldsymbol{W}}$, we can substitute the expression for $\tilde{\boldsymbol{W}}$ into (4) to get the relationship between $\tilde{\boldsymbol{U}}_i$ and $\hat{\boldsymbol{U}}_i$ as follows,

$$
\begin{aligned}
\tilde{\boldsymbol{U}}_i &= (\tilde{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\tilde{\boldsymbol{W}}^{\mathrm{T}})^{-1}\tilde{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \\
&= (\boldsymbol{O}^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\hat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{O})^{-1}\boldsymbol{O}^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \\
&= \boldsymbol{O}^{-1}(\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\hat{\boldsymbol{W}}^{\mathrm{T}})^{-1}(\boldsymbol{O}^{-1})^{\mathrm{T}}\boldsymbol{O}^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \\
&= \boldsymbol{O}^{-1}\hat{\boldsymbol{U}}_i.
\end{aligned}
$$

Suppose $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{U}}_i$ are the asymptotically unbiased estimators for $\boldsymbol{W}^*$ and $\boldsymbol{U}_i^*$ respectively. Here we want to clarify that we use $\boldsymbol{S}_u^*$ and $\boldsymbol{U}_i^*$ to differentiate the unmeasured confounders before the transformation from that of transformed model but this does not imply that $\boldsymbol{S}_u^*$ and $\boldsymbol{U}_i^*$ are population parameters. we have $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{U}}_i$ to be the asymptotically unbiased estimators for $\boldsymbol{O}^{\mathrm{T}}\boldsymbol{W}^*$ and $\boldsymbol{O}^{-1}\boldsymbol{U}_i^*$.

When the working identifiability condition does not hold, that is, we have $\boldsymbol{S}_u^* \neq \boldsymbol{I}_K$, and/or $p^{-1}\boldsymbol{W}^*\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{W}^*)^{\mathrm{T}}$ is not diagonal matrix with distinct entries, we can find an invertible matrix $\boldsymbol{O}$ to transform the true parameters to satisfy the assumption. Then we build a correspondence between the true model and the model corresponding to the transformed parameters, showing that the two models have parameter of interest θ to be identical.

The transformed parameters are $\boldsymbol{W}^r = \boldsymbol{O}_2^{-1}\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*$ and $\boldsymbol{U}_i^r = \boldsymbol{O}_2^{\mathrm{T}}(\boldsymbol{O}_1^{-1})\boldsymbol{U}_i^*$, which are constructed in two steps. At first step, for any $\boldsymbol{S}_u^* \neq \boldsymbol{I}_K$, we can find a matrix $\boldsymbol{O}_1$ such that

$$
n^{-1}\sum_{i=1}^{n}\boldsymbol{O}_1^{-1}(\boldsymbol{U}_i^* - \bar{\boldsymbol{U}}^*)(\boldsymbol{U}_i^* - \bar{\boldsymbol{U}}^*)^{\mathrm{T}}(\boldsymbol{O}_1^{-1})^{\mathrm{T}} = \boldsymbol{O}_1^{-1}\boldsymbol{S}_u^*(\boldsymbol{O}_1^{-1})^{\mathrm{T}} = \boldsymbol{I}_K.
$$

At next step, given the matrix $p^{-1}\boldsymbol{W}^*\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{W}^*)^{\mathrm{T}}$ is symmetrical, there exists an orthogonal matrix $\boldsymbol{O}_2$ whose columns correspond to the eigenvectors of $p^{-1}\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*)^{\mathrm{T}}$ as we could decompose this symmetric matrix into $p^{-1}\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*)^{\mathrm{T}} = \boldsymbol{O}_2\boldsymbol{\Lambda}\boldsymbol{O}_2^{\mathrm{T}}$ where $\boldsymbol{\Lambda}$ has distinct eigenvalues in the diagonal.

We verify the transformed parameters satisfy the identifiability condition as follows. At $\boldsymbol{W}^r = \boldsymbol{O}_2^{-1}\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*$ and $\boldsymbol{U}_i^r = \boldsymbol{O}_2^{\mathrm{T}}(\boldsymbol{O}_1^{-1})\boldsymbol{U}_i^*$, we have

$$
\boldsymbol{S}_u^r = n^{-1}\sum_{i=1}^{n}\boldsymbol{O}_2^{\mathrm{T}}\boldsymbol{O}_1^{-1}(\boldsymbol{U}_i^* - \bar{\boldsymbol{U}}^*)(\boldsymbol{U}_i^* - \bar{\boldsymbol{U}}^*)^{\mathrm{T}}(\boldsymbol{O}_1^{-1})^{\mathrm{T}}\boldsymbol{O}_2 = \boldsymbol{O}_2^{\mathrm{T}}\boldsymbol{I}_K\boldsymbol{O}_2 = \boldsymbol{I}_K,
$$

and

$$p^{-1}\boldsymbol{W}^r\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{W}^r)^{\mathrm{T}} = p^{-1}\boldsymbol{O}_2^{-1}\boldsymbol{O}_1^{\mathrm{T}}\boldsymbol{W}^*\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{O}_1(\boldsymbol{O}_2^{-1})^{\mathrm{T}} = \boldsymbol{O}_2^{-1}\boldsymbol{O}_2\boldsymbol{\Lambda}\boldsymbol{O}_2^{\mathrm{T}}(\boldsymbol{O}_2^{-1})^{\mathrm{T}} = \boldsymbol{\Lambda},$$

is a diagonal matrix with distinct entries.

We let $\boldsymbol{O} = \boldsymbol{O}_2^{\mathrm{T}}(\boldsymbol{O}_1^{-1})$, so accordingly the parameters for the transformed factor model are $\boldsymbol{U}^r = \boldsymbol{O}\boldsymbol{U}^*$ and $(\boldsymbol{W}^r)^{\mathrm{T}} = (\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{O}^{-1}$. The relationship between the true model and the transformed model is as follows. The factor model structure corresponding to the true parameter is the same as the model corresponding to the transformed parameters as

$$\boldsymbol{X} = (\boldsymbol{W}^r)^{\mathrm{T}}\boldsymbol{U}^r + \boldsymbol{E} = (\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{O}^{-1}\boldsymbol{O}\boldsymbol{U}^* + \boldsymbol{E} = (\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{U}^* + \boldsymbol{E}.$$

The generalized linear framework according to the rotated true parameter is

$$f(y) = \exp\left([y\{\theta^r D + (\boldsymbol{v}^r)^{\mathrm{T}}\boldsymbol{Q} + (\boldsymbol{\beta}^r)^{\mathrm{T}}\boldsymbol{U}\} - b\{\theta^r D + (\boldsymbol{v}^r)^{\mathrm{T}}\boldsymbol{Q} + (\boldsymbol{\beta}^r)^{\mathrm{T}}\boldsymbol{U}]/a(\phi) + c(y,\phi)\right),$$

whereas the framework according to true parameters is

$$f(y) = \exp\left([y\{\theta^* D + (\boldsymbol{v}^*)^{\mathrm{T}}\boldsymbol{Q} + (\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{U}\} - b\{\theta^* D + (\boldsymbol{v}^*)^{\mathrm{T}}\boldsymbol{Q} + (\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{U}\}]/a(\phi) + c(y,\phi)\right).$$

At $\boldsymbol{\beta}^r = \boldsymbol{O}^{-1}\boldsymbol{\beta}^*$, $\theta^r = \theta^*$ and $\boldsymbol{v}^r = \boldsymbol{v}^*$, the two frameworks are identical. That is, when the confounders are not identifiable, only the coefficient $\boldsymbol{\beta}$ will be affected accordingly; the parameter of interest will not change and thus the theoretical results on $\tilde{\theta}$ are not affected.

Appendix C. Estimation of the Dimension of Unmeasured Confounders

In the numerical studies of this paper, we use parallel analysis (Horn, 1965) to estimate the dimension of unmeasured confounders. Because in factor analysis, parallel analysis is a popular approach to selecting the number of factors as it is accurate and easy to use (Hayton et al., 2004; Costello and Osborne, 2005; Brown, 2015). In the existing literature, several authors have conducted extensive simulation studies to assess the performance of parallel analysis relative to other existing approaches (Zwick and Velicer, 1986; Peres-Neto et al., 2005). They have shown that parallel analysis has better numerical performance in terms of selecting K than many existing approaches (Zwick and Velicer, 1986; Peres-Neto et al., 2005). Furthermore, parallel analysis is also a commonly used statistical tool for dimension reduction (Lin et al., 2016), multiple testing dependence (Leek and Storey, 2008), and finds wide applications in other scientific disciplines including virology (Quadeer et al., 2014) and genetic studies (Leek and Storey, 2007).

The implementation of parallel analysis is as follows. With our given matrix $\boldsymbol{X}_{n\times p} = (D, \boldsymbol{Q}^{\mathrm{T}})^{\mathrm{T}}$, we denote p columns in the design matrix as $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, respectively. Then we repeatedly generate matrices $\boldsymbol{X}_\pi$'s where each matrix is generated by randomly permuting every column $\boldsymbol{X}_j$ for $j = 1, \ldots, p$. Next, we select the first factor when the top singular value of $\boldsymbol{X}$ is larger than a certain percentile of the top singular value of the permuted matrices $\boldsymbol{X}_\pi$'s. If the first factor is selected, we repeat this procedure to determine whether the second factor can be selected. The process is repeated until no more factor is selected. The main intuition of this approach is that the factor model is considered a summation of the signal (factors) and noise (random error). The permutation destroys the original signal

structure and turns it into a matrix of random noise. Thus, identifying factors based on large singular values of $\boldsymbol{X}$ can be interpreted as selecting factors that are above the noise level.

Besides parallel analysis, there are various methods to estimate the dimension of unmeasured confounders such as scree plot (Cattell, 1966), which empirically chooses the elbow point in the plot of descending eigenvalues of factors; method based on cross validation (Owen and Wang, 2016), which uses random held-out matrices of data to choose the number of factors; method based on information criteria including AIC and BIC to select the number of factors in high-dimensional factor model (Bai and Ng, 2002); the eigenvalue ratio method (Lam and Yao, 2012; Ahn and Horenstein, 2013), which chooses $\hat{K}$ by $\hat{K} = \arg\max_{K \leq \mathcal{K}} \lambda_k(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})/\lambda_{k+1}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ where $\lambda_k(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ denotes the k-th eigenvalue of $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\mathcal{K}$ is a prespecified threshold which is often set to be $\mathcal{K} = p/2$ in practice. Among these methods, the information-criteria-based method and eigenvalue ratio method have the theoretical guarantees to be consistent under similar assumptions as Assumption 1 in our paper. These assumptions follow the common conditions in theoretical analysis for the approximate factor model that allows weak correlation among the random error. Besides the method to determine the number of factors for linear factor models, Chen and Li (2021) proposed a method based on joint-likelihood-based information criterion to determine the number of factors for generalized linear factor models. All these selection methods are well-established tools and can possibly be used to select K. Nonetheless, our theoretical results hold as long as the dimension of unmeasured confounders is consistently estimated. In our manuscript, we use the parallel analysis for good empirical performance, as illustrated in Peres-Neto et al. (2005).

To further investigate how the accuracy of the estimation of K affects the asymptotic distribution of the debiased estimation, we conduct simulation studies where we keep all the estimation settings to be the same as in Section 5.1 in the paper except manually replacing the $\hat{K}$ estimated by parallel analysis with specified values: 2, 4, 5, and 10. Recall that we set the true $K = 3$, so this simulation provides us with some insights on how the overestimation ($\hat{K} = 4, 5$ and 10) and underestimation ($\hat{K} = 2$) of unmeasured confounder dimension affect the inference results.

The results under the logistic regression model at (a) $n = 500$ and p vary from 30 to 1500; (b) $p = 600$ and n varies from 100 to 1500 are presented in Fig. 6. The results under the linear model at the same regime of n and p are presented in Fig. 7. From the results, we find that the overestimation of K appears not to affect the asymptotic normality results but the underestimation of K can influence the asymptotic distribution of debiased estimation and further affects confidence interval estimation. Intuitively, as long as the corresponding linear combinations of the true underlying factors $\boldsymbol{U}$ in the considered models can be well approximated by those of the estimated $\hat{\boldsymbol{U}}$, the developed inference results for θ^* would still hold.

