Inference of Heterogeneous Treatment Effects Using Observational Data with High-Dimensional Covariates

Yumou Qiu

Iowa State University, Ames, Iowa, USA.

Jing Tao

University of Washington, Seattle, Washington, USA.

Xiao-Hua Zhou

Peking University, Beijing, China.

E-mail: azhou@uw.edu

Summary. This study proposes novel estimation and inference approaches for heterogeneous local treatment effects using high-dimensional covariates and observational data without a strong ignorability assumption (Rosenbaum and Rubin, 1983). To achieve this, lasso estimation under a non-convex objective function is developed for a two-stage regression model. With a binary instrumental variable, the parameters of interest are identified on an unobservable subgroup of the population (compliers). A debiased estimator is proposed to construct confidence intervals for treatment effects conditioned on covariates. Notably, this approach simultaneously corrects the biases due to high-dimensional estimations at both stages. The finite sample performance is evaluated via extensive simulation studies, and real data analysis is performed on the Oregon Health Insurance Experiment to illustrate the feasibility of the procedures. This approach can be used for both continuous and categorical response variables under the framework of generalized linear models.

Keywords: causal inference; high-dimensional data; instrumental variable; non-convexity; two-stage regression.

1. Introduction

1.1 Overview. At present, estimating the effect of a treatment on some outcome after conditioning on a vector of covariates is one of the focus areas of causal inference and program evaluation research. Statistical inference of covariate-specific

treatment effects is particularly important in the recently emerging fields of personalized medicine and dynamic treatment regime. With observational data, a common requirement for identification is a strong ignorable treatment assignment assumption (Rosenbaum and Rubin, 1983). This means that given a set of observed covariates, the treatment assignment (the nature of data collection) is independent of the potential outcomes. However, this assumption could be violated in empirical applications. When it does not hold, the selection problem must be solved by distinguishing the treatment effect from the confounding effects generated by the treatment selection. For observational data with fixed-dimensional covariates, a variety of methods based on instrumental variables (IVs) have been proposed to overcome the selection problem (Imbens and Angrist, 1994; Angrist et al., 1996; Heckman and Vytlacil, 1999; Abadie, 2003; Tan, 2006). These methods primarily focus on the local average treatment effect (LATE), and are widely used in statistics, economics, and other social sciences. Further details can be found in reviews in Angrist and Pischke (2008) and Imbens and Rubin (2015).

The present work focuses on heterogenous treatment effects using observational data with high-dimensional covariates and endogeneity. Novel estimation and inference methods are developed for treatment-covariate interaction effects and covariate-specific treatment effects with the help of an instrumental variable to deal with the endogeneity. The covariate-specific treatment effects represent the expected difference between potential outcomes given a set of covariates. The instrument induces exogeneity between the treatment and the potential outcomes given the covariates under the "complier" subgroup of the population. The identification and estimation of the local treatment effect on compliers with fixed-dimensional data were studied extensively in Imbens and Angrist (1994); Abadie (2003); Tan (2006); Hong and Nekipelov (2010); Ogburn et al. (2015). However, the existing methods for fixed-dimensional covariates do not yield valid inference in data with high-dimensional covariates. The challenges in the causal inference problem considered herein are multi-fold. Firstly, the compliers are unobserved. Following the identification technique used in Abadie (2003), a two-stage regression approach, in which both stages involve high-dimensional covariates, is needed. Secondly, the sample objective function is non-convex in optimization parameters. The unconstrained estimation may yield unstable estimates even under fixed-dimensional cases. Finally, the second-stage estimation involves plug-in estimates from the first-stage model, which contributes additional variation in the second-stage estimator.

To tackle these challenges, under the framework of generalized linear models (GLMs), this study proposes regularized estimation for each regression coefficient

under a non-convex objective function. Further, its consistency is theoretically established. Based on the initial regularized estimator, a debiased estimator is proposed for the regression coefficients, which eliminates the impact of regularization bias from both first- and second-stage regressions. The proposed procedure can be considered as a debiased lasso estimation (van de Geer et al., 2014) under a non-convex two-stage model. The asymptotic normality results are provided for both the debiased estimator and its functionals. Based on these results, confidence intervals could be constructed for the treatment, the covariates of interest, their interaction effects and the covariate-specific treatment effects. The proposed method can be applied to both continuous and categorical responses, corresponding to linear and non-linear second-stage regression models, respectively. Finally, the proposed methods are evaluated and illustrated using simulations and real data from the Oregon Health Insurance Experiment (OHIE). The heterogeneous effects of insurance on healthcare use and well-being of low-income adults are explored.

The main contributions of this work are as follows. (i) A regularized two-stage estimation procedure is proposed for models on the compliers under data endogeneity. It addresses the instability of unconstrained estimation due to non-convexity. The consistency of the regularized estimator under the non-convex objective function is established. (ii) A novel approach to simultaneously correct the biases due to regularized estimation at both stages is proposed. (iii) A novel statistical inference procedure based on the debiased estimator is developed for covariate effects and (local) heterogeneous treatment effects with high-dimensional data. To the best of our knowledge, there have been no conclusive studies that construct debiased lasso estimators under a two-stage model with a *non-convex* objective function. The statistical inference results for this type of model are of independent interest in studies with high-dimensional data.

1.2 Related literature. The interaction effects and covariate-specific treatment effects are key areas of interest in personalized medicine and social science. They reveal how the effect of treatment depends on participants' characteristics, and enable the selection of accurate, individualized treatment plans (Schulte et al., 2014). Substantial research have been conducted on modeling nonlinear and interaction effects (Chipman et al., 1998; Tian and Tibshirani, 2010; Zhao et al., 2012) as well as heterogeneous treatment effects in randomized trials (Bonetti and Gelber, 2004; Tian et al., 2014; Ma and Zhou, 2017). The present study builds upon this by offering novel inference procedures for observational data and non-randomized trials without the strong ignorability assumption.

Considerable efforts have been made to study high-dimensional observational data (Belloni et al., 2014, 2017a; Cattaneo et al., 2019). In particular, Belloni et al. (2017a) provided efficient estimators with honest confidence bands for LATE and local quantile treatment effects under high-dimensional covariates. The parameters of interest in the present study differ from this line of work. LATE considers the single average treatment effect on compliers over the joint distribution of the covariates, whereas this study focuses on high-dimensional covariate and interaction effects and covariate-specific treatment effects. Studying heterogeneous treatment effects conditional on covariates requires new orthogonal scores compared with those in Belloni et al. (2017a). Additionally, methodological innovation is needed for inference procedures of the target parameters.

Under the framework of high-dimensional GLMs, several pioneering works, including Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014), developed inference procedures for regression coefficients using convex objective functions with lasso penalty. Loh and Wainwright (2012); Belloni et al. (2017b); Datta et al. (2017); Loh (2017) studied penalized estimation under non-convex objective functions. The present study requires estimation and inference for a two-stage non-convex optimization problem, which largely remains an open question in the current literature.

1.3 Organization. The remainder of this paper is organized as follows. Section 2 introduces the proposed model, the parameters of interest, and provides comparisons with existing approaches. Section 3 presents the proposed estimator for a two-stage non-convex regression model. Section 4 describes the debiased estimator and its inference procedure, while Section 5 provides their theoretical properties. Sections 6 and 7 discuss the conducted simulation study and real data study, respectively. All the technical proofs are relegated to the Supplementary Material (SM).

2. Model and Background

The data in this study consist of n observations on an outcome variable Y, a binary treatment indicator D, a binary instrument Z and p-dimensional covariates X. For example, for the OHIE data in Section 7, Y represents different health outcomes, D indicates whether or not a low-income person has health insurance, and Z is the indicator of insurance eligibility. In this experiment, Z and D are not equal because not every person eligible to apply for health insurance actually purchased it. The covariates X consist of individual demographic, social, and financial information.

Let $Y(0)$ and $Y(1)$ be the potential outcomes when $D = 0$ and 1, respectively, and $Y = Y(1)D + Y(0)(1 - D)$. Let $D(0)$ and $D(1)$ be the treatment under $Z = 0$ and 1, respectively, and $D = D(1)Z + D(0)(1 - Z)$. Let $Y(z, d)$ be the response variable under $Z = z$ and $D = d$ for $z = \{0, 1\}$ and $d = \{0, 1\}$. Following the terminology in the literature, the population is divided into four different subgroups by $D(0)$ and $D(1)$ as follows. Let *always-takers* and *never-takers* be defined by $D(0) = D(1) = 1$ and $D(0) = D(1) = 0$, respectively. Further, let *compliers* and *defiers* be those having $D(0) = 0$ and $D(1) = 1$ and those having $D(0) = 1$ and $D(1) = 0$, respectively. The goal of causal inference is to learn about the characteristics of the distribution of $\{Y(1), Y(0)\}$ for the complier subgroup.

2.1. Assumptions and Identification

Because the compliers with $\{D(1) > D(0)\}$ are unobservable, the following assumption is made on the identification of the treatment effect, which is thoroughly discussed in Imbens and Angrist (1994) and Angrist et al. (1996).

Assumption 1. (i) Independence of the instrument: conditional on X, the random vector $\{Y(0,0), Y(0,1), Y(1,1), D(0), D(1)\}$ is independent of Z; (ii) exclusion of the instrument: $\mathbb{P}\{Y(1, d) = Y(0, d)|X\} = 1$ for $d \in \{0, 1\}$; (iii) first-stage: $0 < \mathbb{P}(Z = 1|X) < 1$ and $\mathbb{P}\{D(1) = 1|X\} > \mathbb{P}\{D(0) = 1|X\}$; and (iv) monotonicity: $\mathbb{P}\{D(1) \geq D(0)|X\} = 1$.

Let the local average response function (LARF) be $\mathbb{E}\{Y|X, D, D(1) > D(0)\}$. This function describes the average response for compliers given the treatment variable and the covariates. It is equal to $\mathbb{E}\{Y(0)|X, D(1) > D(0)\}$ and $\mathbb{E}\{Y(1)|X, D(1) > D(0)\}$ when $D = 0$ and $D = 1$, respectively, under Assumption 1. The local covariate-specific treatment effect (LCSTE) function is defined as

$$\text{LCSTE}(X) = \mathbb{E}\{Y(1) - Y(0)|X, D(1) > D(0)\}, \tag{2.1}$$

which gives the expected treatment effect for compliers conditional on covariates.

To study heterogeneous treatment effects, a parametric model is assumed for the LARF such that for some known function $h(\cdot)$ and unknown parameters θ_0,

$$\mathbb{E}\{Y|X, D, D(1) > D(0)\} = h(D, X; \theta_0). \tag{2.2}$$

Under Assumption 1 and (2.2), $\text{LCSTE}(X) = h(1, X; \theta_0) - h(0, X; \theta_0)$. This study focuses on the estimation and inference of the unknown parameter θ_0 and its functionals. Let $W = (D, X^{\mathrm{T}}, DX^{\mathrm{T}})^{\mathrm{T}}$ with $\theta = (\alpha, \beta^{\mathrm{T}}, \delta^{\mathrm{T}})^{\mathrm{T}}$ and $W = (D, X^{\mathrm{T}})^{\mathrm{T}}$ with

$\theta = (\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}$ when the interactions between the treatment and covariates are included and absent in the model, respectively. Here, α, β and δ denote the treatment effect, the main covariates effects, and the interaction effects between the covariates and the treatment, respectively. We could also consider the interaction terms and higher order terms among the covariates. However, the empirical performance of lasso may deteriorate as more interactions are added to the model. In the real data study considered in Section 7, the high dimensionality is mainly in the main covariate effects. The followings are three examples of the LARF and LCSTE(X).

Example 1. Linear models on compliers:

$$\mathbb{E}\{Y|X, D, D(1) > D(0)\} = \alpha_0 D + \beta_0^{\mathrm{T}} X + \delta_0^{\mathrm{T}} DX := W^{\mathrm{T}}\theta_0 \tag{2.3}$$

and $\mathrm{LCSTE}(X) = \alpha_0 + \delta_0^{\mathrm{T}} X$, where $\theta_0 = (\alpha_0, \beta_0^{\mathrm{T}}, \delta_0^{\mathrm{T}})^{\mathrm{T}}$.