Appendix D. Proof of Theorem 2

D.1 Estimation Consistency of $\hat{\boldsymbol{\eta}}$

Theorem 2 shows that the estimators $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{w}}$ can be consistently estimated. Before proving these results, we first introduce some lemmas that will be used in the proofs.
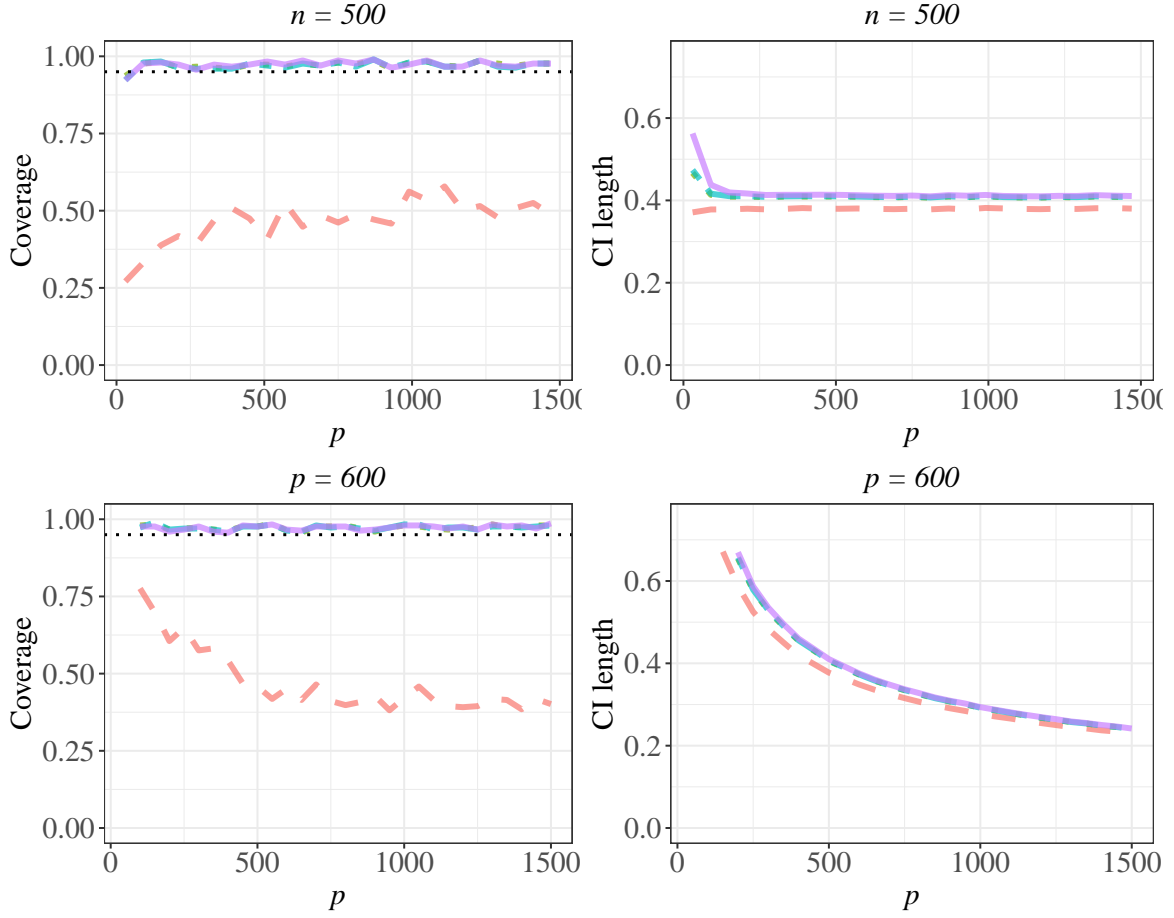
Figure 6: Coverage and length of the confidence interval under logistic regression with pre-specified $\hat{K} = 2, 4, 5$ and 10, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dash-lines indicate the 0.95 level. Orange dashed lines (– –) represent the results at $\hat{K} = 2$. Green dotted lines ($\cdot \cdot \cdot$) represent the results at $\hat{K} = 4$. Blue two-dashed lines (– —) represent the results at $\hat{K} = 5$. Purple solid lines (——) represent the results at $\hat{K} = 10$.

Lemma 7 (Concentration of the Gradient and Hessian) *Under Assumptions 1–2 and the scaling condition $n, p \to \infty$, we have*

   (i) $\|\nabla l(\boldsymbol{\eta}^*)\|_\infty = \|n^{-1} \sum_{i=1}^n [y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}] \dot{\boldsymbol{Z}}_i\|_\infty = O_p\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\};$

   (ii) $\|(\boldsymbol{\tau}^*)^{\mathrm{T}} \nabla^2 l(\boldsymbol{\eta}^*) - E_{\boldsymbol{\eta}^*}\{(\boldsymbol{\tau}^*)^{\mathrm{T}} \nabla^2 l(\boldsymbol{\eta}^*)\}\|_\infty$

$= \|n^{-1} \sum_{i=1}^n (\boldsymbol{\tau}^*)^{\mathrm{T}} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \dot{\boldsymbol{Z}}_i \dot{\boldsymbol{Z}}_i^{\mathrm{T}} - E_{\boldsymbol{\eta}^*}[(\boldsymbol{\tau}^*)^{\mathrm{T}} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \dot{\boldsymbol{Z}}_i \dot{\boldsymbol{Z}}_i^{\mathrm{T}}]\|_\infty$

$= O_p\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}.$

Lemma 7 shows that there exist sub-exponential type of bounds on the gradient and linear combination of the Hessian of the loss functions. This is motivated by Assumption 3.2 in Ning and Liu (2017). They construct the loss function based on the observed covariates
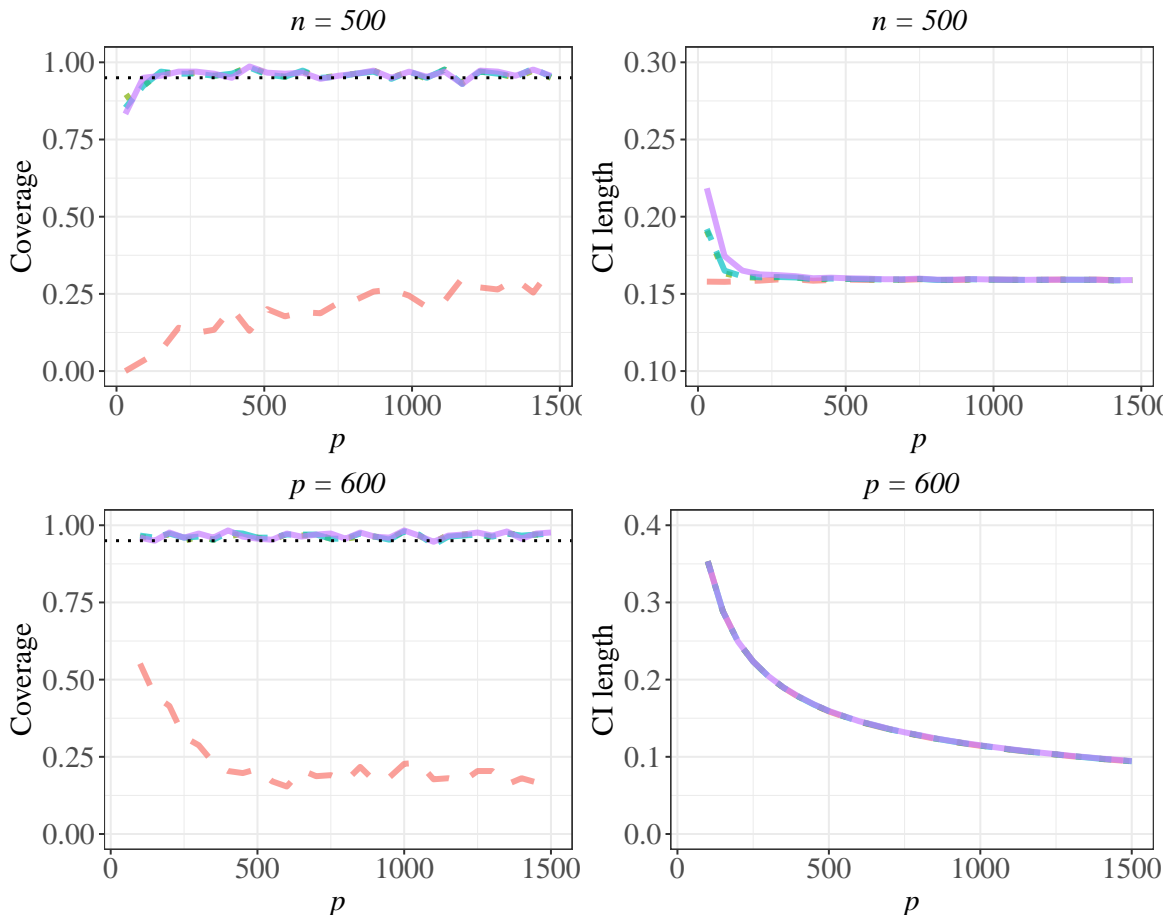
Figure 7: Coverage and length of the confidence interval under linear model with prespecified $\hat{K} = 2, 4, 5$ and 10, averaged over 300 replications, with varying p and fixed $n = 500$ (top), and with varying n and fixed $p = 600$ (bottom). Black dashlines indicate the 0.95 level. Orange dashed lines (**– –**) represent the results at $\hat{K} = 2$. Green dotted lines (**· ·**) represent the results at $\hat{K} = 4$. Blue two-dashed lines (**·– –·**) represent the results at $\hat{K} = 5$. Purple solid lines (——) represent the results at $\hat{K} = 10$.

whereas our results are established upon the fact that the gradient and the hessian of the loss function involve not only observed covariates but also the estimated unmeasured confounders.

Remark 8 *As the decomposition techniques commonly used in linear models may not be applicable in generalized linear model settings, it is necessary to establish stronger and more general intermediate results as the foundation of our theoretical analysis. For instance, different from linear models, where average estimation consistency on unmeasured confounders suffices, stronger uniform estimation consistency is necessary for the generalized linear framework. Specifically, in Fan et al. (2023), the linear model form and projection-*

based techniques enable the reduction of the gradient max-norm into $\|\hat{\boldsymbol{E}}^{\mathrm{T}}(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{E}}\boldsymbol{\gamma}^*)\|_\infty$, where $\boldsymbol{\gamma}^* = (\theta^*, (\boldsymbol{v}^*)^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\boldsymbol{y}} = (\boldsymbol{I}_n - n^{-1}\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^{\mathrm{T}})\boldsymbol{y}$ is the residual of the response $\boldsymbol{y}$ after projecting in onto space of $\hat{\boldsymbol{U}}$. To show the concentration of the gradient, a key step is to upper bound $\|(\boldsymbol{U}^*)^{\mathrm{T}}\hat{\boldsymbol{E}}\boldsymbol{\phi}\|_\infty$ with $\|\boldsymbol{\phi}\|_2 = 1$ by $\|(\hat{\boldsymbol{E}} - \boldsymbol{E})^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{H}^{\mathrm{T}} - \hat{\boldsymbol{U}})\boldsymbol{H}^{-\mathrm{T}}\boldsymbol{\phi}\|_\infty$ and $\|\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{H}^{\mathrm{T}} - \hat{\boldsymbol{U}})\boldsymbol{H}^{-\mathrm{T}}\boldsymbol{\phi}\|_\infty$, where they apply Cauchy-Schwartz inequality and use the Frobenius norm of estimated unmeasured confounders $\|(\boldsymbol{U}\boldsymbol{H}^{\mathrm{T}} - \hat{\boldsymbol{U}})\|_F$. Here $\boldsymbol{H}_{K \times K}$ is some transformation matrix and with a slight abuse of notation, we use $\boldsymbol{E} \in \mathbb{R}^{n \times p}$, $\boldsymbol{U} \in \mathbb{R}^{n \times K}$ and $\boldsymbol{y} \in \mathbb{R}^n$ to denote the matrix (vector) form of random error, unmeasured confounders, and responses. However in the generalized linear model, to derive the bound for gradient max-norm to be $\|n^{-1}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\dot{\boldsymbol{Z}}_i\|_\infty$, we instead apply Bernstein inequality which requires the uniform estimation bound of unmeasured confounders, that is, $\max_i \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i\|_\infty$. We leave the detailed proof of Lemma 7 in Appendix G.1.

We next obtain the upper bound for the estimation error of $\hat{\boldsymbol{\eta}}$. To this end, we define the first-order approximation to the loss difference as $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) = (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\}$. We will obtain upper and lower bounds of $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$. As we will show, the ℓ_2 norm $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_2$ is involved in both upper and lower bounds, combining which will give an inequality of $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_2$ and thus result in a bounded estimation error of $\hat{\boldsymbol{\eta}}$.

Upper Bound for $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$: recall that $\hat{\boldsymbol{\eta}}^{\mathrm{T}} = (\hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}^{\mathrm{T}})$ is a solution obtained from solving the convex optimization problem in (A10). And thus, we have the following optimality conditions $-\nabla l(\hat{\boldsymbol{\gamma}}) = \lambda\boldsymbol{\varrho}$ and $\nabla l(\hat{\boldsymbol{\beta}}) = 0$, where $\boldsymbol{\gamma}$ is a $p \times 1$ vector with entries

$$\boldsymbol{\varrho}_j = \left\{ \begin{array}{ll} \mathrm{sign}(\hat{\gamma}_j), & \hat{\gamma}_j \neq 0 \\ [-1, 1], & \hat{\gamma}_j = 0 \end{array} \right., \quad (j = 1, \ldots p).$$

This implies $\|\nabla l(\hat{\boldsymbol{\eta}})\|_\infty = \|n^{-1}\sum_{i=1}^{n}\dot{\boldsymbol{Z}}_i\{y_i - b'(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\}\|_\infty \leq \lambda$.

Recall that we denote the support set for $\boldsymbol{\eta}^*$ as $\mathcal{S}_\eta = \{j : \eta_j^* \neq 0\}$. We denote the difference between the estimator $\hat{\boldsymbol{\eta}}$ and the true parameter $\boldsymbol{\eta}^*$ as $\hat{\Delta} = \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*$, and its two sub-vectors are $\hat{\Delta}_\mathcal{S} = (\hat{\eta}_j - \eta_j^* : j \in \mathcal{S}_\eta)$ and $\hat{\Delta}_{\bar{\mathcal{S}}} = (\hat{\eta}_j - \eta_j^* : j \notin \mathcal{S}_\eta)$. Similarly, we denote the sub-vectors of $\dot{\boldsymbol{Z}}_i$ corresponding to non-zero entries as $\dot{\boldsymbol{Z}}_{i,\mathcal{S}} = \{\dot{\boldsymbol{Z}}_{ij} : j \in \mathcal{S}_\eta\}$ and that corresponding to zero entries as $\dot{\boldsymbol{Z}}_{i,\bar{\mathcal{S}}} = \{\dot{\boldsymbol{Z}}_{ij} : j \notin \mathcal{S}_\eta\}$. With the notations introduced, the quadratic difference $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$ can then be written as

$$
\begin{aligned}
D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) &= (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_\mathcal{S}^{\mathrm{T}}\dot{\boldsymbol{Z}}_{i,\mathcal{S}}\{y_i - b'(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\} - \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_{\bar{\mathcal{S}}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_{i,\bar{\mathcal{S}}}\{y_i - b'(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\} - \hat{\Delta}^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*) \\
&\leq \lambda\|\hat{\Delta}_S\|_1 - \lambda\|\hat{\Delta}_{\bar{S}}\|_1 + \|\hat{\Delta}\|_1\|\nabla l(\boldsymbol{\eta}^*)\|_\infty
\end{aligned}
$$

where the last inequality is by Hölder's inequality.

From Lemma 7, we have $\|\nabla l(\boldsymbol{\eta}^*)\|_\infty \lesssim n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$. Since $\|\hat{\Delta}\|_1 \leq \|\hat{\Delta}_\mathcal{S}\|_1 + \|\hat{\Delta}_{\bar{\mathcal{S}}}\|_1$ and by taking $\lambda = 2c\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}$ where $c > 0$ is some

constant, we further have

$$
\begin{aligned}
D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) & \leq c\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right)(3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{\bar{S}}\|_1) \\
& \leq 3cs_\eta^{1/2}\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right)\|\hat{\Delta}\|_2,
\end{aligned}
\tag{A12}
$$

where the last inequality is because $\|\hat{\Delta}_{\mathcal{S}}\|_1 \leq s_\eta^{1/2}\|\hat{\Delta}_{\mathcal{S}}\|_2 \leq s_\eta^{1/2}\|\hat{\Delta}\|_2$.

Lower Bound for $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$: after establishing the upper bound for the quadratic difference, we next obtain a lower bound for it. Because $l(\boldsymbol{\eta}^*)$ is convex function, $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) \geq 0$. We have $\|\hat{\Delta}_{\bar{S}}\|_1 \leq 3\|\hat{\Delta}_S\|_1$ from (A12), and based on this result and the restricted strong convexity condition for generalized linear model in Proposition 1 of Loh and Wainwright (2015),

$$
D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) \geq \kappa_2\|\hat{\Delta}\|_2^2,
\tag{A13}
$$

for some constant $\kappa_2 > 0$.

Then combining the upper bound (A12) and the lower bound (A13) for $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$, we have

$$
\|\hat{\Delta}\|_2 \leq \frac{3cs_\eta^{1/2}}{\kappa_2}\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right),
\tag{A14}
$$

and so

$$
\begin{aligned}
\|\hat{\Delta}\|_1 & \leq \|\hat{\Delta}_{\mathcal{S}}\|_1 + \|\hat{\Delta}_{\bar{\mathcal{S}}}\|_1 \leq 4\|\hat{\Delta}_S\|_1 \leq 4s_\eta^{1/2}\|\hat{\Delta}_S\|_2 \leq 4s_\eta^{1/2}\|\hat{\Delta}\|_2 \\
& \leq \frac{12cs_\eta}{\kappa_2}\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right),
\end{aligned}
$$

which completes the first part proof.

D.2 Estimation Consistency of $\hat{\boldsymbol{w}}$

Before we present the proof for the estimation consistency of the estimator $\hat{\boldsymbol{w}}$, we introduce additional results established based on the estimation consistency of $\hat{\boldsymbol{\eta}}$ and used in proving the estimation consistency of $\hat{\boldsymbol{w}}$.

Lemma 9 *Under Assumptions 1–2, with $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$ and $(s_\eta \vee s_w)(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}) = o_p(1)$, then we have*

$$
\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) = O_p\left\{s_\eta\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\},
\tag{A15}
$$

$$
\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{M}}_i^{\mathrm{T}}(\hat{\boldsymbol{w}} - \boldsymbol{w}^*) = O_p\left\{(s_\eta \vee s_w)\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\},
\tag{A16}
$$

$$
\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i b''\{\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{M}}_i^{\mathrm{T}}(\hat{\boldsymbol{w}} - \boldsymbol{w}^*) = O_p\left\{(s_\eta \vee s_w)\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\}.
\tag{A17}
$$

Remark 10 *For Lemma 9, we prove the results* (A16) *and* (A17) *along the way in proving our main result of the estimation consistency of* $\hat{\boldsymbol{w}}$ *as the two inequalities are direct consequences of certain intermediate steps. We leave the proof for* (A15) *in Appendix G.2.*

To prove the consistency of $\hat{\boldsymbol{w}}$, we denote $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{w}} - \boldsymbol{w}^*$ and establish the estimation error bound for $\|\hat{\boldsymbol{\delta}}\|_1$. Recall that $\hat{\boldsymbol{w}}$ is a solution to (A11), that is

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2n} \sum_{i=1}^{n} \{\boldsymbol{w}^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) \dot{\boldsymbol{M}}_i \dot{\boldsymbol{M}}_i^T \boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) D_i \dot{\boldsymbol{M}}_i\} + \lambda' \|\boldsymbol{w}\|_1.$$

By definition, we have

$$\frac{1}{2n} \sum_{i=1}^{n} \{\hat{\boldsymbol{w}}^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) \dot{\boldsymbol{M}}_i \dot{\boldsymbol{M}}_i^T \hat{\boldsymbol{w}} - 2\hat{\boldsymbol{w}}^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) D_i \dot{\boldsymbol{M}}_i\} + \lambda' \|\hat{\boldsymbol{w}}\|_1$$

$$\leq \frac{1}{2n} \sum_{i=1}^{n} \{(\boldsymbol{w}^*)^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) \dot{\boldsymbol{M}}_i \dot{\boldsymbol{M}}_i^T \boldsymbol{w}^* - 2(\boldsymbol{w}^*)^{\mathrm{T}} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) D_i \dot{\boldsymbol{M}}_i\} + \lambda' \|\boldsymbol{w}^*\|_1.$$

Write $(\hat{\boldsymbol{\delta}}^{\mathrm{T}} \dot{\boldsymbol{M}}_i)^2 = (\hat{\boldsymbol{w}}^{\mathrm{T}} \dot{\boldsymbol{M}}_i)^2 - \{(\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\}^2 - 2\hat{\boldsymbol{w}}^{\mathrm{T}} \dot{\boldsymbol{M}}_i \dot{\boldsymbol{M}}_i^{\mathrm{T}} \boldsymbol{w}^* + 2\{(\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\}^2$, then the above inequality can be rearranged as

$$\frac{1}{2n} \sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i)(\hat{\boldsymbol{\delta}}^{\mathrm{T}} \dot{\boldsymbol{M}}_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i)\{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}} + \lambda' \|\boldsymbol{w}^*\|_1 - \lambda' \|\hat{\boldsymbol{w}}\|_1.$$

$$\text{(A18)}$$

The proof techniques are mostly motivated by Ning and Liu (2017). To present our proof, we define two quadratic difference terms as $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) = (\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}} \nabla^2 l(\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{w}} - \boldsymbol{w}^*) = n^{-1} \sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i)(\dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}})^2$ and $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) = (\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}} \nabla^2 l(\boldsymbol{\eta}^*)(\hat{\boldsymbol{w}} - \boldsymbol{w}^*) = n^{-1} \sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}(\dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}})^2$. The left hand side of the above inequality is $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$, and next we investigate the upper bound for $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ in details.

For the right hand side of (A18), we first consider $\lambda' \|\boldsymbol{w}^*\|_1 - \lambda' \|\hat{\boldsymbol{w}}\|_1$. Recall that we denote the support set for $\boldsymbol{w}^*$ as $\mathcal{S}_w = \{j : w_j^* \neq 0\}$. We also denote $\boldsymbol{w}_{\mathcal{S}}^* = (w_j^* : j \in \mathcal{S}_w)$, $\boldsymbol{w}_{\bar{\mathcal{S}}}^* = (w_j^* : j \notin \mathcal{S}_w)$, $\hat{\boldsymbol{\delta}}_{\mathcal{S}} = (\hat{w}_j - w_j^* : j \in \mathcal{S}_w)$ and $\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}} = (\hat{w}_j - w_j^* : j \notin \mathcal{S}_w)$. So $\|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1 = \|\hat{\boldsymbol{w}}_{\bar{\mathcal{S}}}\|_1$ and $\|\boldsymbol{w}_{\bar{\mathcal{S}}}^*\|_1 = 0$. Therefore we have

$$
\begin{aligned}
\lambda' \|\boldsymbol{w}^*\|_1 - \lambda' \|\hat{\boldsymbol{w}}\|_1 &= \lambda' \|\boldsymbol{w}_{\mathcal{S}}^*\|_1 + \lambda' \|\boldsymbol{w}_{\bar{\mathcal{S}}}^*\|_1 - \lambda' \|\hat{\boldsymbol{w}}_{\mathcal{S}}\|_1 - \lambda' \|\hat{\boldsymbol{w}}_{\bar{\mathcal{S}}}\|_1 \\
&= \lambda' \|\boldsymbol{w}_{\mathcal{S}}^*\|_1 - \lambda' \|\hat{\boldsymbol{w}}_{\mathcal{S}}\|_1 - \lambda' \|\hat{\boldsymbol{w}}_{\bar{\mathcal{S}}}\|_1 \\
&\leq \lambda' \|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \lambda' \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1.
\end{aligned}
\tag{A19}
$$

And according to Lagranian duality theory, an equivalent problem for (A11) is

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \quad \text{s.t.} \quad \frac{1}{2n} \sum_{i=1}^{n} \{\boldsymbol{w}^{\mathrm{T}} \nabla_{\zeta\zeta} \ell_i(\hat{\boldsymbol{\eta}}) \boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}} \nabla_{\zeta\theta} \ell_i(\hat{\boldsymbol{\eta}})\} \leq b^2,$$

for $b > 0$. This gives $\|\hat{\boldsymbol{w}}\|_1 \leq \|\boldsymbol{w}^*\|_1$, which further results in $\|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1 \leq \|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1$ from (A19).

We next consider the first term in (A18). Denote $I_1 = n^{-1} \sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}}$, which is a summation of two terms I_{11} and I_{12} denoted as

$$
\begin{aligned}
I_1 &= \frac{1}{n} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} [b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) - b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}] \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}} \\
&= I_{11} + I_{12}.
\end{aligned}
$$

For I_{11}, using a similar argument as in the proof of Lemma 7, we have $\|n^{-1} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}}\|_\infty \lesssim n^{-1/2} (\log p)^{1/2} + p^{-1/2} (\log n)^{1/2}$, hence

$$
\begin{aligned}
|I_{11}| &\leq \|\hat{\boldsymbol{\delta}}\|_1 \|n^{-1} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}}\|_\infty \\
&\lesssim \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \|\hat{\boldsymbol{\delta}}\|_1. \quad (A20)
\end{aligned}
$$

For I_{12}, by Assumption 2 that $|b''(t_1) - b''(t)| \leq B|t_1 - t| b''(t)$ with $t_1 = \hat{\eta}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i$ and $t = (\eta^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i$ and applying Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
|I_{12}| &\leq \left| n^{-1} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} |\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i - (\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i| \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \dot{\boldsymbol{M}}_i\} \dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}} \right| \\
&\leq \left| n^{-1} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} (\dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}})^2 \right|^{1/2} \left| n^{-1} \sum_{i=1}^{n} b'' \{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} \{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}^2 \right|^{1/2} \\
&\lesssim |Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right), \quad (A21)
\end{aligned}
$$

where the last inequality is from (A15) in Lemma 9.

Combining the upper bounds in (A19), (A20) and (A21) into (A18), we have

$$
\begin{aligned}
Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) &\lesssim \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \|\hat{\boldsymbol{\delta}}\|_1 \\
&\quad + |Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \\
&\quad + \lambda' \|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \lambda' \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1. \quad (A22)
\end{aligned}
$$

The above upper bound of $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ involves the expression of $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$. We next investigate the relation between $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ and $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$. We apply Assumption 2 again

on the difference between $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ and $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ as we do for bounding the term I_{12}, and have

$$
\begin{aligned}
|Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) - Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)| &= \left| \frac{1}{n} \sum_{i=1}^{n} [b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}] (\dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}})^2 \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} (\dot{\boldsymbol{M}}_i^{\mathrm{T}} \hat{\boldsymbol{\delta}})^2 (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i \right| \\
&\leq Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 \max_{i=1,\dots,n} \|\dot{\boldsymbol{Z}}_i\|_\infty \\
&\lesssim Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) s_\eta \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right),
\end{aligned}
$$

where the last inequality is by the estimation consistency results of $\hat{\boldsymbol{\eta}}$ in Appendix D.1 and by Proposition 1 that $\|\dot{\boldsymbol{Z}}_i\|_\infty = M + O_p(n^{-1/2} + p^{-1/2} (\log n)^{1/2})$ and $s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}) = o_p(1)$. From the above result, we further apply triangular inequality and have

$$
Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \left\{ 1 - s_\eta \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right) \right\} \lesssim Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*). \quad \text{(A23)}
$$

Combining the above result on $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ with (A22) and taking $\lambda' = 2\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}$, we get

$$
\begin{aligned}
Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) &\lesssim |Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \\
&\quad + \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) (3\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1). \quad \text{(A24)}
\end{aligned}
$$

The above result will be used to prove (A17), which will later be used to derive the error bound $\|\hat{\boldsymbol{\delta}}\|_1$. Consider the following two cases.