The approach proposed herein can be applied to GLMs. Suppose the conditional density of the response Y on the compliers takes the form

$$f\{y|X, D, D(1) > D(0); \theta_0\} = \exp[\{y\eta - b(\eta)\}/\phi + c(y, \phi)] \quad \text{for} \quad \eta = W^{\mathrm{T}}\theta_0, \tag{2.4}$$

where ϕ is the dispersion parameter, $b(\eta)$ and $c(y, \phi)$ are smooth functions, and θ_0 represents the true regression coefficients on compliers. If the dispersion parameter is known, this is the natural exponential family with η as the natural parameter. Under (2.4), the relationship between Y and W is modeled as $g_c[\mathbb{E}\{Y|X, D, D(1) > D(0)\}] = \eta = W^{\mathrm{T}}\theta_0$, where $g_c(\cdot)$ is the canonical link function, the inverse of the derivative of $b(\eta)$. To simplify the notation, write $f(y|X, D; \theta_0) = f\{y|X, D, D(1) > D(0); \theta_0\}$. Note that $f(y|X, D; \theta_0)$ may not be the density function for all the data. The exponential dispersion family (2.4) is widely used in nonlinear regression models. The following examples are special cases of (2.4).

Example 2. Logistic regression models on compliers: $\mathbb{P}\{Y = 1|X, D, D(1) > D(0)\} = \Lambda(\alpha_0 D + \beta_0^{\mathrm{T}} X + \delta_0^{\mathrm{T}} DX)$, and $\mathrm{LCSTE}(X) = \Lambda\{\alpha_0 + (\beta_0 + \delta_0)^{\mathrm{T}} X\} - \Lambda(\beta_0^{\mathrm{T}} X)$ for $\Lambda(a) = \exp(a)/\{1 + \exp(a)\}$.

Example 3. Poisson regression models on compliers: $Y|X, D, D(1) > D(0) \sim \mathrm{Poisson}(\lambda)$ for $\lambda = \exp(\alpha_0 D + \beta_0^{\mathrm{T}} X + \delta_0^{\mathrm{T}} DX)$, and $\mathrm{LCSTE}(X) = \exp\{\alpha_0 + (\beta_0 + \delta_0)^{\mathrm{T}} X\} - \exp(\beta_0^{\mathrm{T}} X)$.

To identify θ_0 in the LARF, let $g(Y, D, X; \theta) = (Y - \theta^{\mathrm{T}} W)^2$ for the linear model in (2.3) or $g(Y, D, X; \theta) = -\log\{f(Y|X, D; \theta)\}$, the negative log likelihood function, for the GLM in (2.4). Define the weight function

$$\kappa := \kappa^{(0)}\mathbb{P}(Z = 0|X) + \kappa^{(1)}\mathbb{P}(Z = 1|X) = 1 - \frac{D(1 - Z)}{\mathbb{P}(Z = 0|X)} - \frac{(1 - D)Z}{\mathbb{P}(Z = 1|X)}, \tag{2.5}$$

where $\kappa^{(0)} = (1-D)\frac{\mathbb{P}(Z=1|X)-Z}{\mathbb{P}(Z=0|X)\mathbb{P}(Z=1|X)}$ and $\kappa^{(1)} = D\frac{Z-\mathbb{P}(Z=1|X)}{\mathbb{P}(Z=0|X)\mathbb{P}(Z=1|X)}$. From Theorem 3.1 in Abadie (2003), under Assumption 1, there exists a relationship between the conditional expectation (given compliers) and the unconditional expectation (over the entire population) such that the true parameter θ_0 is identified by

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathbb{E}\{g(Y,D,X;\theta)|D(1) > D(0)\} = \arg\min_{\theta\in\Theta} \mathbb{E}\{\kappa g(Y,D,X;\theta)\}, \quad (2.6)$$

where $g(\cdot)$ is any measurable real function of (Y, D, X). Notice that κ is 1 when $D = Z$ and takes a negative value when $D \neq Z$. Intuitively, the negative values of κ should offset non-compliers in the unconditional expectation in (2.6). However, those negative weights also make the sample objective function of (2.6) non-convex, imposing challenges in the estimation of θ_0.

2.2. Connection to Existing Methods

Under Assumption 1, Imbens and Angrist (1994); Angrist et al. (1996); Tan (2006) have shown that LATE on the compliers can be identified via the ratio

$$\mathbb{E}\{Y(1) - Y(0)|D(1) > D(0)\} = \frac{\mathbb{E}\{\mathbb{E}(Y|Z=1,X)\} - \mathbb{E}\{\mathbb{E}(Y|Z=0,X)\}}{\mathbb{E}\{\mathbb{E}(D|Z=1,X)\} - \mathbb{E}\{\mathbb{E}(D|Z=0,X)\}}, \quad (2.7)$$

where the outer expectations in the ratio are taken respect to the covariates X. Similarly, LCSTE(X) can be identified by the same ratio without the expectations for X. Based on (2.7), LATE can be estimated by inverse propensity score weighting, using an estimated $\mathbb{P}(Z = 1|X)$ for the denominator and numerator in (2.7) (Tan, 2006). This method only requires modeling $\mathbb{P}(Z = 1|X)$; however, it cannot be used to estimate the covariate-specific effects LCSTE(X) given X. Tan (2006) also proposed a regression approach for LCSTE(X) based on parametrical models for $\mathbb{P}(D = 1|Z, X)$ and $\mathbb{E}(Y|D = d, Z, X)$ for $d = 0, 1$. Note that both the aforementioned methods are for fixed-dimensional data, which cannot be applied to high-dimensional covariates.

Belloni et al. (2017a) combined both regression and inverse weighting approaches to estimate LATE under a high-dimensional setting. Their approach is based on regularized regression estimation (or consistent machine learning estimation) for $\mathbb{P}(Z = 1|X)$ and four conditional expectations $\mathbb{E}(Y|Z = z, X)$ and $\mathbb{E}(D|Z = z, X)$ for $z = 0, 1$. Those regressions do not involve plug-in estimates and can be fitted by regularized convex optimization, which are advantages for this approach. In fact, their estimator has the same structure as the doubly robust estimator for LATE discussed in Tan (2006) under fixed-dimensional settings. However, the doubly robust inference for this estimator has not been studied with high-dimensional control

variables. Despite its flexibility in constructing estimators for the aforementioned conditional expectations, it is not clear whether the inference procedure for LATE can be extended to LCSTE(X) for covariate-specific treatment effects.

This study employs a different modeling approach as compared to Tan (2006) and Belloni et al. (2017a). Here, a GLM for the response is directly built on the compliers. This approach does not require estimating the four conditional expectations of Y and D given Z and X. In stead, the parsimonious model for LARF is estimated, which enables easy interpretation for the impact of covariates and treatment on LARF and LCSTE(X). Notice that the covariate effect on LATE is not straight forward from (2.7). Further discussions on the comparison between modeling on the compliers versus modeling the overall conditional expectations of Y and D can be found in Abadie (2003); Ogburn et al. (2015).

One disadvantage of a parametric approach is misspecification. Ogburn et al. (2015) considered a parametric model for LCSTE(X) on the covariates X or a subset of X with an identification equation based on the weight $\kappa^{(1)} - \kappa^{(0)}$. Under a fixed-dimensional setting, they proposed a doubly robust estimation method. Specifically, as long as the model of LCSTE(X) is correct, their method can produce consistent estimators for its regression coefficients, even if the propensity score $\mathbb{P}(Z = 1|X)$ is mis-specified. However, to satisfy Assumption 1, applied researchers generally need to consider a large number of covariates X. This is because the treatment may be confounded, in the sense that Z may not be independent of the potential outcome variables given a limited number of covariates (Lee et al., 2017). Thus, a large set of conditioning variables generally need to be employed. This requires statistical inference approaches for LCSTE(X) with high-dimensional covariates. As an extension of the proposed approach, doubly robust estimation for high-dimensional data is discussed in Section 8.

3. Estimation Method

We begin with notation and definitions. Throughout the paper, for a vector $a = (a_1, \ldots, a_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $1 \leq d < \infty$, let $|a|_d = (\sum_{i=1}^p |v_i|^d)^{1/d}$ be the vector ℓ_d norm, and $|a|_\infty = \max_{1\leq j\leq p} |a_j|$ be the vector maximum norm. For a matrix $A = (a_{j_1 j_2}) \in \mathbb{R}^{p\times q}$, the elementwise ℓ_1 and ℓ_∞ norm are defined as $\|A\|_1 = \sum_{j_1=1}^p \sum_{j_2=1}^q |a_{j_1 j_2}|$ and $\|A\|_\infty = \max_{1\leq j_1, j_2\leq q} |a_{j_1 j_2}|$, respectively. The matrices ℓ_1, ℓ_∞ and ℓ_2 (spectral) norm are denoted as $\|A\|_{\ell_1} = \max_{1\leq j_2\leq q} \sum_{j_1=1}^p |a_{j_1 j_2}|$, $\|A\|_{\ell_\infty} = \max_{1\leq j_1\leq p} \sum_{j_2=1}^q |a_{j_1 j_2}|$ and $\|A\|_{\ell_2} = \sup_{|x|_2\leq 1} |Ax|_2$, respectively. The

Frobenius norm is $\|A\|_F = (\sum_{j_1 j_2} a_{j_1 j_2}^2)^{1/2}$. Let I_p be a $p \times p$ identity matrix. For any random variables $U_1, \ldots, U_n$ and any function $h(\cdot)$, let $\mathbb{E}_n[h(U_i)] = \sum_{i=1}^n h(U_i)/n$ denote the empirical average of $\{h(U_i)\}$. For a function $h(\cdot)$, let $\dot{h}(\cdot)$ and $\ddot{h}(\cdot)$ be its first and second order derivatives.

The proposed estimation consists of two stages. In the first stage, an estimator is developed for κ in (2.5) by estimating $\mathbb{P}(Z = 1|X)$. In the second stage, θ_0 and the smooth functional of the linear combination $w_c^{\mathrm{T}}\theta_0$ are estimated for a pre-specified vector w_c. Without loss of generality, our analysis is based on the likelihood estimation $g(Y, D, X; \theta) = -\log\{f(Y|X, D; \theta)\}$ in (2.6), although least square estimation can be used for the linear model (2.3).

Suppose the observed data $(Y_i, D_i, Z_i, X_i^{\mathrm{T}})_{i=1}^n$ are independent and identically distributed (i.i.d.) random vectors with the conditional density $f(Y_i|X_i, D_i; \theta_0)$ on the compliers satisfying (2.4) and Assumption 1. We assume Z_i follows the logistic regression given X_i where $\mathbb{P}(Z_i = 1|X_i = x_i) = \Lambda(x_i^{\mathrm{T}}\gamma_0)$ for $\Lambda(a) = \exp(a)/\{1 + \exp(a)\}$ and $\gamma_0 = (\gamma_{0,1}, \ldots, \gamma_{0,p})^{\mathrm{T}} \in \mathbb{R}^p$. Other models can be applied in a similar manner. The lasso estimator $\tilde{\gamma}$ of the logistic regression solves

$$\tilde{\gamma} = \arg\min_{\gamma} -\mathbb{E}_n[\rho_\gamma(X_i, Z_i)] + \lambda_1|\gamma|_1 \quad \text{for} \tag{3.1}$$

$$\rho_\gamma(X_i, Z_i) = Z_i \log \Lambda(X_i^{\mathrm{T}}\gamma) + (1 - Z_i)\log\{1 - \Lambda(X_i^{\mathrm{T}}\gamma)\}. \tag{3.2}$$

The weight κ_i for the ith observation is estimated by plugging $\tilde{\gamma}$ into (2.5),

$$\tilde{\kappa}_i = \kappa_i(\tilde{\gamma}) = \frac{(1 - D_i)\big\{\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}) - Z_i\big\}}{\Lambda(X_i^{\mathrm{T}}\tilde{\gamma})} + \frac{D_i\big\{Z_i - \Lambda(X_i^{\mathrm{T}}\tilde{\gamma})\big\}}{1 - \Lambda(X_i^{\mathrm{T}}\tilde{\gamma})}. \tag{3.3}$$

To avoid the estimated weight being too small, we could set the estimated propensity score to be c_{ps} and $1 - c_{\mathrm{ps}}$ if $\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}) < c_{\mathrm{ps}}$ and $\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}) > 1 - c_{\mathrm{ps}}$, respectively, where c_{ps} is a small positive constant. In the simulation, we set $c_{\mathrm{ps}} = 1/21$ so that the estimated weights $\tilde{\kappa}_i$ are bounded by -20 from below. Notice that trimming is only need for $D_i \neq Z_i$ as $\tilde{\kappa}_i = 1$ if $D_i = Z_i$.