Case 1: $|Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} \lesssim s_\eta^{1/2} (n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})$. We have (A17) naturally hold.

Case 2: $|Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} \gtrsim s_\eta^{1/2} (n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})$. Then we can have

$$
Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) - |Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) > 0.
$$

combining this result with (A24), we have $3\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \geq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$. We next use this cone condition to derive the lower bound for $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ for case 2.

Denote $\tilde{\boldsymbol{U}}_i^{\mathrm{T}} = (\mathbf{0}_{p-1}, \hat{\boldsymbol{U}}_i^{\mathrm{T}} - \boldsymbol{U}_i^{\mathrm{T}})_{p+K-1}$ as a vector including a vector of zeros and the differences between estimated unmeasured confounders and the true confounders. Hence we have $\dot{\boldsymbol{M}}_i = \boldsymbol{M}_i + \tilde{\boldsymbol{U}}_i$. We establish the lower bounds for $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)/\|\hat{\boldsymbol{\delta}}\|_2^2$ first.

$$
\begin{aligned}
\frac{Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)}{\|\hat{\boldsymbol{\delta}}\|_2^2} &= \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\{n^{-1}\sum_{i=1}^n b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\dot{\boldsymbol{M}}_i\dot{\boldsymbol{M}}_i^{\mathrm{T}}\}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&= \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\{n^{-1}\sum_{i=1}^n b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}\}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} + \frac{2\hat{\boldsymbol{\delta}}^{\mathrm{T}}\{n^{-1}\sum_{i=1}^n b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}\}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&\quad + \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\{n^{-1}\sum_{i=1}^n b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\tilde{\boldsymbol{U}}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}\}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&= Q_1 + Q_2 + Q_3. \tag{A25}
\end{aligned}
$$

We next derive the bounds for the three terms in (A25) separately. For Q_1, we have

$$
Q_1 = \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}]\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} + \frac{1}{n}\sum_{i=1}^n \frac{(\boldsymbol{M}_i^{\mathrm{T}}\hat{\boldsymbol{\delta}})^2}{\|\hat{\boldsymbol{\delta}}\|_2^2}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]. \tag{A26}
$$

For the second term in (A26), consider Assumption 2 that $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ by letting $t_1 = \hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$ and $t = (\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$ and applying Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\max_{i=1,\ldots,n} |b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}| \\
&\leq \max_{i=1,\ldots,n} Bb''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}|\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i - (\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i| \\
&\leq \max_{i=1,\ldots,n} Bb''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}|\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i - (\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i + (\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i - (\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i| \\
&\leq \max_{i=1,\ldots,n} Bb''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}|(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i + (\boldsymbol{\eta}^*)^{\mathrm{T}}(\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i)|. \\
&\leq \max_{i=1,\ldots,n} Bb''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1\|\dot{\boldsymbol{Z}}_i\|_\infty + \|\boldsymbol{\eta}^*\|_1\|\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i\|_\infty). \tag{A27}
\end{aligned}
$$

From Assumption 2, $b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} \in [0, B]$. From the result for estimation consistency of $\hat{\boldsymbol{\eta}}$ in Appendix D.1, we have $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 \lesssim s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})$. In addition, we have $\|\dot{\boldsymbol{Z}}_i\|_\infty = M + O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. Based on these results and by the scaling condition $n, p \to \infty$ and $s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}) = o_p(1)$, we have

$$
\max_{i=1,\ldots,n} |b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}| = o_p(1). \tag{A28}
$$

Hence we have

$$
Q_1 \geq \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}]\hat{\boldsymbol{\delta}}}{4\|\hat{\boldsymbol{\delta}}\|_2^2} \tag{A29}
$$

with probability tending to 1 as the second term in (A26) goes to 0. Recall that we denote $\boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}} = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}]$. Because

$$
\begin{aligned}
&\frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}}]\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&= \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} + \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}|n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}|\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&\geq \lambda_{\min}(\boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}) - \left|\frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}]\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2}\right| \\
&\geq \kappa - \frac{\|\hat{\boldsymbol{\delta}}\|_1^2\|n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\|_{\max}}{\|\hat{\boldsymbol{\delta}}\|_2^2}
\end{aligned}
$$

where the last inequality is by the definition and properties of eigenvalue as well as matrix operations. Recall that we have obtained the cone condition $3\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \geq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$, which results in $\|\hat{\boldsymbol{\delta}}\|_1^2 \leq 16\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1^2 \leq 16s_w\|\hat{\boldsymbol{\delta}}\|_2^2$. Moreover, using Bernstein's inequality, it can be shown that $\|n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\|_{\max} = O_p\{n^{-1/2}(\log p)^{1/2}\}$. As $(s_\eta \vee s_w)(n^{-1/2}(\log p)^{1/2}+p^{-1/2}(\log n)^{1/2}) = o_p(1)$, the term $\|\hat{\boldsymbol{\delta}}\|_1^2\|n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\boldsymbol{M}_i^{\mathrm{T}} - \boldsymbol{I}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\|_{\max}/\|\hat{\boldsymbol{\delta}}\|_2^2 = o_p(1)$, which together with (A29) shows $Q_1 \geq \kappa/4$ with probability tending to 1.

For Q_2, we can write it as

$$
Q_2 = \frac{2\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}]\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} + \frac{2}{n}\sum_{i=1}^n \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]. \quad \text{(A30)}
$$

Based on (A28) and

$$
\frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|_2^2} \leq \frac{\|\hat{\boldsymbol{\delta}}\|_1^2\|\boldsymbol{M}_i\|_\infty\|\tilde{\boldsymbol{U}}_i\|_\infty}{\|\hat{\boldsymbol{\delta}}\|_2^2} \lesssim s_w\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right),
$$

and under the scaling condition $(s_\eta \vee s_w)(n^{-1/2}(\log p)^{1/2}+p^{-1/2}(\log n)^{1/2}) = o_p(1)$, we have the second term in (A30) goes to 0 and thus

$$
\begin{aligned}
Q_2 &\geq \frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}]\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|_2^2} \\
&\geq \frac{\lambda_{\min}(\boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}})}{2} - \left|\frac{\hat{\boldsymbol{\delta}}^{\mathrm{T}}[n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}]\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|_2^2}\right| \\
&\geq \frac{\kappa}{2} - \frac{\|\hat{\boldsymbol{\delta}}\|_1^2\|n^{-1}\sum_{i=1}^n b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}} - \boldsymbol{I}^*_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\|_{\max}}{\|\hat{\boldsymbol{\delta}}\|_2^2}. \quad \text{(A31)}
\end{aligned}
$$

Because we have

$$\|n^{-1}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}} - \boldsymbol{I}_{\zeta\zeta}^*\|_{\max}$$

$$\leq \|n^{-1}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}} - \mathbb{E}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}]\|_{\max}$$

$$+\|\mathbb{E}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}] - \boldsymbol{I}_{\zeta\zeta}^*\|_{\max}.$$

$$\lesssim \sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}, \tag{A32}$$

where the last inequality is from techniques similar to the proof of Lemma 7 Condition (ii) that $\|n^{-1}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}} - \mathbb{E}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}]\|_{\max} = O_p\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}$ and from Proposition 1 that $\|\mathbb{E}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{M}_i\tilde{\boldsymbol{U}}_i^{\mathrm{T}}] - \boldsymbol{I}_{\zeta\zeta}^*\|_{\max} = O_p\{n^{-1/2} + p^{-1/2}(\log n)^{1/2}\}$. Because $\|\hat{\boldsymbol{\delta}}\|_1^2 \leq 16s_w\|\hat{\boldsymbol{\delta}}\|_2^2$, under the scaling condition $(s_\eta \vee s_w)\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\} = o_p(1)$, combining (A31) and (A32) gives $Q_2 \geq \kappa/2$ with probability tending to 1. Similarly, we have $Q_3 \geq \kappa/4$.

Summarizing the above results, we have that with a high probability $Q_1 + Q_2 + Q_3 \geq \kappa$ as $n, p \to \infty$ and $(s_\eta \vee s_w)\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\} = o_p(1)$. Substituting this result for the terms Q_1, Q_2 and Q_3 into (A25), and because $\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1^2 \leq s_w\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2 \leq s_w\|\hat{\boldsymbol{\delta}}\|_2^2$, we have

$$Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \geq \kappa s_w^{-1}\|\hat{\boldsymbol{\delta}}_S\|_1^2. \tag{A33}$$

Recall we have proved (A24), which can be written as

$$Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2}\left\{Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2} - s_\eta^{1/2}\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2}\right\}$$

$$\lesssim \left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2}(3\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1)$$

$$\lesssim 3\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2}\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1.$$

From (A33), we have $\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1^2 \leq s_w\kappa^{-1}Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ with high probability tending to 1. Substitute this result into the above inequality, we have

$$Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2}\left\{Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2} - s_\eta^{1/2}\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2}\right\}$$

$$\lesssim \frac{3s_w^{1/2}}{\kappa^{1/2}}\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2}Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2}.$$

Cancelling the term $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2}$, we have

$$Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2} - s_\eta^{1/2}\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2} \lesssim \frac{3s_w^{1/2}}{\kappa^{1/2}}\left(\frac{\log p}{n} + \frac{\log n}{p}\right)^{1/2},$$

which gives an upper bound for $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2}$ that

$$Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2} \lesssim (s_w \vee s_\eta)^{1/2} \left(\frac{\log p}{n} + \frac{\log n}{p} \right)^{1/2}. \tag{A34}$$

which completes the proof of (A17) in case 2. Therefore, we prove (A17) holds in both cases. In both cases 1 and 2, by replacing $Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ with $Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$, we get (A16) as

$$Q^*(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) = O_p \left\{ (s_w \vee s_\eta) \left(\frac{\log p}{n} + \frac{\log n}{p} \right) \right\}.$$

To finally prove the estimation consistency of $\hat{\boldsymbol{w}}$, or equivalently, to derive the error bound $\|\hat{\boldsymbol{\delta}}\|_1$, we also consider two situations. If cone condition $6\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \geq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$ holds, we then apply similar techniques as in case 2 to derive (A33). Therefore

$$
\begin{aligned}
\|\hat{\boldsymbol{\delta}}\|_1 \leq \|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1 \leq 7\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \quad &\lesssim \quad s_w^{1/2} Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)^{1/2} \\
&\lesssim \quad s_w^{1/2}(s_\eta \vee s_w)^{1/2} \left(\frac{\log p}{n} + \frac{\log n}{p} \right)^{1/2} \\
&\lesssim \quad (s_\eta \vee s_w) \left(\frac{\log p}{n} + \frac{\log n}{p} \right)^{1/2}.
\end{aligned}
$$

Otherwise, we have $6\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \leq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$. From (A24), we have

$$
\begin{aligned}
Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \quad \lesssim \quad & |Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \\
& + \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) (3\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1). \\
\lesssim \quad & |Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) - \frac{1}{2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1,
\end{aligned}
$$

which together with $6\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \leq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$ gives

$$
\begin{aligned}
\|\hat{\boldsymbol{\delta}}\|_1 \quad \leq \quad & \frac{7}{6} \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1 \\
\leq \quad & \frac{7}{3} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right)^{-1} \left\{ |Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)|^{1/2} s_\eta^{1/2} \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) - Q(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \right\} \\
\lesssim \quad & (s_\eta \vee s_w) \left(\frac{\log p}{n} + \frac{\log n}{p} \right)^{1/2},
\end{aligned}
$$

where the last inequality is from (A17).

In conclusion, under either $6\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \geq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$ or $6\|\hat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \leq \|\hat{\boldsymbol{\delta}}_{\bar{\mathcal{S}}}\|_1$, we prove

$$\|\hat{\boldsymbol{\delta}}\|_1 = O_p \left\{ (s_w \vee s_\eta) \left(\frac{\log p}{n} + \frac{\log n}{p} \right)^{1/2} \right\}.$$

Appendix E. Proof of Theorem 3

Throughout the rest of appendix, we denote

$$\hat{H} = (\hat{W}\hat{\Sigma}_e^{-1}\hat{W}^{\mathrm{T}})^{-1}, \quad \hat{H}_p = p \times \hat{H} = (p^{-1}\hat{W}\hat{\Sigma}_e^{-1}\hat{W}^{\mathrm{T}})^{-1}.$$

From the Corollary S.1 in Bai and Li (2016), we have $\hat{H}_p = O_p(1)$ and $\hat{H} = O_p(p^{-1})$. These results will play an important role in the following proofs. We next introduce a few technical lemmas as tools in the proof of Theorem 3.

Lemma 11 (Smoothness of loss function) *Suppose that Assumptions 1–2 hold. With $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$ and $(s_w \vee s_\eta) (n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, the following conditions hold.*
(iii) $(\tau^*)^{\mathrm{T}}\{\nabla l(\hat{\eta}) - \nabla l(\eta^*) - \nabla^2 l(\eta^*)(\hat{\eta} - \eta^*)\} = o_p(n^{-1/2});$
(iv) $(\hat{\tau} - \tau^*)^{\mathrm{T}}\{\nabla l(\hat{\eta}) - \nabla l(\eta^*)\} = o_p(n^{-1/2}).$

Lemma 11 shows the loss functions are smooth in a sense that they are second-order differentiable around the true parameter value. The conditions hold for quadratic loss functions as well as non-quadratic functions given the function $b(\cdot)$ is properly constrained.

Lemma 12 (Central limit theorem of score function) *Under Assumptions 1–2, with $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$ and $(s_w \vee s_\eta) (n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, it holds that*

$$n^{1/2}(\tau^*)^{\mathrm{T}}\nabla l(\eta^*)(I_{\theta|\zeta}^*)^{-1/2} \to_d N(0,1). \tag{A35}$$

Lemma 12 implies that a linear combination of the gradient of loss function, in other words, the score function is asymptotically normal.

Lemma 13 (Partial information estimator consistency) *Suppose that Assumptions 1–2 hold. With $\lambda \asymp \lambda' \asymp n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}$, if $n, p \to \infty$ and $(s_w \vee s_\eta) (n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, then the estimator for the partial information $\hat{I}_{\theta|\zeta} = \nabla_{\theta\theta}^2 l(\hat{\eta}) - \hat{w}\nabla_{\zeta\theta}^2 l(\hat{\eta})$ is consistent,*

$$\hat{I}_{\theta|\zeta} - I_{\theta|\zeta}^* = o_p(1).$$

With these result, we now prove the asymptotic normality of the debiased estimator, which generalizes the proof of Theorem 3.2 in Ning and Liu (2017) to the setting with unmeasured confounders.

The goal is to show the debiased estimator $\tilde{\theta} = \hat{\theta} - \hat{I}_{\theta|\zeta}^{-1}\hat{S}(\hat{\eta})$ is asymptotically normal. First, by Lemma 12, we have (A35) hold. Therefore, it suffices to show that

$$n^{1/2}|(\tilde{\theta} - \theta^*)(I_{\theta|\zeta}^*)^{1/2} + (\tau^*)^{\mathrm{T}}\nabla l(\eta^*)(I_{\theta|\zeta}^*)^{-1/2}| = o_p(1).$$

which is equivalent to show

$$n^{1/2}|(\tilde{\theta} - \theta^*)I_{\theta|\zeta}^* + (\tau^*)^{\mathrm{T}}\nabla l(\eta^*)| = o_p(1).$$

since $I^*_{\theta|\zeta}$ is constant. Note that we have $\hat{S}(\hat{\boldsymbol{\eta}}) = \hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})$ by the definition of estimated decorrelated score function, we next decompose the left hand side of the above expression and apply triangular inequality as

$$
\begin{aligned}
& n^{1/2}|\{\hat{\theta} - \hat{I}^{-1}_{\theta|\zeta}\hat{S}(\hat{\boldsymbol{\eta}}) - \theta^*\}I^*_{\theta|\zeta} + (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*)| \\
= \ & n^{1/2}|(\hat{\theta} - \theta^*)I^*_{\theta|\zeta} - (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}}) + (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}}) + \hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}}) \\
& - I^*_{\theta|\zeta}\hat{I}^{-1}_{\theta|\zeta}\hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}}) + (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*)| \\
\leq \ & n^{1/2}|(\hat{\theta} - \theta^*)I^*_{\theta|\zeta} - (\boldsymbol{\tau}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\}| \\
& + n^{1/2}|(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})| + n^{1/2}|(I^*_{\theta|\zeta}\hat{I}^{-1}_{\theta|\zeta} - 1)\hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})| \\
=: \ & I_1 + I_2 + I_3.
\end{aligned}
$$

By an application of Lemma 11 condition (iii), we write I_1 as

$$
\begin{aligned}
I_1 = \ & n^{1/2}|(\hat{\theta} - \theta^*)I^*_{\theta|\zeta} - (\boldsymbol{\tau}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\}| \\
\leq \ & n^{1/2}|(\hat{\theta} - \theta^*)I^*_{\theta|\zeta} - (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla^2 l(\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)| + o_p(1) \\
\leq \ & n^{1/2}|(\hat{\theta} - \theta^*)I^*_{\theta|\zeta} - \{\nabla^2_{\theta\theta}l(\boldsymbol{\eta}^*) - (\boldsymbol{w}^*)^{\mathrm{T}}\nabla^2_{\zeta\theta}l(\boldsymbol{\eta}^*)\}(\hat{\theta} - \theta^*) \\
& - \{\nabla^2_{\theta\zeta}l(\boldsymbol{\eta}^*) - \boldsymbol{w}^{\mathrm{T}}\nabla^2_{\zeta\zeta}l(\boldsymbol{\eta}^*)\}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*)| + o_p(1) \\
\leq \ & n^{1/2}\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1\|T\|_\infty + o_p(1)
\end{aligned}
$$

where the last inequality is by Hölder inequality with $T = \{I^*_{\theta|\zeta} - \nabla^2_{\theta\theta}l(\boldsymbol{\eta}^*) + (\boldsymbol{w}^*)^{\mathrm{T}}\nabla^2_{\zeta\theta}l(\boldsymbol{\eta}^*),$ $\nabla^2_{\theta\zeta}l(\boldsymbol{\eta}^*) - (\boldsymbol{w}^*)^{\mathrm{T}}\nabla^2_{\zeta\zeta}l(\boldsymbol{\eta}^*)\}^{\mathrm{T}}$. As a consequence of Lemma 7 condition (ii), we have $\|T\|_\infty = O_p\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}$. In addition, Theorem 2 gives $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 = O_p\{s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})\}$. Hence under the scaling condition of $(s_w \vee s_\eta)$ $(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, we have

$$
I_1 \lesssim \sqrt{n}s_\eta\left(\frac{\log p}{n} + \frac{\log n}{p}\right) + o_p(1) = o_p(1). \tag{A36}
$$

By an application of Lemma 11 condition (iv), we write I_2 as

$$
\begin{aligned}
I_2 = \ & n^{1/2}|(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})| \\
\leq \ & n^{1/2}|(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*)| + o_p(1) \\
\leq \ & n^{1/2}\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*\|_1\|\nabla l(\boldsymbol{\eta}^*)\|_\infty + o_p(1). \tag{A37}
\end{aligned}
$$

From Theorem 2, we have $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*\|_1 = \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_1 = O_p\{(s_w \vee s_\eta)(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})\}$. From Lemma 7 Condition (i), we have $\|\nabla l(\boldsymbol{\eta}^*)\|_\infty = O_p\{n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}\}$. Hence under the condition of $(s_w \vee s_\eta)$ $(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, we have

$$
I_2 \lesssim \sqrt{n}(s_w \vee s_\eta)\left(\frac{\log p}{n} + \frac{\log n}{p}\right) + o_p(1) = o_p(1). \tag{A38}
$$

For I_3, since $\hat{I}_{\theta|\zeta} - I^*_{\theta|\zeta} = o_p(1)$ from Lemma 13, we have $I^*_{\theta|\zeta}\hat{I}^{-1}_{\theta|\zeta} = O_p(1)$. From (A37) and $\boldsymbol{\tau}^*$ being fixed, the term $I_2 = o_p(1)$ implies that $n^{1/2}|\hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})| = o_p(1)$. Hence we have

$$
I_3 = n^{1/2}|(I^*_{\theta|\zeta}\hat{I}^{-1}_{\theta|\zeta} - 1)\hat{\boldsymbol{\tau}}^{\mathrm{T}}\nabla l(\hat{\boldsymbol{\eta}})| = o_p(1). \tag{A39}
$$

Combining (A36), (A38) and (A39), we obtain

$$n^{1/2}|(\tilde{\theta} - \theta^*)I_{\theta|\zeta}^* + (\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*)| \leq I_1 + I_2 + I_3 = o_p(1),$$

which completes the proof.

Appendix F. Proof of Proposition 1

To prove Proposition 1, we show that the estimated unmeasured confounders are bounded with convergence guarantees. From the expression of the estimator for unmeasured confounders in (4) of the main text and recall that $\hat{H} = (\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\hat{\boldsymbol{W}}^{\mathrm{T}})^{-1}$ and $\hat{\boldsymbol{H}}_p = p \times \hat{\boldsymbol{H}} = (p^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\hat{\boldsymbol{W}}^{\mathrm{T}})^{-1}$, we apply triangular inequality and have

$$
\begin{aligned}
\max_{i=1,\ldots,n} \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i\|_\infty &= \max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\{(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{U}_i + \boldsymbol{E}_i\}\|_\infty \\
&\leq \max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{U}_i\|_\infty + \max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\boldsymbol{E}_i\|_\infty.
\end{aligned}
$$
(A40)

For the first term in (A40), we have

$$
\begin{aligned}
\max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{U}_i\|_\infty &\leq \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\|_{1,\infty} \max_{i=1,\ldots,n} \|\boldsymbol{U}_i\|_1 \\
&\leq \sqrt{K}\|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\|_F \max_{i=1,\ldots,n} K\|\boldsymbol{U}_i\|_\infty.
\end{aligned}
$$

From Assumption 2, we have $\|\boldsymbol{U}_i\|_\infty \leq M$ for all $i = 1, \ldots, n$ and some constant $M > 0$. For the norm $\|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\|_F$, we apply the Cauchy-Schwarz inequality to bound the matrix norm

$$
\begin{aligned}
\left\|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\right\|_F &= \left\|\hat{\boldsymbol{H}}_p\frac{1}{p}\sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}\hat{\boldsymbol{W}}_j(\boldsymbol{W}_j^* - \hat{\boldsymbol{W}}_j)^{\mathrm{T}}\right\|_F \\
&\leq \|\hat{\boldsymbol{H}}_p\|_F \left(\frac{1}{p}\sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}\|\hat{\boldsymbol{W}}_j\|_2^2\right)^{1/2} \left(\frac{1}{p}\sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}\|\boldsymbol{W}_j^* - \hat{\boldsymbol{W}}_j\|_2^2\right)^{1/2} \\
&= O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{p}\right)
\end{aligned}
$$

where the last equality follows because $\|\hat{\boldsymbol{H}}_p\|_F = O_p(1)$ from Corollary S.1(b), $(p^{-1}\sum_{j=1}^p \hat{\sigma}_j^{-2}\|\hat{\boldsymbol{W}}_j\|_2^2)^{1/2} = O_p(1)$ from Corollary S.1(a) and $(p^{-1}\sum_{j=1}^p \hat{\sigma}_j^{-2}\|\boldsymbol{W}_j^* - \hat{\boldsymbol{W}}_j\|_2^2)^{1/2} = O_p(n^{-1/2} + p^{-1})$ from Proposition 1 in Bai and Li (2016). Therefore

$$\max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^{\mathrm{T}}\boldsymbol{U}_i\|_\infty = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{p}\right). \tag{A41}$$

For the second term in (A40), we have

$$
\begin{aligned}
\max_{i=1,\ldots,n} \|\hat{\boldsymbol{H}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\boldsymbol{E}_i\|_\infty &\leq \|\hat{\boldsymbol{H}}_p\|_{1,\infty} \max_{i=1,\ldots,n} \|p^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\boldsymbol{E}_i\|_1 \\
&\leq \sqrt{K}\|\hat{\boldsymbol{H}}_p\|_F \max_{i=1,\ldots,n} K\|p^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\boldsymbol{E}_i\|_\infty. \tag{A42}
\end{aligned}
$$

41

For $p^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{\Sigma}}_e^{-1}\boldsymbol{E}_i$, we express the K-dimensional vector as

$$
\begin{aligned}
\frac{1}{p}\sum_{j=1}^p \hat{\sigma}_j^{-2}\hat{\boldsymbol{W}}_j E_{ij} &= \frac{1}{p}\sum_{j=1}^p (\sigma_j^*)^{-2}\boldsymbol{W}_j^* E_{ij} + \frac{1}{p}\sum_{j=1}^p \left\{\frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2}\right\}\boldsymbol{W}_j^* E_{ij} \\
&\quad + \frac{1}{p}\sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}(\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j)E_{ij}. \\
&=: \ R_1 + R_2 + R_3. \quad\quad\quad\quad\quad\quad (A43)
\end{aligned}
$$

We next investigate the bound for the term $R_1 = p^{-1}\sum_{j=1}^p (\sigma_j^*)^{-2}\boldsymbol{W}_j^* E_{ij}$ and then show the other two terms R_2 and R_3 are of smaller order and dominated by R_1. From Assumption 2, we have $\|\boldsymbol{U}_i\|_\infty \le M$, $\|\boldsymbol{W}_j^*\|_2 \le C$ and $\|\boldsymbol{X}_i\|_\infty \le M$. We apply triangular inequality and have

$$
\begin{aligned}
\|\boldsymbol{E}_i\|_\infty &= \|\boldsymbol{X}_i - (\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{U}_i\|_\infty \\
&\le \|\boldsymbol{X}_i\|_\infty + \|(\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{U}_i\|_\infty \\
&\le M + \max_{j=1,\dots,p}\|\boldsymbol{W}_j^*\|_1\|\boldsymbol{U}_i\|_1 \\
&\le M + \max_{j=1,\dots,p}\sqrt{K}\|\boldsymbol{W}_j^*\|_2 K\|\boldsymbol{U}_i\|_\infty \\
&\le M(1 + CK^{3/2}),
\end{aligned}
$$

where the last three inequalities follow from matrix operation and properties of norms.