In the second-stage estimation, Because θ_0 is the minimizer of $\mathbb{E}[\kappa g(Y, D, X; \theta)]$, a natural estimator of θ_0 is obtained by minimizing the penalized sample analog of (2.6). Let $\mathcal{A}$ be the index set of important covariates not penalized, and $\theta_{\mathcal{A}^c}$ be the sub-vector of θ with components not in $\mathcal{A}$. A common choice of $\mathcal{A}$ is $\mathcal{A} = \{1\}$, indicating that only the treatment variable is not penalized in the estimation. Let

$$\mathcal{L}_{n,\tilde{\gamma}}(\theta) = \frac{1}{n}\sum_{i=1}^n \kappa_i(\tilde{\gamma}) g(Y_i, D_i, X_i; \theta) \tag{3.4}$$

be the empirical estimate of $\mathbb{E}\{\kappa g(Y, D, X; \theta)\}$. However, the weight $\kappa_i(\tilde{\gamma})$ is negative when $Z_i = 1$ and $D_i = 0$ or $Z_i = 0$ and $D_i = 1$. These negative weights could make the empirical loss function $\mathcal{L}_{n,\tilde{\gamma}}(\theta)$ non-convex, and make the lasso solution of $\arg\min\{\mathcal{L}_{n,\tilde{\gamma}}(\theta) + \lambda_2|\theta_{\mathcal{A}^c}|_1\}$ diverges to $-\infty$.

To solve this problem, following Loh and Wainwright (2012) and Loh (2017), the restricted lasso program is considered as follows:

$$\tilde{\theta} \in \underset{|\theta|_1 \leq R}{\arg\min}\{\mathcal{L}_{n,\tilde{\gamma}}(\theta) + \lambda_2|\theta_{\mathcal{A}^c}|_1\}, \tag{3.5}$$

where λ_2 is the lasso penalty parameter, and R is a positive constant satisfying $R \geq |\theta_0|_1$ to guarantee that the true regression parameter θ_0 is feasible for the minimization program (3.5). The constraint $|\theta|_1 \leq R$ ensures the existence of local/global optima for (3.5). Notice that Theorem 3.1 in Abadie (2003) shows $\mathbb{E}\{\kappa_i(\gamma_0)g(Y_i, D_i, X_i; \theta)\} = \mathbb{E}\{g(Y_i, D_i, X_i; \theta)|D(1) > D(0)\}$, which implies the convexity of the population loss function. The non-convexity of $\mathcal{L}_{n,\tilde{\gamma}}(\theta)$ is due to the negative weight $\kappa_i(\tilde{\gamma})$ on some observations. This cause of non-convexity in the proposed model differs from that of the linear model with measurement errors and missing covariates considered in Loh and Wainwright (2012) and that of the robust linear regression considered in Loh (2017). Moreover, those works only considered one-stage estimation problems, while this study considers a two-stage problem.

In practice, it may not be desired to drop the main covariates effects but keeping their interactions with the treatment. To this end, we could use a combination of the lasso and group lasso penalties in the form

$$p_{\lambda,a_0}(\theta) = \lambda\left\{a_0\sum_{j=1}^{p}|\delta_j| + (1-a_0)\sum_{j=1}^{p}(\beta_j^2 + \delta_j^2)^{1/2}\right\}$$

in (3.5) for $\lambda > 0$ and $a_0 \in (0, 1)$, where β_j is the main effect of the jth covariate and δ_j is its interaction effect with the treatment variable. The group lasso penalty simultaneously penalizes the main effect β_j and the interaction effect δ_j, while the lasso penalty imposes sparsity on the interaction effects.

Specifically, for the linear model in Example 1, let $\hat{M}_{\tilde{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tilde{\gamma})W_iW_i^{\mathrm{T}}$, $\hat{M}_{\gamma_0} = \frac{1}{n}\sum_{i=1}^{n}\kappa_i(\gamma_0)W_iW_i^{\mathrm{T}}$ and $M_{\gamma_0} = \mathbb{E}\{\kappa_i(\gamma_0)W_iW_i^{\mathrm{T}}\}$. The program (3.5) has a specific form

$$\underset{|\theta|_1 \leq R}{\arg\min}\{\theta^{\mathrm{T}}\hat{M}_{\tilde{\gamma}}\theta/2 - \hat{\omega}^{\mathrm{T}}\theta + \lambda_2|\theta_{\mathcal{A}^c}|_1\}, \tag{3.6}$$

where $\hat{\omega} = \sum_{i=1}^{n}\kappa_i(\tilde{\gamma})Y_iW_i/n$. Because of the negative weights, the matrix $\hat{M}_{\tilde{\gamma}}$ could be non-positive-semi-definite, leading to deteriorated solutions over iterative

optimization, despite its population counterpart M_{γ_0} being positive definite. Therefore, the proposed estimator (3.5) obtains the minimum of the objective function within a constraint ℓ_1 ball of θ.

Let p_0 be the dimension of θ and ∇ denote the gradient or subgradient operator of a function. Let $\langle \cdot, \cdot \rangle$ be the inner product operator on the Euclidean space. $\check{\theta}$ is called a stationary point of the optimization problem (3.5) if

$$\langle \nabla \mathcal{L}_{n,\tilde{\gamma}}(\check{\theta}) + \lambda_2 \nabla |\check{\theta}_{\mathcal{A}^c}|_1, \theta - \check{\theta} \rangle \geq 0 \tag{3.7}$$

for all feasible $\theta \in \mathbb{R}^{p_0}$. For $\check{\theta}$ lying in the interior of the ℓ_1 ball $|\theta|_1 < R$, the condition (3.7) becomes the usual zero subgradient condition $\nabla \mathcal{L}_{n,\tilde{\gamma}}(\check{\theta}) + \lambda_2 \nabla |\check{\theta}_{\mathcal{A}^c}|_1 = 0$. Note that the set of stationary points may also include local maxima. Here, we will show that any stationary point of (3.5) in the neighborhood of θ_0 is consistent to θ_0 as $n \to \infty$ in Section 5.

To obtain a stationary solution for the optimization program (3.5), the composite gradient descent is applied. This iteratively solves the following minimization problem:

$$\theta^{t+1} = \underset{|\theta|_1 \leq R}{\arg\min} \big\{ \langle \nabla \mathcal{L}_{n,\tilde{\gamma}}(\theta^t), \theta \rangle + r_u |\theta - \theta^t|_2^2 / 2 + \lambda_2 |\theta_{\mathcal{A}^c}|_1 \big\} \tag{3.8}$$

for a positive constant r_u. The solution to (3.8) can be computed in two steps. In the first step, $\tilde{\theta}^{t+1} = \theta^t - \nabla \mathcal{L}_{n,\tilde{\gamma}}(\theta^t)/r_u$ is updated, and the components of $\tilde{\theta}^{t+1}$ are soft thresholded at the level λ_2/r_u if they are not in $\mathcal{A}$. This is the solution of (3.8) without the constraint $|\theta|_1 \leq R$. If the resulting thresholded vector has ℓ_1 norm greater than R, it is projected onto the ℓ_1 ball centered at zero with radius R, based on the algorithm provided in Duchi et al. (2008).

4. Inference Method

This section describes the construction of confidence intervals for θ_0 and its linear combinations based on the stationary point $\tilde{\theta}$ of the optimization (3.5). Compared to the debiased lasso procedure for linear regression model, inference for the local treatment effect model brings two additional challenges. First, the weights $\{\kappa_i(\gamma_0)\}$ are unknown, and the inference procedure for θ_0 must account for the variation of estimating γ_0 using $\tilde{\gamma}$. Second, the nodewise regression for debiasing $\tilde{\theta}$ cannot be applied owing to the negative weights in $\{\kappa_i(\gamma_0)\}$ and the non-convexity of the empirical objective function $\mathcal{L}_{n,\tilde{\gamma}}(\theta)$.

We first construct the debiased lasso estimator for $\tilde{\gamma}$ in (3.1) in the first-stage regression following the method by van de Geer et al. (2014). Note that, for the logistic regression, the first and second derivatives of the function $\Lambda(\cdot)$ are $\dot{\Lambda} = \Lambda(1-\Lambda)$ and $\ddot{\Lambda} = \dot{\Lambda}(1-2\Lambda)$, where $\dot{\Lambda}(X_i^{\mathrm{T}}\gamma_0)$ is the conditional variance of Z_i given X_i. Let $\mathcal{I}_\gamma = \mathbb{E}\{\dot{\rho}_\gamma(X_i, Z_i)\dot{\rho}_\gamma^{\mathrm{T}}(X_i, Z_i)\} = \mathbb{E}\{\dot{\Lambda}(X_i^{\mathrm{T}}\gamma)X_iX_i^{\mathrm{T}}\}$ be the FI matrix of γ in the first-stage model, where $\rho_\gamma(X_i, Z_i)$ is defined in (3.2). Let $\Xi_\gamma = \mathcal{I}_\gamma^{-1}$ be the inverse of the FI matrix.

Let $g_{i,\gamma}^2 = \dot{\Lambda}(X_i^{\mathrm{T}}\gamma)$ and $G_\gamma = \mathrm{diag}\{g_{1,\gamma}, \ldots, g_{n,\gamma}\}$. Note that $\mathcal{I}_\gamma$ can be viewed as the expected inner product matrix of the weighted covariates $g_{i,\gamma}X_i$. Its inverse Ξ_γ can be estimated by the nodewise regressions (van de Geer et al., 2014; Janková and van de Geer, 2016). Define

$$\hat{\phi}_{\tilde{\gamma},j} = \underset{\phi_j\in\mathbb{R}^p, \phi_{jj}=-1}{\arg\min} |G_{\tilde{\gamma}}\mathbf{X}\phi_j|_2^2/n + 2\lambda_{1,j}(|\phi_j|_1 - 1), \tag{4.1}$$

$$\hat{\tau}_{\tilde{\gamma},j}^2 = |G_{\tilde{\gamma}}\mathbf{X}\hat{\phi}_{\tilde{\gamma},j}|_2^2/n + \lambda_{1,j}(|\hat{\phi}_{\tilde{\gamma},j}|_1 - 1). \tag{4.2}$$

Let $\hat{\Xi}_{\tilde{\gamma}}$ be the estimator of Ξ_{γ_0}, with the jth row being $\hat{\Xi}_{\tilde{\gamma},j} = -\hat{\phi}_{\tilde{\gamma},j}/\hat{\tau}_{\tilde{\gamma},j}^2$. The debiased lasso estimator of $\tilde{\gamma}$ is

$$\hat{\gamma} = \tilde{\gamma} + \hat{\Xi}_{\tilde{\gamma}}\mathbb{E}_n\big[\{Z_i - \Lambda(X_i^{\mathrm{T}}\tilde{\gamma})\}X_i\big]. \tag{4.3}$$

Because of the negative weights in $\{\kappa_i(\tilde{\gamma})\}$, the nodewise regression method cannot be applied to the inference of θ_0. A Taylor expansion of the gradient $\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta)$ of the empirical objective function is considered to construct a debiased estimator of $\tilde{\theta}$ in (3.5) for statistical inference. To this end, we introduce the influence matrix $N_{\theta,\gamma}$ of the first stage estimation and the FI matrix $M_{\theta,\gamma}$ of the second stage regression.

Recall that p_0 is the dimension of θ and W_i. Let

$$z_{\theta,i} = -\frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \theta) = \phi^{-1}\big\{Y_i - \dot{b}(W_i^{\mathrm{T}}\theta)\big\}W_i$$

for $z_{\theta,i} = (z_{\theta,i1}, \ldots, z_{\theta,ip_0})^{\mathrm{T}}$, where $\dot{b}(\cdot)$ denotes the first derivative of $b(\cdot)$. Let

$$\kappa_0(a_1) = \kappa_0(a_1, D, Z) = \frac{(Z-a_1)(a_1+D-1)}{a_1(1-a_1)}$$

such that κ in (2.5) can be written as $\kappa_i(\gamma) = \kappa_0\{\Lambda(X_i^{\mathrm{T}}\gamma)\}$. The influence matrix of the estimator $\tilde{\gamma}$ on estimating θ is $N_{\theta,\gamma} = -\mathbb{E}[z_{\theta,i}\dot{\kappa}_0\{\Lambda(X_i^{\mathrm{T}}\gamma)\}\dot{\Lambda}(X_i^{\mathrm{T}}\gamma)X_i^{\mathrm{T}}]$. Let

$$\hat{N}_{\theta,\gamma} = (\hat{N}_{\theta,\gamma,j_1j_2})_{p_0\times p} = -\frac{1}{n}\sum_{i=1}^n z_{\theta,i}\dot{\kappa}_0\{\Lambda(X_i^{\mathrm{T}}\gamma)\}\dot{\Lambda}(X_i^{\mathrm{T}}\gamma)X_i^{\mathrm{T}}$$

be the sample counterpart of $N_{\theta,\gamma}$, and $\check{N}_{q_0}$ be a regularized estimator of N_{θ_0,γ_0} by thresholding each component $\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}$ of $\hat{N}_{\tilde{\theta},\tilde{\gamma}}$, formulated as

$$\check{N}_{q_0} = \big\{\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}(|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3)\big\}_{p_0\times p} \tag{4.4}$$

for a threshold level $q_0 > 0$. Here, $\lambda_3 = \sqrt{s_1}\max(\sqrt{s_2}, \log p_0)\{\log(p_0)/n\}^{1/2}$, where s_1 and s_2 are the number of the nonzero elements in γ_0 and θ_0, respectively.