We have E_{ij} to be sub-exponential random variable since E_{ij} is bounded with $|E_{ij}| \le M(1 + CK^{3/2})$. Combining the bound for E_{ij} with that $|W_{jk}^*| \le \|\boldsymbol{W}_j^*\|_\infty \le \|\boldsymbol{W}_j^*\|_2 \le C$ and $C^{-2} \le (\sigma_j^*)^{-2} \le C^2$, we have $|(\sigma_j^*)^{-2}W_{jk}^* E_{ij}| \le MC^3(1 + CK^{3/2})$, for $j = 1,\dots,p$ and $k = 1,\dots,K$. So $(\sigma_j^*)^{-2}W_{jk}^* E_{ij}$ is sub-Gaussian random variable and thus is sub-exponential. As E_{ij} has mean zero, $(\sigma_j^*)^{-2}W_{jk}^* E_{ij}$ has mean zero and by Bernstein inequality

$$
\begin{aligned}
&P\left(\left|\frac{1}{p}\sum_{j=1}^p (\sigma_j^*)^{-2}W_{jk}^* E_{ij}\right| \ge t\right) \\
&\le 2\exp\left[-C''\min\left\{\frac{t^2}{M^2 C^6(1 + CK^{3/2})^2}, \frac{t}{MC^3(1 + CK^{3/2})}\right\}p\right].
\end{aligned}
$$

Apply union bound inequality, we have

$$
\begin{aligned}
&P\left(\max_{i=1,\dots,n}\left\|\frac{1}{p}\sum_{j=1}^p (\sigma_j^*)^{-2}\boldsymbol{W}_j^* E_{ij}\right\|_\infty \ge t\right) \\
&\le 2nK\exp\left[-C''\min\left\{\frac{t^2}{M^2 C^6(1 + CK^{3/2})^2}, \frac{t}{MC^3(1 + CK^{3/2})}\right\}p\right],
\end{aligned}
$$

where $C'' > 0$ is a constant. At $t = MC^3(1 + CK^{3/2})p^{-1/2}(\log n)^{1/2}$, the inequality

$$
\max_{i=1,\dots,n}\left\|\frac{1}{p}\sum_{j=1}^p (\sigma_j^*)^{-2}W_j^* E_{ij}\right\|_\infty \le MC^3(1 + CK^{3/2})\left(\frac{\log n}{p}\right)^{1/2}, \quad\quad (A44)
$$

holds with probability at least $1 - n^{-1}$.

For the term R_2, we apply Cauchy-Schwarz inequality and have

$$\left\| \frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\} \boldsymbol{W}_j^* E_{ij} \right\|_2$$

$$\leq \quad \left[\frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\}^2 \right]^{1/2} \left\{ \frac{1}{p} \sum_{j=1}^{p} E_{ij}(\boldsymbol{W}_j^*)^\intercal \boldsymbol{W}_j^* E_{ij} \right\}^{1/2}.$$

By Assumption 1(b), $\hat{\sigma}_j^{-2}$ and $(\sigma_j^*)^{-2}$ are bounded in $[C^{-2}, C^2]$ and as a result, we have

$$\left[\frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\}^2 \right]^{1/2} \leq 2C^2.$$

and

$$\max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\} \boldsymbol{W}_j^* E_{ij} \right\|_\infty \leq \max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\} \boldsymbol{W}_j^* E_{ij} \right\|_2$$

$$\leq \quad 2C^2 \max_{i=1,\dots,n} \left\{ \frac{1}{p} \sum_{j=1}^{p} E_{ij}(\boldsymbol{W}_j^*)^\intercal \boldsymbol{W}_j^* E_{ij} \right\}^{1/2}.$$

Because $|W_{jk}^* W_{jk'}^* E_{ij}^2| \leq M^2 C^2 (1 + CK^{3/2})^2$ and by Berstein inequality,

$$P\left(\left| \frac{1}{p} \sum_{j=1}^{p} W_{jk}^* W_{jk'}^* E_{ij}^2 \right| \geq t \right)$$

$$\leq 2 \exp\left[-C'' \min\left\{ \frac{t^2}{M^4 C^4 (1+CK^{3/2})^4}, \frac{t}{M^2 C^2 (1+CK^{3/2})^2} \right\} p \right].$$

and then applying union bound gives

$$P\left(\max_{i=1,\dots,n} \left\{ \frac{1}{p} \sum_{j=1}^{p} E_{ij}(\boldsymbol{W}_j^*)^\intercal \boldsymbol{W}_j^* E_{ij} \right\} \geq t \right)$$

$$\leq 2nK^2 \exp\left[-C'' \min\left\{ \frac{t^2}{M^4 C^4 (1+CK^{3/2})^4}, \frac{t}{M^2 C^2 (1+CK^{3/2})^2} \right\} p \right].$$

From the above probability bound, we have $\max_{i=1,\dots,n} \left\{ p^{-1} \sum_{j=1}^{p} E_{ij}(\boldsymbol{W}_j^*)^\intercal \boldsymbol{W}_j^* E_{ij} \right\} = O_p(p^{-1/2}(\log n)^{1/2})$ and

$$\max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{1}{\hat{\sigma}_j^2} - \frac{1}{(\sigma_j^*)^2} \right\} \boldsymbol{W}_j^* E_{ij} \right\|_\infty = O_p\left\{ \left(\frac{\log n}{p} \right)^{1/4} \right\} \tag{A45}$$

For the term R_3, we apply Cauchy-Schwarz inequality and get

$$\left\| \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} (\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j) E_{ij} \right\|_2 \leq \left(\frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} \|\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j\|_2^2 \right)^{1/2} \left(\frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} E_{ij}^2 \right)^{1/2}. \quad \text{(A46)}$$

By Proposition 1 in Bai and Li (2016), the first term

$$\left(\frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} \|\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j\|_2^2 \right)^{1/2} = O_p \left(\frac{1}{\sqrt{n}} + \frac{1}{p} \right).$$

By a similar technique as in bounding R_1 and R_2 using the combination of Bernstein inequality and union bound, we can show $\max_{i=1,\dots,n} (p^{-1} \sum_{j=1}^{p} \hat{\sigma}_j^{-2} E_{ij}^2)^{1/2} = O_p(p^{-1/4}(\log n)^{1/4})$, and therefore

$$\max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} (\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j) E_{ij} \right\|_\infty$$

$$\leq \max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} (\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j) E_{ij} \right\|_2$$

$$\leq \left(\frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} \|\hat{\boldsymbol{W}}_j - \boldsymbol{W}_j\|_2^2 \right)^{1/2} \max_{i=1,\dots,n} \left(\frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2} E_{ij}^2 \right)^{1/2}$$

$$= O_p \left\{ \frac{(\log n)^{1/4}}{p^{1/4}} \left(\frac{1}{\sqrt{n}} + \frac{1}{p} \right) \right\} \quad \text{(A47)}$$

Combining (A43), (A44), (A45) and (A47), we have

$$\max_{i=1,\dots,n} \left\| \frac{1}{p} \sum_{j=1}^{p} \hat{\sigma}_j^{-2} \hat{\boldsymbol{W}}_j E_{ij} \right\|_\infty = O_p \left\{ \sqrt{\frac{\log n}{p}} \right\}.$$

In addition, from Corollary S.1 (a) in Bai and Li (2016), we have $\hat{\boldsymbol{H}}_p = (\boldsymbol{\Gamma}^*)^{-1} + o_p(1)$ with $\|(\boldsymbol{\Gamma}^*)^{-1}\|_{\mathrm{sp}} = \lambda_{\max}^{1/2} \{(\boldsymbol{\Gamma}^*)^{-\mathsf{T}} (\boldsymbol{\Gamma}^*)^{-1}\} = \lambda_{\max} \{(\boldsymbol{\Gamma}^*)^{-1}\}$ being finite constant as $(\boldsymbol{\Gamma}^*)^{-1}$ is symmetric and positive definite. Substituting these results into (A42), we have

$$\max_{i=1,\dots,n} \|\hat{\boldsymbol{H}} \hat{\boldsymbol{W}} \hat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{E}_i\|_\infty \lesssim K^{3/2} \lambda_{\max} \{(\boldsymbol{\Gamma}^*)^{-1}\} O_p \left\{ \sqrt{\frac{\log n}{p}} \right\} \quad \text{(A48)}$$

Combining (A40) and (A48), we show

$$\max_{i=1,\dots,n} \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i\|_\infty = O_p \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right).$$

Appendix G. Proofs of Lemmas

G.1 Proof of Lemma 7

Proof of Condition (i). To prove the Condition (i) in Lemma 7, we need to show that

$$\left\| -\frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\dot{\boldsymbol{Z}}_i \right\|_{\infty} = O_p\left\{ \left(\frac{\log p}{n}\right)^{1/2} + \left(\frac{\log n}{p}\right)^{1/2} \right\}.$$

From Assumption 2, we have $y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}$ to be sub-exponential variable with mean 0 and $\|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\|_{\varphi_1} \le M$. In addition, it is assumed that $a_1 \le (\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i \le a_2$ and $0 \le b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} \le B$. Denote $\max_i \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i^*\|_{\infty} = \|\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i\|_{\infty} = \epsilon$. From Proposition 1, $\epsilon = O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. Since $\|\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i\|_{\infty} = \epsilon$, it can be shown that $|(\boldsymbol{\eta}^*)^{\mathrm{T}}(\boldsymbol{Z}_i - \dot{\boldsymbol{Z}}_i)| \le a_2 + s_{\eta}M\epsilon$.

To prove condition (i), we focus on finding the appropriate sequence of t such that the following probability tends to 0.

$$P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\dot{\boldsymbol{Z}}_i \right\|_{\infty} \ge t \right)$$

$$\le P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\hat{\boldsymbol{U}}_i \right\|_{\infty} \ge t \right) + P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\boldsymbol{X}_i \right\|_{\infty} \ge t \right)$$

$$\le P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{U}_i \right\|_{\infty} \ge t_1 - \delta_1 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\boldsymbol{U}_i \right\|_{\infty} \ge \delta_1 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}](\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \right\|_{\infty} \ge t_2 - \delta_2 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}](\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \right\|_{\infty} \ge \delta_2 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{X}_i \right\|_{\infty} \ge t_3 - \delta_3 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\boldsymbol{X}_i \right\|_{\infty} \ge \delta_3 \right)$$

$$=: P_1 + P_2 + P_3 + P_4 + P_5 + P_6, \tag{A49}$$

where $t = \max\{t_1 + t_2, t_3\}$

We let $\delta_1 = \delta_3 = BM(a_2 + s_{\eta}M\epsilon)$ and $\delta_2 = B\epsilon(a_2 + s_{\eta}M\epsilon)$, then it can be verified that the probability P_2, P_4 and P_6 tends to 0 under $\epsilon = O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. Next, we look at P_1, P_3 and P_5 to determine t_1, t_2 and t_3 separately.

For P_1, as $[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]U_{ik}$ is independent mean 0 sub-exponential random variables and $\|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]U_{ik}\|_{\varphi_1} \leq 2M^2$, by Bernstein inequality, we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n}|y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]U_{ik}| \geq t_1 - \delta_1\right) \leq 2\exp\left\{-\tilde{M}''\min\left(\frac{(t_1-\delta_1)^2}{4M^4}, \frac{t_1-\delta_1}{2M^2}\right)n\right\}.$$

Then applying union bound inequality, we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{U}_i\right\|_{\infty} \geq t_1 - \delta_1\right)$$
$$\leq \quad 2K\exp\left\{-\tilde{M}''\min\left(\frac{(t_1-\delta_1)^2}{4M^4}, \frac{t_1-\delta_1}{2M^2}\right)n\right\},$$

so at $t_1 - \delta_1 = M^2 n^{-1/2}$, the probability P_1 tends to 0. Using similar techniques, we can show that at $t_2 - \delta_2 = M\epsilon n^{-1/2}$, the probability P_3 tends to 0; at $t_3 - \delta_3 = M^2 n^{-1/2}(\log p)^{1/2}$, the probability P_5 tends to 0. Hence at

$$t = \max\{t_1 + t_2, t_3\}$$
$$= \max\left\{\frac{M^2}{\sqrt{n}} + BM(a_2 + s_\eta M\epsilon) + \frac{M\epsilon}{\sqrt{n}} + B\epsilon(a_2 + s_\eta M\epsilon),\right.$$
$$\left. M^2\sqrt{\frac{\log p}{n}} + BM(a_2 + s_\eta M\epsilon)\right\},$$

the probability (A49) tends to 0. As t_3 dominates $t_1 + t_2$ in the above expression, we eventually have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[y_i - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\dot{\boldsymbol{Z}}_i\right\|_{\infty} = O_p\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right),$$

which completes the proof for condition (i).