Taking the second derivative of $g(Y_i, D_i, X_i; \theta)$ with respect to θ, let

$$M_{\theta,\gamma} = \mathbb{E}\bigg\{\kappa_i(\gamma)\frac{\partial^2}{\partial\theta\partial\theta}g(Y_i, D_i, X_i; \theta)\bigg\} = \mathbb{E}\big\{\phi^{-1}\kappa_i(\gamma)\ddot{b}(W_i^{\mathrm{T}}\theta)W_iW_i^{\mathrm{T}}\big\} \tag{4.5}$$

be the FI matrix of θ on the complier group, where $\ddot{b}(\cdot)$ denotes the second derivative of $b(\cdot)$. Let $B_{\theta_0,\gamma_0} = M^{-1}_{\theta_0,\gamma_0}$. To obtain the estimator of B_{θ_0,γ_0}, the CLIME procedure, originally proposed by Cai et al. (2011) for estimating sparse precision matrices, is applied. Specifically, let $\tilde{B}_{\tilde{\theta},\tilde{\gamma}} = (\tilde{b}_{j_1j_2})$ be the solution of the optimization problem:

$$\min \|B_{\tilde{\theta},\tilde{\gamma}}\|_1 \text{ subject to } \|B_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I_{p_0}\|_\infty \le \lambda_4,\ B_{\tilde{\theta},\tilde{\gamma}} \in \mathbb{R}^{p_0\times p_0}, \tag{4.6}$$

where λ_4 is a tuning parameter that satisfies $\lambda_4 \ge \|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty$, and

$$\hat{M}_{\tilde{\theta},\tilde{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\kappa_i(\tilde{\gamma})\frac{\partial^2}{\partial\theta\partial\theta}g(Y_i, D_i, X_i; \tilde{\theta}) \tag{4.7}$$

is the estimated FI matrix. To obtain a symmetrical estimator, we construct $\hat{B}_{\tilde{\theta},\tilde{\gamma}} = (\hat{b}_{j_1j_2})$ for $\hat{b}_{j_1j_2} = \hat{b}_{j_2j_1} = \tilde{b}_{j_1j_2}\mathbb{I}(|\tilde{b}_{j_1j_2}| \le |\tilde{b}_{j_2j_1}|) + \tilde{b}_{j_2j_1}\mathbb{I}(|\tilde{b}_{j_1j_2}| > |\tilde{b}_{j_2j_1}|)$. The advantage of CLIME lies in the convex optimization (4.6), which can be solved column-by-column (S.2.7) by linear programming despite the matrix $\hat{M}_{\tilde{\theta},\tilde{\gamma}}$ may not be positive semi-definite. However, the solution of graphical lasso is not guaranteed to converge due to the non-positive-semi-definiteness of $\hat{M}_{\tilde{\theta},\tilde{\gamma}}$; see Remark 3 in Mazumder and Hastie (2012). Moreover, CLIME only requires weak sparsity of B_{θ_0,γ_0}, whereas graphical lasso generally needs exact sparsity.

As $\tilde{\gamma}$ and $\tilde{\theta}$ are consistent to γ_0 and θ_0, it can be shown that $\hat{M}_{\tilde{\theta},\tilde{\gamma}}$ is an element-wise consistent estimator of M_{θ_0,γ_0}. The optimization (4.6) follows the CLIME procedure; however, the theoretical justification for $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ is more involved compared to the original CLIME built on sample covariance matrices. This is because of the additional variation introduced by the plugging-in estimate $\tilde{\gamma}$ for γ_0.

By the Taylor expansion of $\kappa_i(\tilde{\gamma})$ at γ_0 and $\frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \theta_0)$ at $\tilde{\theta}$,

$$\begin{aligned}\sqrt{n}(\tilde{\theta} - \theta_0) &+ \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big] \\ &= \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] - \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma} - \gamma_0) + o_p(1).\end{aligned} \tag{4.8}$$

This is shown in the SM. Note that the bias brought by $\tilde{\gamma}$ in the term $\sqrt{n}(\tilde{\gamma}-\gamma_0)$ must also be corrected for the inference of θ_0. Using the de-sparsified/debiased lasso estimator $\hat{\gamma}$ in (4.3) and its asymptotic expansion (given in Lemma 1 in Section 5), we have

$$\begin{aligned}&\sqrt{n}(\tilde{\theta}-\theta_0)+\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big]+\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\hat{\gamma})\\&\quad=\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big]-\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}\Xi_{\gamma_0}\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big]+o_p(1).\quad(4.9)\end{aligned}$$

Motivated by the equality (4.9), a debiased estimator $\hat{\theta}$ for θ_0 is proposed as follows:

$$\hat{\theta}\;=\;\tilde{\theta}+\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big]+\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\hat{\gamma}).\tag{4.10}$$

Based on the expansion result in (4.9) for $\hat{\theta}$, confidence intervals are constructed for θ_0 and its linear combinations. Let

$$V_0=\mathbb{E}\big\{\kappa_i(\gamma_0)z_{\theta_0,i}-N_{\theta_0,\gamma_0}\Xi_{\gamma_0}\dot{\rho}_{\gamma_0}(X_i,Z_i)\big\}\big\{\kappa_i(\gamma_0)z_{\theta_0,i}-N_{\theta_0,\gamma_0}\Xi_{\gamma_0}\dot{\rho}_{\gamma_0}(X_i,Z_i)\big\}^{\mathrm{T}}.\tag{4.11}$$

Then, $V=B_{\theta_0,\gamma_0}V_0B_{\theta_0,\gamma_0}^{\mathrm{T}}$ can be estimated by $\hat{V}=\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{V}_0\hat{B}_{\tilde{\theta},\tilde{\gamma}}^{\mathrm{T}}$ for

$$\hat{V}_0=\frac{1}{n}\sum_{i=1}^{n}\big\{\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}-\check{N}_{q_0}\hat{\Xi}_{\tilde{\gamma}}\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\big\}\big\{\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}-\check{N}_{q_0}\hat{\Xi}_{\tilde{\gamma}}\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\big\}^{\mathrm{T}}.\tag{4.12}$$

The diagonal values $\{\hat{V}_{jj}\}$ of $\hat{V}$ are the estimated variances of $\sqrt{n}\hat{\theta}_j$ for $j=1,\ldots,p_0$. Theorem 2 in Section 5 provides the asymptotic normality of $\sqrt{n}(\hat{\theta}_j-\theta_{0,j})\hat{V}_{jj}^{-1/2}$. Based on this result, the $1-\tau$ confidence interval for $\theta_{0,j}$ can be constructed as

$$(\hat{\theta}_j-z_{\tau/2}\hat{V}_{jj}^{1/2}n^{-1/2},\hat{\theta}_j+z_{\tau/2}\hat{V}_{jj}^{1/2}n^{-1/2}),\tag{4.13}$$

where $z_{\tau/2}$ is the upper $\tau/2$ quantile of the standard normal distribution. Given a vector of covariates w_c, $w_c^{\mathrm{T}}\theta_0$ can be estimated by $w_c^{\mathrm{T}}\hat{\theta}$ with the confidence interval

$$(w_c^{\mathrm{T}}\hat{\theta}-z_{\tau/2}(w_c^{\mathrm{T}}\hat{V}w_c/n)^{1/2},w_c^{\mathrm{T}}\hat{\theta}+z_{\tau/2}(w_c^{\mathrm{T}}\hat{V}w_c/n)^{1/2}).\tag{4.14}$$

Under the linear model in Example 1, $w_c^{\mathrm{T}}\hat{\theta}$ together with (4.14) provide the estimate and confidence interval for the LCSTE, conditioned on $W=w_c$. For the nonlinear models in Examples 2 and 3, given $W=w_c$, the LARF $h(w_c^{\mathrm{T}}\theta_0)$ can be estimated by $h(w_c^{\mathrm{T}}\hat{\theta})$. Recall that $\dot{h}(\cdot)$ is the first order derivative of $h(\cdot)$. The confidence interval of $h(w_c^{\mathrm{T}}\theta_0)$ can be constructed by the delta method as

$$(h(w_c^{\mathrm{T}}\hat{\theta})-z_{\tau/2}|\dot{h}(w_c^{\mathrm{T}}\tilde{\theta})|(w_c^{\mathrm{T}}\hat{V}w_c/n)^{1/2},h(w_c^{\mathrm{T}}\hat{\theta})+z_{\tau/2}|\dot{h}(w_c^{\mathrm{T}}\tilde{\theta})|(w_c^{\mathrm{T}}\hat{V}w_c/n)^{1/2}).\tag{4.15}$$

Let w_0 and w_1 be the values of W under $D = 0$ and $D = 1$ with the same X, respectively. The LCSTE $= h(w_1^{\mathrm{T}}\theta_0) - h(w_0^{\mathrm{T}}\theta_0)$ in Examples 2 and 3 can be estimated by $h(w_1^{\mathrm{T}}\hat{\theta}) - h(w_0^{\mathrm{T}}\hat{\theta})$, with the $1-\tau$ level confidence interval

$$(h(w_1^{\mathrm{T}}\hat{\theta}) - h(w_0^{\mathrm{T}}\hat{\theta}) - z_{\tau/2}(w_a^{\mathrm{T}}\hat{V}w_a/n)^{1/2}, h(w_1^{\mathrm{T}}\hat{\theta}) - h(w_0^{\mathrm{T}}\hat{\theta}) + z_{\tau/2}(w_a^{\mathrm{T}}\hat{V}w_a/n)^{1/2}) \tag{4.16}$$

for $w_a = \dot{h}(w_1^{\mathrm{T}}\tilde{\theta})w_1 - \dot{h}(w_0^{\mathrm{T}}\tilde{\theta})w_0$.

5. Asymptotic Results

This section derives the asymptotic normality of $\hat{\theta}$ and its linear combination $w_c^{\mathrm{T}}\hat{\theta}$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the supports of the regression parameters γ_0 and θ_0, respectively. Let $s_1 = |\mathcal{S}_1|$ and $s_2 = |\mathcal{S}_2|$ be the number of nonzero elements in γ_0 and θ_0. For two positive sequences $\{a_{1,n}\}$ and $\{a_{2,n}\}$, the notation $a_{1,n} \asymp a_{2,n}$ indicates that $a_{1,n}$ and $a_{2,n}$ are of the same order, such that $C_1 \le a_{1,n}/a_{2,n} \le C_2$ for all n and two constants C_1 and C_2. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum eigenvalues of a non-negative definite matrix A. Let $|\mathcal{A}|$ be the cardinality of a set $\mathcal{A}$. Further, let C be a positive constant which may change from case to case.

The asymptotic expansion of $\hat{\gamma}$ is presented in (4.3). A similar result on the de-sparsified/debiased lasso estimator has been given in van de Geer et al. (2014) under GLMs with bounded covariates. Here, the following assumptions are made for the first-stage regression with sub-Gaussian distributed X.

Assumption 2. The vector X_i is a sub-Gaussian random vector such that

$$\sup_{|v|_2 \le 1} \mathbb{E}\big[\exp\{(v^{\mathrm{T}}X_i)^2/C_1\}\big] \le C_2$$

for two positive constants C_1 and $C_2 < \infty$, and $i = 1, \ldots, n$.

Assumption 3. $\log(p) = o(n^{1/5})$ and $s_1\{\log^3(p)/n\}^{1/2} = o(1)$.

Assumption 4. The probabilities $\mathbb{P}(Z = 1|X = X_i) = \Lambda(X_i^{\mathrm{T}}\gamma_0)$ are bounded from 0 and 1, equivalently, $\max_{1\le i\le n} |X_i^{\mathrm{T}}\gamma_0| \le C$, where $\Lambda(a) = \exp(a)/\{1+\exp(a)\}$.