Proof of Condition (ii). To show condition (ii) in Lemma 7 holds, we use a similar technique as in proving condition (i) and decomposing the probability as follows. Recall that $\boldsymbol{\tau}^* = (1, -(\boldsymbol{w}^*)^{\mathrm{T}})^{\mathrm{T}}$ and the two sub-vectors of $\boldsymbol{w}^*$ are denoted as $\boldsymbol{w}_q^* = (w_2^*, \ldots, w_q^*)^{\mathrm{T}}$ and $\boldsymbol{w}_u^* = (w_{p+1}^*, \ldots, w_{p+K}^*)^{\mathrm{T}}$. Denote $\boldsymbol{\tau}_q^* = (1, -(\boldsymbol{w}_q^*)^{\mathrm{T}})^{\mathrm{T}} = (1, -w_2^*, \ldots, -w_q^*)^{\mathrm{T}}$.

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}}\right]\right\|_{\infty} \geq t\right)$$
$$\leq P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}}\right]\right\|_{\infty} \geq t\right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t \right)$$

$$\leq P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}} - E_{\eta^*}\left[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t_4 - \delta_4 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}} \right. \right.$$

$$\left. \left. - E_{\eta^*}\left[(\boldsymbol{w}_u^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\hat{\boldsymbol{U}}_i\boldsymbol{X}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq \delta_4 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} - E_{\eta^*}\left[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t_5 - \delta_5 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right. \right.$$

$$\left. \left. - E_{\eta^*}\left[(\boldsymbol{w}_u^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq \delta_5 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t_6 - \delta_6 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right. \right.$$

$$\left. \left. - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{X}_i\hat{\boldsymbol{U}}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq \delta_6 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq t_7 - \delta_7 \right)$$

$$+ P\left(\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \right. \right.$$

$$\left. \left. - E_{\eta^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_q^*)^{\mathrm{T}}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} \right] \right\|_{\infty} \geq \delta_7 \right)$$

$$=: R_1 + R_2 + R_3 + R_4 + R_5 + R_6 + R_7 + R_8, \tag{A50}$$

where $t = \max\{t_4, t_5, t_6, t_7\}$.

Similarly as in proof of condition (i), we let $\max_i \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i^*\|_{\infty} = \|\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i\|_{\infty} = \epsilon$ with $\epsilon = O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. From Assumption 2, we have $b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\} \in [0, B]$,

$|(\boldsymbol{\tau}_q^*)^{\mathrm{T}}\boldsymbol{X}_i| \leq 2M$ and as a result, $|(\boldsymbol{w}_u^*)^{\mathrm{T}}\hat{U}_i| \leq K\|\boldsymbol{w}^*\|_\infty(M+\epsilon)$. In addition, $|b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}| \leq |b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}|(\boldsymbol{\eta}^*)^{\mathrm{T}}(\dot{\boldsymbol{Z}}_i - \boldsymbol{Z}_i) \leq B(a_2 + s_\eta M\epsilon)$.

We let

$$\delta_4 = 2BM(a_2 + s_\eta M\epsilon)\{K\|\boldsymbol{w}^*\|_\infty(M+\epsilon)\};$$
$$\delta_5 = 2B(M+\epsilon)(a_2 + s_\eta M\epsilon)\{K\|\boldsymbol{w}^*\|_\infty(M+\epsilon)\};$$
$$\delta_6 = 4BM(M+\epsilon)(a_2 + s_\eta M\epsilon);$$
$$\delta_7 = 4BM^2(a_2 + s_\eta M\epsilon).$$

It can be shown that R_2, R_4, R_6 and R_8 tends to 0 under $\epsilon = O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. Then we determine t_4, t_5, t_6 and t_7 such that the probability (A50) tends to 0.

For R_1, as $(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij} - E_{\eta^*}[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij}]$ is independent mean 0 sub-exponential random variables and

$$\|(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij} - E_{\eta^*}[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij}]\|_{\varphi_1}$$
$$\leq 2BMK\|\boldsymbol{w}^*\|_\infty(M+\epsilon) =: M_c(M+\epsilon),$$

where we denote $M_c = 2BMK\|\boldsymbol{w}^*\|_\infty$ for notational simplicity. By Bernstein inequality, we have

$$P\left(\frac{1}{n}\sum_{i=1}^n\left|(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij} - E_{\eta^*}[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_iX_{ij}]\right| \geq t_4 - \delta_4\right)$$
$$\leq 2\exp\left\{-\tilde{M}''\min\left(\frac{(t_4 - \delta_4)^2}{M_c^2(M+\epsilon)^2}, \frac{t_4 - \delta_4}{M_c^2(M+\epsilon)^2}\right)n\right\}.$$

Then applying union bound inequality, we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_i\boldsymbol{X}_i - E_{\eta^*}[(\boldsymbol{w}_u^*)^{\mathrm{T}}b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}\hat{U}_i\boldsymbol{X}_i]\right\|_\infty \geq t_4 - \delta_4\right)$$
$$\leq 2p\exp\left\{-\tilde{M}''\min\left(\frac{(t_4 - \delta_4)^2}{M_c^2(M+\epsilon)^2}, \frac{t_4 - \delta_4}{M_c^2(M+\epsilon)^2}\right)n\right\},$$

so at $t_4 - \delta_4 = M_c(M+\epsilon)n^{-1/2}(\log p)^{1/2}$, the probability R_1 tends to 0. Using similar techniques, we can show that at $t_5 - \delta_5 = 2BK\|\boldsymbol{w}^*\|_\infty(M+\epsilon)^2n^{-1/2}$, the probability R_3 tends to 0; at $t_6 - \delta_6 = 4BM(M+\epsilon)n^{-1/2}$, the probability R_5 tends to 0; at $t_7 - \delta_7 = 4BM^2n^{-1/2}(\log p)^{1/2}$, the probability R_7 tends to 0. Hence at

$$\begin{aligned}
t &= \max\{t_4, t_5, t_6, t_7\} \\
&= \max\Big\{M_c(M+\epsilon)n^{-1/2}(\log p)^{1/2} + 2BM(a_2 + s_\eta M\epsilon)\{K\|\boldsymbol{w}^*\|_\infty(M+\epsilon)\}, \\
&\quad 2BK\|\boldsymbol{w}^*\|_\infty(M+\epsilon)^2n^{-1/2} + 2B(M+\epsilon)(a_2 + s_\eta M\epsilon)\{K\|\boldsymbol{w}^*\|_\infty(M+\epsilon)\}, \\
&\quad 4BM(M+\epsilon)n^{-1/2} + 4BM(M+\epsilon)(a_2 + s_\eta M\epsilon), \\
&\quad 4BM^2n^{-1/2}(\log p)^{1/2} + 4BM^2(a_2 + s_\eta M\epsilon)\Big\},
\end{aligned}$$

the probability (A50) tends to 0. Since t_7 dominates t_4, t_5 and t_6, we have

$$
\left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}} - E_{\boldsymbol{\eta}^*}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}} \right] \right\|_\infty
$$
$$
= O_p\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right).
$$

This completes the proof for Condition (ii).

G.2 Proof of Lemma 9

In this proof, we denote

$$
H_\eta \;=\; \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)
$$

We continue to use the notations and results defined in Appendix D.1. Recall we define $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) = (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\}$. To show (A15), we consider the difference of $D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$ and H_η and apply the mean value theorem with $\tilde{\boldsymbol{\eta}} = \xi\hat{\boldsymbol{\eta}} + (1-\xi)\boldsymbol{\eta}^*$ for $\xi \in [0,1]$ to get

$$
|D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*) - H_\eta|
$$
$$
= \left| (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\} - \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \right|
$$
$$
= \left| (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\left[\nabla^2 l(\tilde{\boldsymbol{\eta}}) - \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\dot{\boldsymbol{Z}}_i\dot{\boldsymbol{Z}}_i^{\mathrm{T}} \right](\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \right|
$$
$$
= \left| \frac{1}{n}\sum_{i=1}^{n}[b''(\tilde{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]\{\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\}^2 \right|
$$
$$
\le B\left| \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\}^2 \right| \max_i\left| (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i \right| \tag{A51}
$$
$$
\lesssim B\left| \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{\dot{\boldsymbol{Z}}_i^{\mathrm{T}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\}^2 \right| \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 \max_{i=1,\ldots,n}\|\dot{\boldsymbol{Z}}_i\|_\infty
$$
$$
\lesssim H_\eta s_\eta\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right)\left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right) \tag{A52}
$$

where the inequality (A51) is based on Assumption 2(3) that $|b''(t_1) - b''(t)| \le B|t_1 - t|b''(t)$ with $t_1 = \tilde{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$ and $t = (\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$. The last inequality (A52) is from the estimation error bound in the proof of Theorem 2 that $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_1 \lesssim s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2})$ and from the result derived from Proposition 1 that $\max_i\|\dot{\boldsymbol{Z}}_i\|_\infty = M + O_p(n^{-1/2} + p^{-1/2}(\log n)^{1/2})$. Under the assumption that $s_\eta(n^{-1/2}(\log p)^{1/2} + p^{-1/2}(\log n)^{1/2}) = o_p(1)$,

we have

$$H_\eta \left\{ 1 - s_\eta \left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}} \right) \left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right) \right\} \lesssim D(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^*)$$

$$\lesssim 9c^2 s_\eta \left(\frac{\log p}{n} + \frac{\log n}{p} \right),$$

where the last inequality is by combining (A12) and (A14). Therefore

$$H_\eta = O_p \left\{ s_\eta \left(\frac{\log p}{n} + \frac{\log n}{p} \right) \right\}.$$

The two inequalities (A16) and (A17) are shown to hold in Appendix D.2. Hence the proof for Lemma 9 is complete.

G.3 Proof of Lemma 11

As shown in the preliminaries, we have

$$\nabla l(\boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^n \{-y_i + b'(\boldsymbol{\eta}^\mathrm{T} \dot{\boldsymbol{Z}}_i)\} \dot{\boldsymbol{Z}}_i, \quad \nabla^2 l(\boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^n \{-y_i + b''(\boldsymbol{\eta}^\mathrm{T} \dot{\boldsymbol{Z}}_i)\} \dot{\boldsymbol{Z}}_i \dot{\boldsymbol{Z}}_i^\mathrm{T}.$$

Proof of Condition (iii). For condition (iii), we apply mean value theorem with $\tilde{\boldsymbol{\eta}} = \xi \hat{\boldsymbol{\eta}} + (1 - \xi) \boldsymbol{\eta}^*$ for $\xi \in [0, 1]$, the left hand side is equivalent to

$$|(\boldsymbol{\tau}^*)^\mathrm{T} \{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*) - \nabla^2 l(\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\}|$$

$$= |(\boldsymbol{\tau}^*)^\mathrm{T} \{\nabla^2 l(\tilde{\boldsymbol{\eta}}) - \nabla^2 l(\boldsymbol{\eta}^*)\}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)|$$

$$= \left| \frac{1}{n} \sum_{i=1}^n [b''(\tilde{\boldsymbol{\eta}}^\mathrm{T} \dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i\}] \dot{\boldsymbol{Z}}_i^\mathrm{T} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)(\boldsymbol{\tau}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i \right|$$

$$\leq B \left| \frac{1}{n} \sum_{i=1}^n \xi(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i b''\{(\boldsymbol{\eta}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i\} \dot{\boldsymbol{Z}}_i^\mathrm{T} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \right| \max_i |(\boldsymbol{\tau}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i| \qquad \text{(A53)}$$

$$\lesssim s_\eta \left(\frac{\log p}{n} + \frac{\log n}{p} \right) \left\{ M + O_p \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right) \right\} \qquad \text{(A54)}$$

where the inequality (A53) is based on Assumption 2(4) that $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ with $t_1 = \tilde{\boldsymbol{\eta}}^\mathrm{T} \dot{\boldsymbol{Z}}_i$ and $t = (\boldsymbol{\eta}^*)^\mathrm{T} \dot{\boldsymbol{Z}}_i$. The last inequality (A54) is from the (A15) of Lemma 9 and from the results of Proposition 1.

As $n, p \to \infty$, under the scaling condition that $(s_w \vee s_\eta)(n^{-1/2} \log p + p^{-1/2} n^{1/2} \log n) = o_p(1)$, we have condition (iii) holds as

$$\sqrt{n}|(\boldsymbol{\tau}^*)^\mathrm{T} \{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*) - \nabla^2 l(\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\}|$$

$$\lesssim \sqrt{n} s_\eta \left(\frac{\log p}{n} + \frac{\log n}{p} \right) \left\{ M + O_p \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}} \right) \right\}$$

$$= o_p(1).$$

Proof of Condition (iv). We multiply $n^{1/2}$ to the left hand side of condition (iv) and apply mean value theorem with $\tilde{\boldsymbol{\eta}} = \xi\hat{\boldsymbol{\eta}} + (1-\xi)\boldsymbol{\eta}^*$. Then we have

$$
n^{1/2}|(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\{\nabla l(\hat{\boldsymbol{\eta}}) - \nabla l(\boldsymbol{\eta}^*)\}|
$$

$$
= n^{1/2}\left|(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n} b'(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\}\dot{\boldsymbol{Z}}_i\right]\right|
$$

$$
= n^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n} b''(\tilde{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\right|
$$

$$
\leq n^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i\}^2\right|^{1/2}
$$

$$
\times \left|\frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}^2\right|^{1/2} \tag{A55}
$$

$$
\lesssim n^{1/2}\left\{s_\eta\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\}^{1/2}\left\{(s_\eta \vee s_w)\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\}^{1/2} \tag{A56}
$$

$$
\lesssim n^{1/2}(s_w \vee s_\eta)\left(\frac{\log p}{n} + \frac{\log n}{p}\right) = o_p(1), \tag{A57}
$$

where (A55) is by applying Cauchy-Schwarz inequality, (A56) is from the (A15) and (A16) in Lemma 9 and (A57) is from the condition $(s_w \vee s_\eta)(n^{-1/2}\log p + p^{-1/2}n^{1/2}\log n) = o_p(1)$. This completes the proof of Lemma 11.

G.4 Proof of Lemma 12

According to the definitions in the preliminaries, we have

$$
(\boldsymbol{\tau}^*)^{\mathrm{T}}\nabla l(\boldsymbol{\eta}^*)(I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}
$$

$$
= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}](I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}.
$$

$$
= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}](I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}
$$

$$
+ \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}](I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}
$$

$$
= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}^*)^{\mathrm{T}}\boldsymbol{Z}_i[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}](I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}
$$

$$
+ \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_u^*)^{\mathrm{T}}(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}](I_{\theta|\boldsymbol{\zeta}}^*)^{-1/2}
$$

$$+ \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\tau}^*)^{\mathrm{T}} \boldsymbol{Z}_i [b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}_u^*)^{\mathrm{T}} (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2}$$

$$=: G_1 + G_2 + G_3 + G_4$$

Because $|(\boldsymbol{\tau}^*)^{\mathrm{T}} \boldsymbol{Z}_i|$ is bounded, $[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}]$ is sub-exponential from Assumption 2, the term $(\boldsymbol{\tau}^*)^{\mathrm{T}} \boldsymbol{Z}_i[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2}$ are independent and has finite moments. We apply Berry-Esseen Theorem and show that $G_1 \to_d N(0, 1)$

For G_2, we apply similar techniques as in the proof of Lemma 7 condition(i) by Bernstein's inequality and it can be verified that

$$G_2 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}_u^*)^{\mathrm{T}} (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2} = O_p\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right).$$

As a result of Proposition 1, we have

$$G_3 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\tau}^*)^{\mathrm{T}} \boldsymbol{Z}_i[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2} = O_p\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right),$$

and

$$G_4 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}_u^*)^{\mathrm{T}} (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)[b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\} - b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\}](I_{\theta|\zeta}^*)^{-1/2} = O_p\left(\frac{1}{n} + \frac{\log n}{p}\right).$$

Under the scaling condition that $n, p \to \infty$ and $(s_w \vee s_\eta)(n^{-1/2} \log p + p^{-1} n^{1/2} \log n) = o_p(1)$, we show that $G_2 \to_p 0$, $G_3 \to_p 0$ and $G_4 \to_p 0$. Applying Slutsky's Theorem, we have

$$\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\tau}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i[-y_i + b'\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \dot{\boldsymbol{Z}}_i\}](I_{\theta|\zeta}^*)^{-1/2} \to_d N(0, 1).$$

This completes the proof of central limit theorem for the score function.

G.5 Proof of Lemma 13

In this proof, we will show the partial information

$$I_{\theta|\zeta}^* = E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}} \boldsymbol{Z}_i\} D_i \{D_i - (\boldsymbol{w}^*)^{\mathrm{T}} \boldsymbol{M}_i\}],$$

is consistently estimated by

$$\hat{I}_{\theta|\zeta} = \frac{1}{n} \sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}} \dot{\boldsymbol{Z}}_i) D_i (D_i - \hat{\boldsymbol{w}}^{\mathrm{T}} \dot{\boldsymbol{M}}_i).$$

To see this, we write the difference between them as

$$\hat{I}_{\theta|\boldsymbol{\zeta}} - I_{\theta|\boldsymbol{\zeta}}^* = \frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)D_i^2 - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2]$$

$$+\frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_iD_i - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{M}_iD_i]$$

$$= L_1 + L_2.$$

For L_1, we decompose it into

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)D_i^2 - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2 - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2]$$

$$+\frac{1}{n}\sum_{i=1}^{n}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]D_i^2$$

$$+\frac{1}{n}\sum_{i=1}^{n}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]D_i^2$$

$$= L_{11} + L_{12} + L_{13}.$$

Applying similar techniques as in the proof of Lemma 7 condition (ii) by Bernstein inequality, we can show that

$$L_{11} = \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2 - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}D_i^2] = O_p\left(\frac{1}{\sqrt{n}}\right).$$

As a result of Proposition 1,

$$L_{12} = \frac{1}{n}\sum_{i=1}^{n}[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\} - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}]D_i^2 = O_p\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right).$$

Using the estimation consistency results in Theorem 2, we have

$$L_{13} = \frac{1}{n}\sum_{i=1}^{n}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}]D_i^2$$

$$= O_p\left\{s_{\eta}\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log n}{p}}\right)\left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right)\right\}$$

Under the condition that $n, p \to \infty$ and $(s_w \vee s_{\eta})\,(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$, we have

$$L_1 = L_{11} + L_{12} + L_{13} = o_p(1). \tag{A58}$$

For L_2, we decompose it into

$$
\begin{aligned}
L_2 &= \frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\hat{\boldsymbol{w}}^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{M}_i D_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}](\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{M}_i D_i] \\
&= L_{21} + L_{22} + L_{23}.
\end{aligned}
$$

We apply Hölder's inequality on L_{21} and get

$$
\begin{aligned}
L_{21} &= \frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i \\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)\{(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i\}^2\right]^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i)D_i^2\right\}^{1/2} \\
&\lesssim \left\{(s_\eta \vee s_w)\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\}^{1/2}\left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right)^{1/2} \qquad (\mathrm{A}59)
\end{aligned}
$$

where the inequality (A59) is from the the equation (A17) in Lemma 9 and similar arguments as in bounding $L_{12} + L_{13}$.

For L_{22}, we apply Assumption 2(3) that $|b''(t_1) - b''(t)| \leq B|t_1 - t|b''(t)$ with $t_1 = \hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$ and $t = (\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i$. Then we apply Hölder's inequality and have

$$
\begin{aligned}
L_{22} &= \frac{1}{n}\sum_{i=1}^{n}[b''(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\dot{\boldsymbol{Z}}_i) - b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}](\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i \\
&\leq \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i(\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i \\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}\{(\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i\}^2 D_i^2\right]^{1/2} \\
&\lesssim \left\{s_\eta\left(\frac{\log p}{n} + \frac{\log n}{p}\right)\right\}^{1/2}\left(M + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{p}}\right)^{3/2} \qquad (\mathrm{A}60)
\end{aligned}
$$

where the inequality (A60) is from the results (A15) in Lemma 9 and from similar arguments as in bounding L_{11}.

For L_{23}, we follow the similar arguments in bounding L_{11} and get

$$
\begin{aligned}
L_{23} &= \frac{1}{n}\sum_{i=1}^{n} b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\dot{\boldsymbol{Z}}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\dot{\boldsymbol{M}}_i D_i - E[b''\{(\boldsymbol{\eta}^*)^{\mathrm{T}}\boldsymbol{Z}_i\}(\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{M}_i D_i] \\
&\lesssim \frac{1}{n} + \frac{\log n}{p}.
\end{aligned}
\tag{A61}
$$

Under the condition that $n, p \to \infty$ and $(s_w \vee s_\eta)\,(n^{-1/2}\log p + p^{-1}n^{1/2}\log n) = o_p(1)$ and (A59), (A60), (A61), we get

$$
L_2 = L_{21} + L_{22} + L_{23} = o_p(1)
\tag{A62}
$$

Combining (A58) and (A62), we have

$$
\hat{I}_{\theta|\boldsymbol{\zeta}} - I^*_{\theta|\boldsymbol{\zeta}} = o_p(1).
$$

This completes the proof of Lemma 13.

References

Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

T. W. Anderson and Yasuo Amemiya. The Asymptotic Normal Distribution of Estimators in Factor Analysis under General Conditions. *The Annals of Statistics*, 16(2):759–771, 1988.

Jennifer L. Ather and Matthew E. Poynter. Serum Amyloid A3 is required for normal weight and immunometabolic function in mice. *PLOS ONE*, 13(2):1–14, 2018.

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

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, 2012.

Jushan Bai and Kunpeng Li. Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, 98(2):298–309, 2016.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619, 2016.

Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26(5): 2333–2355, 2017.

T. Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, 118(542):1319–1332, 2023.

Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020.

Yunxiao Chen and Xiaoou Li. Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, 109(3):769–782, 2021.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532):1756–1770, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Anna B Costello and Jason Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1):7, 2005.

Alexis Dinno. Exploring the sensitivity of horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, 44(3):362–388, 2009.

Edgar Dobriban. Permutation methods for factor analysis and PCA. *The Annals of Statistics*, 48(5):2824–2847, 2020.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4), 2013.

Jianqing Fan, Fang Han, and Han Liu. Challenges of Big Data analysis. *National Science Review*, 1(2):293–314, 2014.

Jianqing Fan, Yuan Ke, and Kaizheng Wang. Factor-adjusted regularized model selection. *Journal of Econometrics*, 216(1):71–85, 2020.

Jianqing Fan, Zhipeng Lou, and Mengxin Yu. Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association, in press*, 2023.

Zoe Fewell, George Davey Smith, and Jonathan A. C. Sterne. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 166(6):646–655, 2007.

Johann Gagnon-Bartsch and Terence Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics (Oxford, England)*, 13:539–52, 2011.

Zijian Guo, Hyunseung Kang, T. Tony Cai, and Dylan S. Small. Confidence intervals for causal effects with invalid instruments by using two–stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B*, 80(4):793–815, 2018.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, 50(3):1320–1347, 2022.

James C Hayton, David G Allen, and Vida Scarpello. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7(2):191–205, 2004.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

Ji-Su Jang, Jun-Ho Lee, Nam-Chul Jung, So-Yeon Choi, Soo-Yeoun Park, Ji-Young Yoo, Jie-Young Song, Han Geuk Seo, Hyun Soo Lee, and Dae-Seog Lim. RSAD2 is necessary for mouse dendritic cell maturation via the IRF7-mediated signaling pathway. *Cell Death & Disease*, 9(8):823, 2018.

Jana Jankova and Sara van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Hyunseung Kang, Anru Zhang, T. Tony Cai, and Dylan S. Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Alois Kneip and Pascal Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011.

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y Weiss-Solís, Robin Duque, Hugues Bersini, and Ann Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics*, 14(4):469–490, 2013.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

Dan Yu Lin and Zhiliang Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

Wei Lin and Jinchi Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

Zhixiang Lin, Can Yang, Ying Zhu, John Duchi, Yao Fu, Yong Wang, Bai Jiang, Mahdi Zamanighomi, Xuming Xu, Mingfeng Li, et al. Simultaneous dimension reduction and adjustment for confounding variation. *Proceedings of the National Academy of Sciences*, 113(51):14662–14667, 2016.

Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, 2010.

Wei Liu, Huazhen Lin, Shurong Zheng, and Jin Liu. Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118(542):1385–1401, 2023.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015.

Rong Ma, T. Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 116(534):984–998, 2021.

J. Müller-Berghaus, K. Kern, A. Paschen, X D Nguyen, H. Klüter, G. Morahan, and D. Schadendorf. Deficient IL-12p70 secretion by dendritic cells based on IL12B promoter genotype. *Genes & Immunity*, 5(5):431–434, 2004.

A. Nemetz, M. Pilar Nosti-Escanilla, Tamás Molnár, Adorján Köpe, Ágota Kovács, János Fehér, Zsolt Tulassay, Ferenc Nagy, M. Asunción García-González, and A. S. Peña. IL1B gene polymorphisms influence the course and severity of inflammatory bowel disease. *Immunogenetics*, 49(6):527–531, 1999.

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.

Art B. Owen and Jingshu Wang. Bi-Cross-Validation for Factor Analysis. *Statistical Science*, 31(1):119–139, 2016.

Debashis Paul, Eric Bair, Trevor Hastie, and Robert Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618, 2008.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77, 2010.

Pedro R. Peres-Neto, Donald A. Jackson, and Keith M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Ahmed A Quadeer, Raymond HY Louie, Karthik Shekhar, Arup K Chakraborty, I-Ming Hsing, and Matthew R McKay. Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis c virus nonstructural protein 3 exposes targets for immunogen design. *Journal of Virology*, 88(13):7628–7644, 2014.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510 (7505):363–369, 2014.

Chengchun Shi, Rui Song, Wenbin Lu, and Runze Li. Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, 116(535):1307–1318, 2021.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Yujiro Toyoshima, Hidemitsu Kitamura, Huihui Xiang, Yosuke Ohno, Shigenori Homma, Hideki Kawamura, Norihiko Takahashi, Toshiya Kamiyama, Mishie Tanino, and Akinobu Taketomi. IL6 modulates the immune status of the tumor microenvironment to facilitate metastatic colonization of colorectal cancer cells. *Cancer Immunology Research*, 7(12): 1944–1957, 2019.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Fa Wang. Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions. *Journal of Econometrics*, 229(1):180–200, 2022.

Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of statistics*, 45(3):1342, 2017.

Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

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.

Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2020.

William R. Zwick and Wayne F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3):432–442, 1986.