Assumption 5. The FI matrix $\mathcal{I}_{\gamma_0}$ is positive definite, and its values are bounded such that $\|\mathcal{I}_{\gamma_0}\|_\infty \le C$ and $\lambda_{\min}(\mathcal{I}_{\gamma_0}) \ge 1/C$. The number of the nonzero entries in each row of Ξ_{γ_0} is controlled by s_3, where Ξ_{γ_0} is from the family of sparse matrices

$$\mathcal{H}_0(s_3) = \left\{\Xi : \max_{1\le j\le p} |\Xi_{\gamma_0,j}|_0 \le s_3, \lambda_{\max}(\Xi_{\gamma_0}) \le C\right\} \tag{5.1}$$

for $s_3\{\log^3(p)/n\}^{1/2} = o(1)$. Here, $\Xi_{\gamma_0,j}$ denotes the jth row of Ξ_{γ_0}.

Assumption 2 considers the covariates to be sub-Gaussian random vectors, including the Gaussian distribution as a special case. Assumptions 3-5 are standard for the lasso inference. Assumption 3 is made for the sparsity of the first-stage regression, which guarantees that the initial estimator $\tilde{\gamma}$ satisfies $|\tilde{\gamma}-\gamma_0|_1 \leq Cs_1\lambda_1$ and $|\tilde{\gamma}-\gamma_0|_2 \leq Cs_1^{1/2}\lambda_1$, with high probability for some $C>0$. Assumption 4 implies the loss function $-\mathbb{E}_n[\rho_\gamma(X_i,Z_i)]$ in (3.1) behaves quadratically near its minimum. This condition is commonly set in the study of high-dimensional GLMs (see Lemma 6.8 in Bühlmann and Van de Geer (2011)). Assumption 5 considers the inverse of the FI matrix to be strongly sparse, such that the number of nonzero elements in each row is bounded by s_3. The debiased estimator $\hat{\gamma}$ in (4.3) also works under a weaker condition (Assumption S5 in the SM) via the ℓ_q norm on Ξ_{γ_0} for $q \in [0,1)$, which allows several small nonzero entries.

It can be shown that the first-stage lasso estimator $\tilde{\gamma}$ in (3.1) and its debiased estimator $\hat{\gamma}$ in (4.3) have the following properties.

Lemma 1. Suppose that Assumptions 2-5 hold. For the lasso parameters $\lambda_1 \asymp \{\log(p)/n\}^{1/2}$ and $\lambda_{1,j} \asymp \log(p)/\sqrt{n}$, we have $|\tilde{\gamma}-\gamma_0|_1 = O_p(s_1\lambda_1)$, $|\tilde{\gamma}-\gamma_0|_2 = O_p(\sqrt{s_1}\lambda_1)$, $\mathbb{E}_n\big[X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\big]^2 = O_p(s_1\lambda_1^2)$; and

$$\sqrt{n}\big(\hat{\gamma}-\gamma_0\big) = \Xi_{\gamma_0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\rho}_{\gamma_0}(X_i,Z_i) + O_p\big\{(s_3+s_1)\log^{3/2}(p)/\sqrt{n}\big\}. \tag{5.2}$$

Lemma 1 is used in (4.9) to develop the debiased estimator $\hat{\theta}$ and the inference procedure for θ_0. A similar result holds under the weakly sparse condition via ℓ_q norm on Ξ_{γ_0} at the price of a slower convergence rate for the smaller order term in (5.2); see Lemma S1 in the SM for details. This will also cause a slower convergence rate for the smaller order term in the expansion of $\hat{\theta}$ in (5.5). Next, we show the consistency of the restricted lasso estimator $\tilde{\theta}$ in (3.5) to θ_0.

Assumption 6. There exist constants $h_1>0$, $r_0>0$ and $C>0$ such that the distribution of the response variable Y_i satisfies (i). $\mathbb{E}\exp\big[t\{Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)\}\big] \leq C$ for any $|t| \leq h_1$; and (ii). $\lambda_{\min}\big[\mathbb{E}_{D(1)>D(0)}(W_iW_i^{\mathrm{T}})\big] > h_1$, and $1/C \leq \max_{1\leq i\leq n}\ddot{b}(W_i^{\mathrm{T}}\theta) \leq C$ for $\theta \in \mathbb{B}_{r_0}(\theta_0)$.

Part (i) of Assumption 6 assumes a sub-exponential distribution on the centered response $Y_i - \dot{b}(W_i^{\mathrm{T}}\theta_0)$. Owing to the boundedness of the weights $\kappa_i(\gamma_0)$ under Assumption 4, this condition indicates the existence of the moment generating function of $\kappa_i(\gamma_0)\{Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)\}$ around zero for all $i=1,\ldots,n$. Note that $\mathbb{E}\big\{\kappa_i(\gamma_0)\frac{\partial}{\partial\theta_j}\log f(Y_i|X_i,D_i;\theta_0)\big\} = \mathbb{E}_{D(1)>D(0)}\big\{\frac{\partial}{\partial\theta_j}\log f(Y_i|X_i,D_i;\theta_0)\big\} = 0$. By

the large deviation results established in the SM, the deviation of $|\nabla \mathcal{L}_{n,\gamma_0}(\theta_0)|_\infty$ from 0 is bounded at the order $\{\log(p_0)/n\}^{1/2}$. Note that the moment generation function of Y_i on compliers is $\mathbb{E}\big\{\exp(t_1 Y_i)\big\} = \exp\big[\{b(W_i^{\mathrm{T}}\theta_0 + t_1\phi) - b(W_i^{\mathrm{T}}\theta_0)\}/\phi\big]$ under the exponential dispersion family (2.4). For $Y_i - \dot{b}(W_i^{\mathrm{T}}\theta_0)$, we have

$$\begin{aligned}\mathbb{E}\exp\big\{t_1 Y_i - t_1\dot{b}(W_i^{\mathrm{T}}\theta_0)\big\} &= \exp\big\{t_1\dot{b}(W_i^{\mathrm{T}}\theta_0 + t_2\phi) - t_1\dot{b}(W_i^{\mathrm{T}}\theta_0)\big\} \\ &= \exp\big\{t_1 t_2\phi\ddot{b}(W_i^{\mathrm{T}}\theta_0 + t_3\phi)\big\},\end{aligned}$$

where t_2, t_3 are between 0 and t_1. Part (i) of Assumption 6 is automatically satisfied for compliers under (2.4) if $\{\ddot{b}(W_i^{\mathrm{T}}\theta_0)\}$ are bounded.

Part (ii) of Assumption 6 assumes the minimum eigenvalue of $\mathbb{E}_{D(1)>D(0)}(W_i W_i^{\mathrm{T}})$ is larger than a positive constant, and $\ddot{b}(W_i^{\mathrm{T}}\theta)$ is bounded away from 0 and ∞ for all $i = 1, \ldots, n$. Note that $\mathbb{V}\mathrm{ar}_{D(1)>D(0)}(Y_i|D_i, X_i; \theta) = \phi\ddot{b}(W_i^{\mathrm{T}}\theta)$. This condition indicates the variance of Y_i is bounded away from 0 and ∞ in the neighborhood of θ_0. It also implies the expected second derivative of the negative log likelihood for compliers is strictly positive definite around θ_0, such that

$$\lambda_{\min}\bigg[-\mathbb{E}_{D(1)>D(0)}\bigg\{\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y_i|X_i, D_i; \theta)\bigg\}\bigg] > h_0 \quad \text{for} \quad \theta \in \mathbb{B}_{r_0}(\theta_0) \tag{5.3}$$

for a positive constant $h_0 = h_1/(C\phi)$. The above inequality (5.3) is used to guarantee the local restricted strong convexity (Loh, 2017) of the objection function $\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta)$. The detailed derivations are provided in the SM.

Theorem 1. Under Assumptions 1-6, $s_2\log(p_0)/\sqrt{n} = o(1)$, $\lambda_2 \asymp \{s_1\log(p_0)/n\}^{1/2}$, $|\mathcal{A}| < c$ and $|\theta_0|_1 \le R \le cs_2$ for some positive constant c, there exists a stationary point $\tilde{\theta}$ of the optimization program (3.5) such that $|\tilde{\theta} - \theta_0|_2 \le r_0$. Furthermore, any stationary point $\tilde{\theta}$ within this r_0-ball of θ_0 satisfies

$$|\tilde{\theta} - \theta_0|_1 \le Cs_2\lambda_2 \quad \text{and} \quad |\tilde{\theta} - \theta_0|_2 \le C\sqrt{s_2}\lambda_2 \tag{5.4}$$

for some positive constant C, with probability approaching 1 as $n \to \infty$. Additionally, if $\lambda_{\max}\big\{\mathbb{E}_{D(1)>D(0)}(W_i W_i^{\mathrm{T}})\big\} \le 1/h_1$ for a positive constant h_1, $\mathbb{E}_n\big[W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)\big]^2 = O_p(s_2\lambda_2^2)$.

Theorem 1 shows the existence of stationary points of (3.5). It also establishes the consistency of a stationary point to the true parameter θ_0. Compared to the penalty rate $\{\log(p_0)/n\}^{1/2}$ of the conventional convex lasso optimization, the proposed estimator (3.5) requires a larger penalty of the order $\{s_1\log(p_0)/n\}^{1/2}$. This results in slower ℓ_1 and ℓ_2 convergence rates for $\tilde{\theta} - \theta_0$, which are increased by a factor $\sqrt{s_1}$. Such a large penalty is used to control the error when estimating the unknown

weights $\kappa_i(\gamma_0)$ in the empirical objective function $\mathcal{L}_{n,\tilde{\gamma}}(\theta)$. Using the estimated $\tilde{\gamma}$ increases the deviation bound of $|\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)|_\infty$ to the order $\{s_1\log(p_0)/n\}^{1/2}$, even though the maximal gradient $|\nabla\mathcal{L}_{n,\gamma_0}(\theta_0)|_\infty$ with the true γ_0 can be controlled at the rate $\{\log(p_0)/n\}^{1/2}$.

Now, the asymptotic normality for each component of the debiased estimator $\hat{\theta}$ in (4.10) and its linear combination are derived. Note that $\hat{N}_{\theta,\gamma}$ and $\hat{B}_{\theta,\gamma}$ depend on both γ and θ. The following assumptions are made on N_{θ_0,γ_0} and B_{θ_0,γ_0}.

Assumption 7. The influence matrix N_{θ_0,γ_0} is weakly sparse, such that $\|N_{\theta_0,\gamma_0}\|_{\ell_\infty} \leq s_4$ and for $0 \leq q_1 < 1$, $\max_{1\leq j_1\leq p_0}\sum_{j_2=1}^{p}|N_{\theta_0,\gamma_0,j_1j_2}|^{q_1} \leq c_1(p)$.

Assumption 8. The minimum and maximum eigenvalues of $\mathbb{E}_{D(1)>D(0)}(W_iW_i^{\mathrm{T}})$ are bounded from below and above by h_1 and $1/h_1$ for $h_1 > 0$, respectively, and $\|B_{\theta_0,\gamma_0}\|_{\ell_1} \leq s_5$ and $\max_{1\leq j_1\leq p_0}\sum_{j_2=1}^{p_0}|B_{\theta_0,\gamma_0,j_1j_2}|^{q_2} \leq c_2(p)$ for $0 \leq q_2 < 1$.

If the interest is only on the local average treatment effect or the local average treatment effect on the treated, the sparsity conditions on Ξ_{γ_0}, N_{θ_0,γ_0} and B_{θ_0,γ_0} are not needed. But the identification, model and distribution assumptions are still needed, as well as the sparsity assumptions for the instrument propensity score model and the four regressions of Y and D on X under $Z = 0, 1$. See the discussion in Section 2.2 and Belloni et al. (2017a) which treats the covariate effects as nuisance parameters. To conduct statistical inference for each of the regression coefficients in LARF, the sparsity conditions for the Fisher Information matrices are needed. This is similar to the conditions of the de-sparsified lasso in van de Geer et al. (2014).

Let $s_0 = \max_{1\leq k\leq 5}\{s_k\}$. The following theorem provides the asymptotic expansion of the proposed debiased estimator $\hat{\theta}$, in preparation for the variance estimation and the asymptotic normality of $\hat{\theta}$ and $w_c^{\mathrm{T}}\hat{\theta}$.

Theorem 2. Under the conditions of Theorem 1, Assumptions 7, 8, $s_0^{7/2}\log^{5/2}(p_0) = o(\sqrt{n})$, $s_5\lambda_3^{1-q_1}c_1(p)(\log p)^{1/2} = o(1)$ for $\lambda_3 = \sqrt{s_1}\max(\sqrt{s_2}, \log p_0)\{\log(p_0)/n\}^{1/2}$, and $s_0(s_5\lambda_4)^{1-q_2}c_2(p)(\log p)^{1/2} = o(1)$ for $\lambda_4 \asymp s_5\{s_1s_2\log(p_0)/n\}^{1/2}$, we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= \sqrt{n}B_{\theta_0,\gamma_0}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] \\ &- \sqrt{n}B_{\theta_0,\gamma_0}N_{\theta_0,\gamma_0}\Xi_{\gamma_0}\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big] + o_p(1). \qquad (5.5)\end{aligned}$$

Compared to the asymptotic expansion of $\hat{\gamma}$ in Lemma 1, Theorem 2 for $\hat{\theta}$ requires more restrictive conditions on the sparsities of θ_0, γ_0, N_{θ_0,γ_0} and B_{θ_0,γ_0}. This is due to the nature of two-stage estimation, where the first-stage estimator influences the second-stage estimation, causing additional variation. It is further caused by the rates of $\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty$ and $\|\hat{N}_{\tilde{\theta},\tilde{\gamma}} - N_{\theta_0,\gamma_0}\|_\infty$ being higher

than $\{\log(p_0)/n\}^{1/2}$, which is the conventional deviation rate of sample covariances from their population counterparts by large deviation results. In fact, the derivation of Theorem 2 shows that $\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty = O_p\big[\{s_1 s_2 \log(p_0)/n\}^{1/2}\big]$ and $\|\hat{N}_{\tilde{\theta},\tilde{\gamma}} - N_{\theta_0,\gamma_0}\|_\infty = O_p\big[\sqrt{s_1}\max(\sqrt{s_2}, \log p_0)\{\log(p_0)/n\}^{1/2}\big]$, where the extra factors s_1 and s_2 are caused by replacing the true parameters γ_0 and θ_0 by their lasso estimates $\tilde{\gamma}$ and $\tilde{\theta}$.

From the expansion result of $\hat{\theta}$ in Theorem 2, it can be observed that $V = B_{\theta_0,\gamma_0} V_0 B_{\theta_0,\gamma_0}^{\mathrm{T}}$ is the variance of the leading order term in (5.5). The following theorems show the asymptotic normality of $\sqrt{n}(\hat{\theta}_j - \theta_{0,j})\hat{V}_{jj}^{-1/2}$ and $\sqrt{n}w_c^{\mathrm{T}}(\hat{\theta} - \theta_0)(w_c^{\mathrm{T}}\hat{V}w_c)^{-1/2}$ for a constant vector w_c.

Theorem 3. Under the conditions of Theorem 2, if $s_0^{13/2}\log^{5/2}(p_0) = o(\sqrt{n})$, $s_0^3\lambda_3^{1-q_1}c_1(p) = o(1)$, $\lambda_{\max}(N_{\theta_0,\gamma_0}N_{\theta_0,\gamma_0}^{\mathrm{T}}) < C$ and $V_{jj} > 1/C$ for a positive constant C, then,

$$\sqrt{n}(\hat{\theta}_j - \theta_{0,j})\hat{V}_{jj}^{-1/2} \to N(0,1) \text{ in distribution as } n \to \infty.$$

Theorem 4. Under the conditions of Theorem 2, if $|w_c|_1^2 s_0^{13/2}\log^{5/2}(p_0)/\sqrt{n} = o(1)$, $|w_c|_1^2 s_0^3\lambda_3^{1-q_1}c_1(p) = o(1)$, $|w_c|_1^2 s_0(s_5\lambda_4)^{1-q_2}c_2(p) = o(1)$, $\lambda_{\max}(N_{\theta_0,\gamma_0}N_{\theta_0,\gamma_0}^{\mathrm{T}}) < C$, $|w_c|_2 < C$, and $w_c^{\mathrm{T}}Vw_c > 1/C$ for a positive constant C, then,

$$\sqrt{n}w_c^{\mathrm{T}}(\hat{\theta} - \theta_0)(w_c^{\mathrm{T}}\hat{V}w_c)^{-1/2} \to N(0,1) \text{ in distribution as } n \to \infty.$$

Theorems 3 and 4 justify the confidence intervals of θ_0 in (4.13) and $w_c^{\mathrm{T}}\theta_0$ in (4.14). The ℓ_1 norm constraint on w_c is commonly set to estimate sparse linear combinations of high-dimensional regression coefficients (Cai and Guo, 2017). In the case of regression analysis on centered covariates, w_c measures the differences to the average values of covariates. The following corollary extends the result of Theorem 4 to LCSTE for the nonlinear models in Section 2. The confidence intervals (4.15) and (4.16) for $h(w_c^{\mathrm{T}}\theta_0)$ and $h(w_1^{\mathrm{T}}\theta_0) - h(w_0^{\mathrm{T}}\theta_0)$ are built based on this corollary.

Corollary 1. Under the conditions of Theorem 4, if w_0 and w_1 satisfy the same conditions as w_c in Theorem 4, $h(\cdot)$ is a smooth function with $|\dot{h}(\cdot)| > 1/C$, and $\max\{|w_c^{\mathrm{T}}\theta_0|, |w_0^{\mathrm{T}}\theta_0|, |w_1^{\mathrm{T}}\theta_0|\} < C$ for a positive constant C, we have

$$\sqrt{n}\{h(w_c^{\mathrm{T}}\hat{\theta}) - h(w_c^{\mathrm{T}}\theta_0)\}\{\dot{h}(w_c^{\mathrm{T}}\tilde{\theta})^2 w_c^{\mathrm{T}}\hat{V}w_c\}^{-1/2} \to N(0,1) \text{ and}$$
$$\sqrt{n}\big[\{h(w_1^{\mathrm{T}}\hat{\theta}) - h(w_0^{\mathrm{T}}\hat{\theta})\} - \{h(w_1^{\mathrm{T}}\theta_0) - h(w_0^{\mathrm{T}}\theta_0)\}\big](w_a^{\mathrm{T}}\hat{V}w_a)^{-1/2} \to N(0,1)$$

in distribution as $n \to \infty$, where $w_a = \dot{h}(w_1^{\mathrm{T}}\tilde{\theta})w_1 - \dot{h}(w_0^{\mathrm{T}}\tilde{\theta})w_0$.

6. Simulation

This section presents evaluations of the root mean square error (RMSE) of the proposed estimator and the empirical coverage of the confidence interval for each component of the regression coefficients in the LARF of (2.2). The linear combinations of those regression coefficients and the interaction effects between the treatment variable and the covariates were also considered. Further, the proposed method was compared with the double debiased machine learning (DDML) approach (Belloni et al., 2017a; Chernozhukov et al., 2018) and the un-weighted de-sparsified lasso (uwlasso) approach (van de Geer et al., 2014), which does not use the weights $\{\kappa_i(\tilde{\gamma})\}$ in the sample objective function in (3.4). Note that DDML is designed to estimate LATE, which cannot estimate the regression coefficients in the LARF and the heterogeneous treatment effect LCSTE(X).

The simulation design was constructed as follows. First, p-dimensional covariates $\{X_1, \ldots, X_n\}$ were generated according to the normal distribution with mean 0 and covariance $\Sigma = (\sigma_{j_1 j_2})_{p\times p}$, where $\sigma_{j_1 j_2} = \sigma_0 0.5^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, p$. Second, the instrument variable $\{Z_i\}$ was simulated via logistic regression, where $\mathbb{P}(Z_i = 1|X_i) = \Lambda(X_i^{\mathrm{T}}\gamma_1)$ for $\gamma_1 = (\gamma_{1,1}, \ldots, \gamma_{1,p})^{\mathrm{T}}$ and $i = 1, \ldots, n$. Third, for the treatment variable D_i, the potential treatments $\{D_i(0), D_i(1)\}$ were generated as

$$D_i(0) = \mathbb{I}(X_i^{\mathrm{T}}\gamma_2 - \varsigma + \epsilon_{d0,i} \geq 0) \quad \text{and} \quad D_i^*(1) = \mathbb{I}(X_i^{\mathrm{T}}\gamma_2 + \varsigma + \epsilon_{d1,i} \geq 0),$$

respectively, where $\gamma_2 = (\gamma_{2,1}, \ldots, \gamma_{2,p})^{\mathrm{T}}$, $\varsigma = 1$, and $\{\epsilon_{d0,i}, \epsilon_{d1,i}\} \overset{i.i.d.}{\sim} N(0,1)$. Let $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$. To guarantee the monotonicity assumption $D_i(0) \leq D_i(1)$, $D_i(1) = \max\{D_i^*(1), D_i(0)\}$ was set for $i = 1, \ldots, n$. The strength of the instrument can be measured by the estimated slope from the regression of D on Z, which is between 0.66 and 0.77 (standard errors 0.032–0.049) for all the scenarios. Finally, two settings were considered for the response variable Y, creating different types of endogeneity between Y and D on the always-takers ($D_i(0) = 1$) and the never-takers ($D_i(1) = 0$).

- DGP1: for each $i = 1, \ldots, n$, generate $\epsilon_{y,i} \sim N(0,1)$ correlated with $\epsilon_{d0,i}, \epsilon_{d1,i}$, where $\mathbb{C}\mathrm{ov}(\epsilon_{y,i}, \epsilon_{d0,i}) = \mathbb{C}\mathrm{ov}(\epsilon_{y,i}, \epsilon_{d1,i}) = \rho_\epsilon$. Also generate $\epsilon_{0,i} \sim N(0,1)$ independent of $(\epsilon_{y,i}, \epsilon_{d0,i}, \epsilon_{d1,i})^{\mathrm{T}}$. Let $\epsilon_i = \epsilon_{y,i}$ if $D_i(0) = D_i(1)$, and $\epsilon_i = \epsilon_{0,i}$ if $D_i(0) < D_i(1)$. The second-stage regression is

$$Y_i = \alpha_0 D_i + X_i^{\mathrm{T}}\beta_0 + \epsilon_i, \tag{6.1}$$

  where $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^{\mathrm{T}}$. Take $\sigma_0 = 0.5$, $\gamma_{1,j} = \gamma_{2,j} = 0.5\mathbb{I}(j \leq 5)$, $\alpha_0 = 1$ and $\beta_{0,j} = \mathbb{I}(j \leq 5)$. The correlation ρ_ϵ is set to be $0.2, 0.3, 0.4$.

- DGP2: for each $i = 1, \ldots, n$, generate $\epsilon_{0,i} \sim N(0,1)$ independent of $\epsilon_{d0,i}, \epsilon_{d1,i}$. Generate potential outcomes using linear models

$$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} = \begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} + \begin{pmatrix} \xi_{3i} \\ \xi_{3i} \end{pmatrix} \mathbb{I}\{D_i(0) = 1\} + \begin{pmatrix} \xi_{4i} \\ \xi_{4i} \end{pmatrix} \mathbb{I}\{D_i(1) = 0\}, \tag{6.2}$$

  where $\xi_{1i} = \alpha_0 + X_i^{\mathrm{T}}\beta_0 + \epsilon_{0,i}$, $\xi_{2i} = X_i^{\mathrm{T}}\beta_0 + \epsilon_{0,i}$, $\xi_{3i} \sim \text{Poisson}(\lambda_{11})$, and $\xi_{4i} \sim \text{Poisson}(\lambda_{00})$. Let $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$. Take $\sigma_0 = 1$, $\gamma_{1,j} = \gamma_{2,j} = j^{-2}$, $\beta_{0,j} = 2j^{-2}$, $\alpha_0 = 1$, $\lambda_{00} = 0.5$, and $\lambda_{11} = 0.5, 0.75, 1, 1.25$.

The first setting creates dependence between D_i and the error ϵ_i for non-compliers in the second stage regression (6.1). The simultaneity of the two random variables is measured by the correlation ρ_ϵ. The treatment is exogenous on compliers for any $\rho_\epsilon \in (0, 1)$. In the second setting, the intercepts of the regression models differ among the subgroups of compliers and non-compliers. The endogeneity originates from the uncontrolled subgroup indicator, which confounds the treatment variable. The two designs also have different sparsity structures, where the strong and weak sparsity of β_0 are constructed under DGP1 and DGP2, respectively. The coefficients are allowed to take small but nonzero values under DGP2. The interaction effects δ_0 between D_i and X_i were also considered, where $D_i X_i^{\mathrm{T}}\delta_0$ was added to the regression model (6.1) for $\delta_{0,j} = \mathbb{I}(j \leq 5)$ and to the term ξ_{1i} in (6.2) for $\delta_{0,j} = 2j^{-2}$, respectively. The penalty parameters λ_1 and λ_2 in the two stage regressions (3.1) and (3.5) were chosen by five-fold cross-validation in the simulation, r_u was set to be 10 so that the step size of the algorithm (3.8) was 0.1, and the tuning parameters q_0 and λ_4 were set such that 10% of the elements in $\check{N}_{q_0}$ and $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ were nonzero. Following Loh (2017), we set $R = 2(\alpha_0 + |\beta_0|_1)$. Sensitivity analysis was conducted. It shows that the proposed procedure is robust to the tuning parameters q_0, λ_4 and R. For uwlasso, the lasso parameter for the regression of the response on the treatment and covariates was chosen by five-fold cross-validation. The inverse of the design matrix was estimated by CLIME with the same penalty parameter as the proposed method. Further, we set $n = 200, 400$ and $p = 50, 100, 200, 300, 400, 500$. Note that the effective sample size on compliers was significantly lower than the overall sample size. All the simulations were repeated 1000 times.

Figures 1 and 2 report the empirical coverages of the confidence intervals for the regression coefficients α_0 and β_0 and the RMSEs of the estimated coefficients by the proposed method $\hat{\theta}$ in (4.10), the DDML and uwlasso methods for DGP1 and DGP2, respectively. The nominal level was set as 95%. The uwlasso method was simply applied to the linear regression of Y_i on D_i and X_i, ignoring the endogeneity in the data generation process. Here, the coverage and RMSEs are reported for three

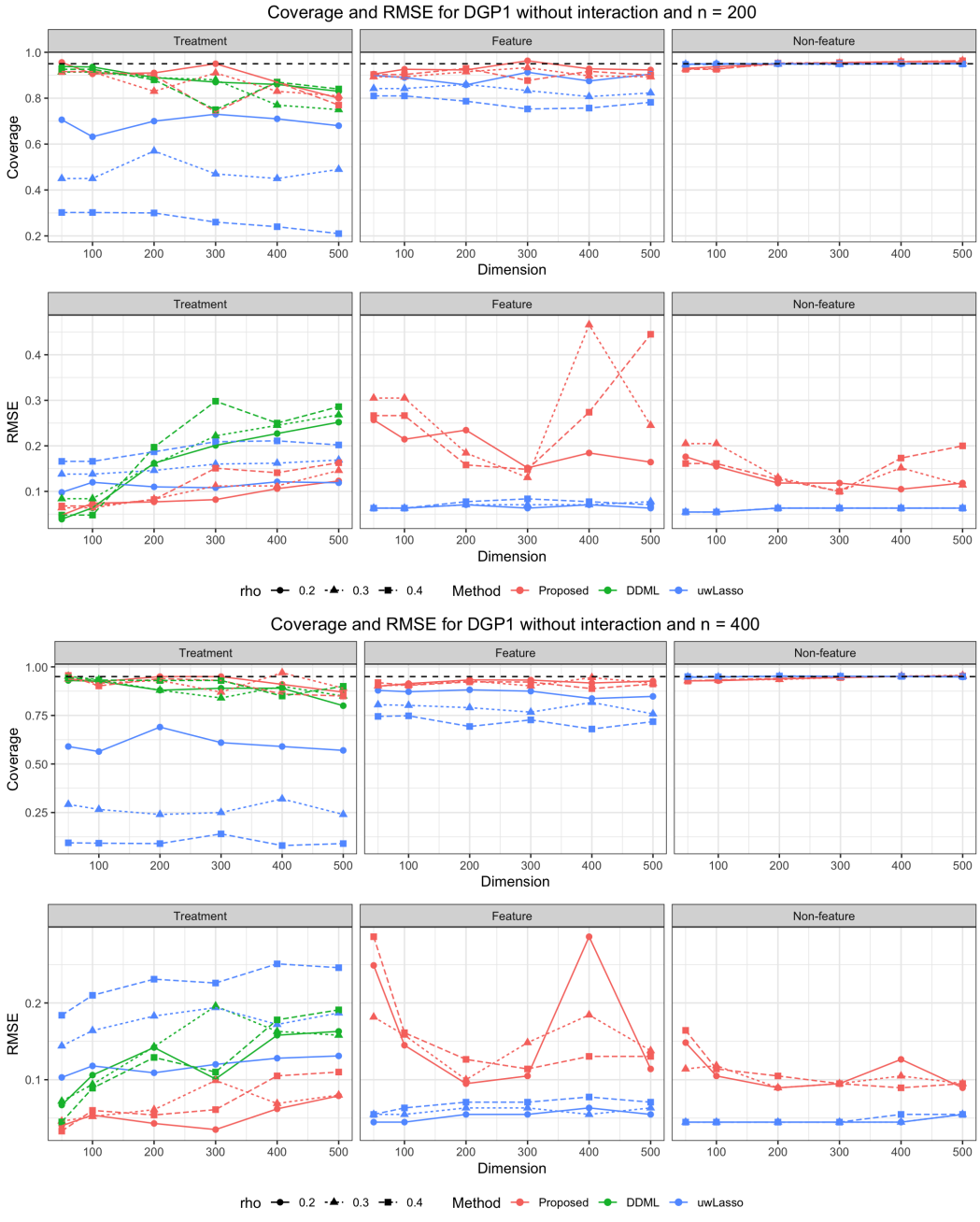

Fig. 1. Empirical coverages of confidence intervals and RMSEs of estimated coefficients for the proposed method, DDML and uwlasso under DGP1 without interaction, $n = 200, 400$ and $\rho_\epsilon = 0.2, 0.3, 0.4$.

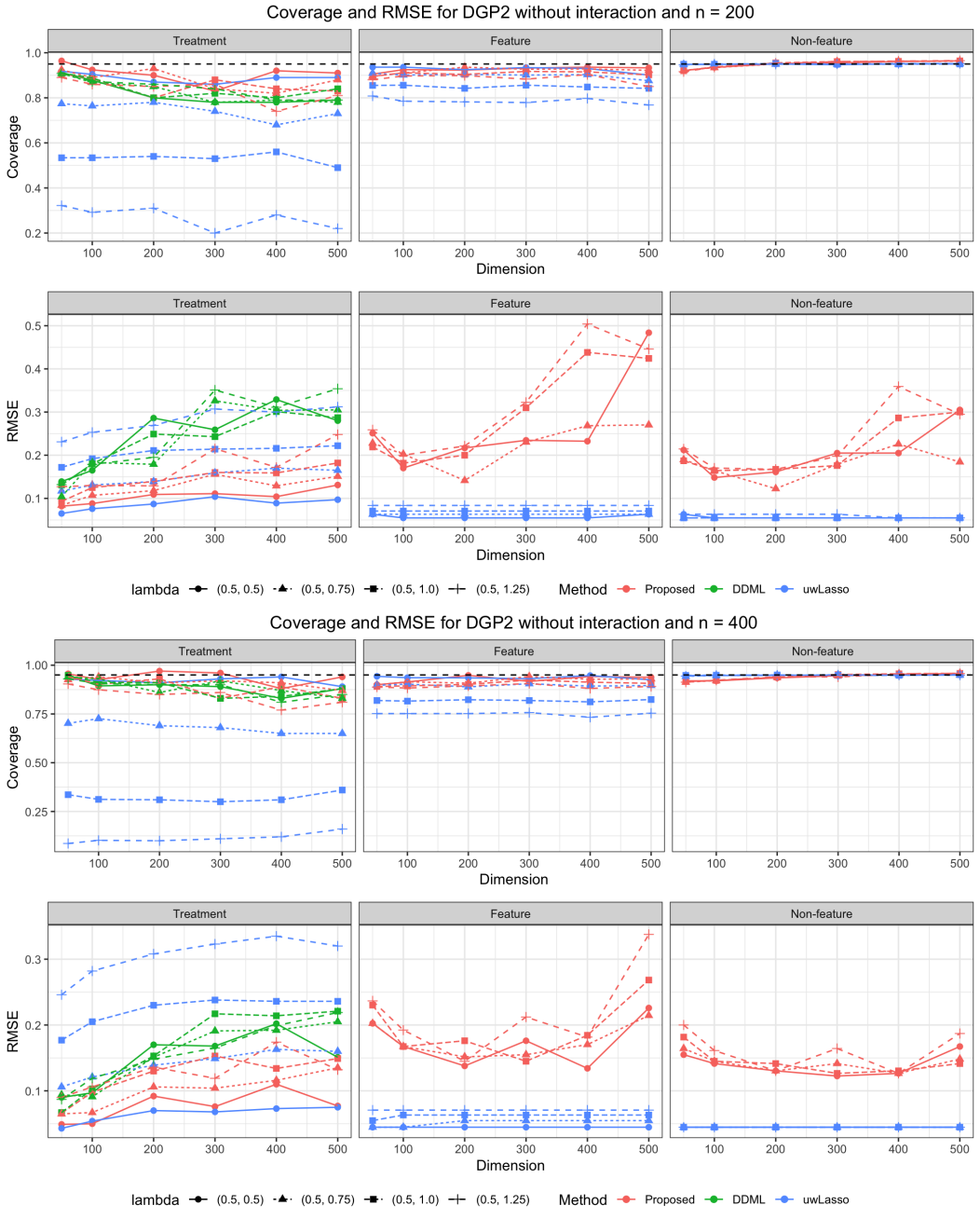

Fig. 2. Empirical coverages of confidence intervals and the RMSEs of estimated coefficients for the proposed method, DDML and uwlasso under DGP2 without interaction, $n = 200, 400$ and $(\lambda_{00}, \lambda_{11}) = (0.5, 0.5), (0.5, 0.75), (0.5, 1), (0.5, 1.25)$.

categories: treatment, covariates with non-zero $\beta_{0,j}$ for DGP1 and $|\beta_{0,j}| > 0.02$ for DGP2 (feature) and covariates with zero $\beta_{0,j}$ for DGP1 and $|\beta_{0,j}| \leq 0.02$ for DGP2 (non-feature). For the latter two categories, the average coverages and RMSEs are calculated. Note that DDML can only be applied to estimate the average treatment effect, which is equal to α_0 under DGP1 and DGP2.

From Figures 1 and 2, it can be seen that the proposed method shows reasonably good performance overall. The method demonstrates coverages close to the nominal level 0.95 for the treatment, features and non-features coefficients in all the given scenarios. The coverage rate departs slightly from the nominal level as the dimension p increases at $n = 200$. As n increases to 400, the empirical coverage improves. Compared to the DDML results, both methods demonstrate similar coverage for the treatment effects at various levels of endogeneity represented by ρ_ϵ. Nevertheless, the RMSE of the proposed method is smaller, as DDML requires the estimation of more parameters compared to the proposed method, which could increase the variation. It should be noted that DDML relies on four fitted regression models of Y and D on X given $Z = 0$ and 1, respectively. On the other hand, the proposed method only needs to estimate one local average response function, as modeled in (2.2). The debiased lasso (uwlasso) approach cannot produce confidence intervals with the desired coverage rate; additionally, it has a large RMSEs for the treatment effect. This is because the lasso without weighting by $\{\kappa_i(\tilde{\gamma})\}$ ignores the endogeneity in the data structure. It is worth mentioning that the uwlasso method yields reasonable coverage for the regression coefficients of features and non-features, as the covariates are independent of the regression error ϵ_i under both DGP1 and DGP2 without interactions. The RMSEs of uwlasso is smaller than that of the proposed method for feature and non-feature coefficients. This is because the propensity score of $\mathbb{P}(Z = 1|X)$ is estimated, and the additional variation of $\tilde{\gamma}$ from the estimated propensity score is included in the proposed method.

Figures S1 and S2 in the SM report the RMSEs of the estimates and the empirical coverages of the proposed approach for the treatment effects and the coefficients of covariates under DGP1 and DGP2 with interactions between D_i and X_i. They also show the relevant results of DDML and uwlasso. The conclusions are similar to those drawn from Figures 1 and 2 without interactions. Figure S3 in the SM provides the results of the linear combination $\alpha_0 + \sum_{j=1}^{10}(10-j)\beta_{0,j}/10$ of the coefficients under DGP1 and DGP2 without interaction. It is observed that the proposed method also achieves accurate coverage in this case, confirming the theoretical results obtained herein. From the above results, it can be concluded that the proposed method performs well for estimating the coefficients in LARF and the LCSTE under the

complier model (2.2). In particular, it demonstrates consistent performance under two data generation processes with different types of endogeneity.

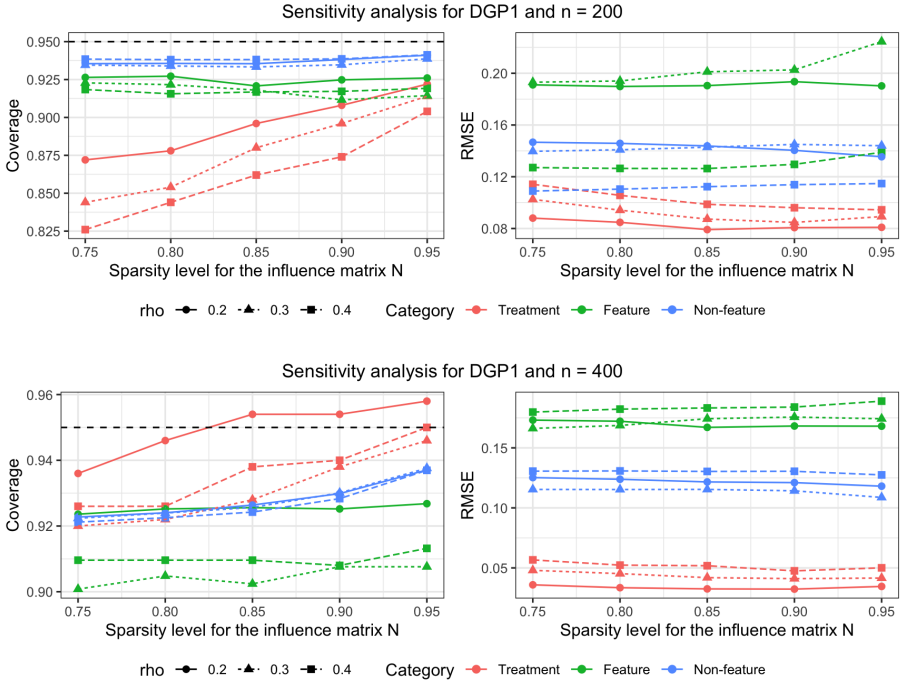

Fig. 3. Empirical coverages of confidence intervals and RMSEs of estimated coefficients for the proposed method with respect to the sparsity of $\check{N}_{q_0}$ (the proportion of zeros in the horizontal axis) under DGP1 without interaction, $p = 100$, and $\rho_\epsilon = 0.2, 0.3, 0.4$.

Figure 3 presents the sensitivity analysis of the estimated influence matrix $\check{N}_{q_0}$ in (4.4) to the proposed method. The empirical coverages and RMSEs of the proposed method are reported for five sparsity levels of $\check{N}_{q_0}$ that correspond to different values of the tuning parameter q_0. It is observed that the coverages and RMSEs are insensitive to changes in the sparsity level in general. Under $n = 200$, the coverage for the treatment effect is slightly lower than the nominal level when $\check{N}_{q_0}$ is less sparse, particularly for $\rho_\epsilon = 0.4$. However, as n increases to 400, the coverages move closer to 0.95 for all sparsity levels. As in $\check{N}_{q_0}$, the impacts of the sparsity of $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ (controlled by λ_4) and the ℓ_1 norm bound R on the proposed procedure are reported in Figure S4 in the SM. This figure shows that the proposed method is robust to the choice of λ_4 and R. These results demonstrate the proposed estimation and inference procedure for the regression coefficients in LARF and the LCSTE are robust to the selection of the tuning parameters.

7. Case Study

In 2008, the state of Oregon conducted eight lottery drawings to randomly select names for its Medicaid program from a waiting list of almost 90,000 uninsured, low-income adults. This created a rare opportunity to study the effects of Medicaid coverage for the uninsured on different health outcomes. One can use this data to examine the effects of expanding public health insurance on healthcare use, financial hardship, health, and labor market outcomes. This section discusses the application of the proposed method to analyze a real data example from the OHIE.

To be eligible for the OHIE, individuals had to be Oregon residents, be otherwise ineligible for Medicaid or other public insurance, have an income below the federal poverty level, have no insurance for at least six months, and be between 19-64 years of age. Among the randomly selected individuals, only those who met the eligibility criteria and completed the application process were enrolled. The treatment variable D and the instrumental variable Z are the 0-1 indicators of Medicaid coverage and lottery selection, respectively; $Z = 1$ if an individual was selected by the lottery, and $D = 1$ if she or he decided to complete the application process and eventually enroll in the Medicaid program after selection. Individuals who did not win the lottery had no chance of Medicaid access, and those who won the lottery could decide not to be enrolled in Medicaid by not completing the application process.

Approximately two years after the experiment, researchers obtained interview data from 5842 adults who were not selected ($Z = 0$) and 6387 adults who were selected for the program ($Z = 1$) (https://www.nber.org/oregon/). The strength of IV can be represented by the slope coefficient 0.240 (standard error 0.008) from the regression of D on Z. This dataset is particularly advantageous for collecting a large set of the demographic characteristics of each individual, including age, gender, education, income, and location. Statistical summary of some key variables of interest are reported in Table 1. For example, about 25% of the respondents were Hispanic or Black, and a little more than half of the participants were female. The average number of emergence department (ED) visits was higher for the treated group. The dependent variable Out-of-Pocket Spending (%) was measured by the percentage of out-of-pocket expenditure to household income. The treated group had less out-of-pocket spending and was less likely to have any catastrophic expenditures or medical debt on average. Happiness was measured on a scale from 1 (unhappiest) to 3 (happiest). Mental and Physical Health Composite Scores were denoted as MCS and PCS, respectively. They were computed using the scores of twelve questions and ranged from zero to one hundred, where one hundred indicated

Table 1. *Some descriptive statistics of the survey respondents in OHIE.*

	Treatment				Control			
	Mean	Std.	Min	Max	Mean	Std.	Min	Max
Age	40.476	11.686	20	64	40.953	11.702	19	71
Female	0.644	0.479	0	1	0.530	0.499	0	1
Hispanic	0.171	0.377	0	1	0.183	0.386	0	1
White	0.684	0.464	0	1	0.687	0.464	0	1
Black	0.120	0.324	0	1	0.095	0.294	0	1
Other	0.136	0.343	0	1	0.147	0.354	0	1
Education (years)	12.170	2.094	9	16	12.351	2.126	9	16
Any doctor visits	0.790	0.408	0	1	0.613	0.486	0	1
Num. of ED visits	1.520	3.791	0	104	0.878	2.118	0	60
Amount of out-of-pocket spending ($)	410.07	1615.436	0	50400	704.912	2356.459	0	92475
Any Catastrophic expenditures	0.039	0.194	0	1	0.046	0.211	0	1
Any medical debt	0.511	0.500	0	1	0.557	0.497	0	1
Happiness	1.921	0.671	1	3	1.961	0.660	1	3
MCS in women $\geq$ 50 yr	40.052	12.221	11.466	66.965	42.889	11.649	13.594	66.404
PCS in women $\geq$ 50 yr	39.087	10.640	12.586	65.352	42.089	10.887	10.269	64.027

the highest level of health while a zero score represented the lowest level of health condition. The control group had higher MCS and PCS compared to the treatment group on average.

Prior analyses of this dataset estimated the LATE of medical insurance on various clinical outcomes (Finkelstein et al., 2012; Baicker et al., 2013, 2014; Finkelstein et al., 2016). They found that compared to the control group without insurance, the treatment group with insurance coverage had substantively higher healthcare utilization, lower out-of-pocket medical expenditures, and better self-reported physical and mental health. However, these studies focused on the LATE and did not explore local heterogeneous treatment effects. Moreover, previous studies only controlled a small set of covariates. We believe that controlling detailed medical histories, personal characteristics, and functional forms of these covariates could lead to more accurate estimates of the effects.

To analyze the data, we consider the following linear model

$$E\left[Y|X_1, X_2, D, D(1) > D(0)\right] = \alpha D + \beta_1^{\mathrm{T}} X_1 + \beta_2^{\mathrm{T}} D X_1 + \beta_3^{\mathrm{T}} X_2,$$

where Y represents different outcomes of interest, the covariates X_1 consists of age, age^2, and race (the baseline group is white and others), and X_2 consists of the second-order polynomials of a large set of covariates that include medical history and demographic variables such as income, family size, education level, location, catastrophic expenditures, and existing borrowed or skipped bills. For the responses Out-of-Pocket Spending and Happiness, the total number of covariates is 40, which

generate 860 polynomial terms. For the dependent variables MCS and PCS, 43 covariates are controlled, which generate 988 polynomial terms excluding the constant. We penalized $\beta_1, \beta_2, \beta_3$ in the estimation (3.5).

Following the strategy in Baicker et al. (2013), results are reported for subgroups divided by age, because the effects are expected to be stronger in subgroups. Three age groups are defined as: below 35, 35 to 49, and 50 to 64 (the oldest eligible group). The sample size for each group is reported in Table 2, excluding observations with missing values. The first two age groups are comparable in size, and about a quarter of the respondents are in the most senior group. For certain clinical outcomes such as MCS and PCS, we focus on 1614 female respondents who were at least 50 years of age. Although the sample sizes are larger than the dimension p in the case study, the number of free parameters in the influence matrix and Fisher Information matrices are at the order of p^2 which is much larger than the sample sizes.

The lasso penalty parameters are chosen by five-fold cross-validation, and all other tuning parameters are kept the same as those in the simulation. The treatment effects of Medicaid coverage on various outcomes of interest are reported in Table 2 (standard errors in parentheses), and $\hat{s}_2$ represents the number of significant covariates selected from the second stage regression by the Benjamini-Hochberg multiple testing procedure on the p-values of each coefficient from the proposed debiased estimator (4.10) at 10% nominal level for FDR. From Table 2, it can be seen that Medicaid coverage led to a significant reduction in financial strain due to medical costs for the younger groups (age$<$35) and senior groups (age$\geq$50) in terms of percentage of out-of-pocket spending. However, the effect was not significant to people whose age were between 35 and 49. This result on each group is consistent with the estimated overall effect of Medicaid coverage on out-of-pocket spending in Baicker et al. (2013). Notably, we also find that the reduction increased with the increase in age, and the effect was more significant for the younger group. For the self-reported levels of happiness, previous research using the same data failed to find a significant overall effect of Medicaid coverage, which was arguably a measure of overall subjective well-being (Baicker et al., 2013). However, as seen in Table 2, the estimation of the heterogeneous effect shows that although Medicaid coverage did not appear to increase the self-reported happiness for younger groups, it significantly increased the happiness levels of individuals above 50 years of age. Our results also show that Medicaid coverage did not significantly increase the overall mental and physical health scores measured by MCS and PCS for elderly women. The results also indicate that age and race had no statistically significant effect on those scores in elderly females. These findings are new to the literature.

Table 2. *Heterogeneous treatment effects of Medicaid coverage on outcomes.*

	Out-of-Pocket Spending (%)			MCS
	age<35	35≤age≤49	age≥ 50	age≥50 & F
D	-0.498 (0.212)	0.246 (0.425)	-1.194 (0.204)	0.206 (0.297)
$female$	-1.932 (0.561)	-1.378 (0.422)	-1.200 (0.785)	–
age	-0.724 (0.335)	-2.789 (5.717)	-1.034 (1.273)	0.116 (0.143)
$Hispanic$	0.072 (0.215)	-0.970 (1.826)	-0.250 (0.560)	0.550 (0.764)
$Black$	0.134 (0.257)	0.535 (0.794)	-1.980 (0.861)	0.113 (0.145)
$D \times age$	-0.385 (0.587)	-0.223 (0.120)	2.596 (0.138)	0.508 (0.417)
$D \times age^2$	-0.280 (0.907)	0.053 (0.107)	-0.239 (0.206)	0.323(0.400)
$D \times Hispanic$	-0.830 (0.881)	0.208 (0.212)	-0.049 (0.234)	0.006 (0.095)
$D \times Black$	-0.830 (0.881)	-0.295 (0.210)	-0.285 (0.191)	0.122 (0.281)
$\hat{s}_2$	66	58	83	38
n	3971	4537	3374	1614
	Happiness			PCS
	age<35	35≤age≤49	age≥ 50	age≥50 & F
D	-0.086 (0.376)	-0.024 (0.056)	0.129 (0.041)	-0.043 (0.384)
$female$	-0.290 (0.614)	-0.275 (0.327)	-0.823 (1.216)	–
age	-0.224 (0.370)	-3.675 (4.434)	-1.244 (1.392)	0.071 (0.255)
$Hispanic$	-0.320 (0.258)	-0.436 (0.657)	1.016 (1.137)	0.033 (0.213)
$Black$	0.050 (0.357)	-0.436 (0.648)	-0.147 (0.568)	-0.094 (0.165)
$D \times age$	0.171 (0.880)	0.022 (0.039)	-0.025 (0.040)	0.234 (0.355)
$D \times age^2$	0.113 (0.083)	0.267 (0.041)	-0.040 (0.056)	0.280 (0.383)
$D \times Hispanic$	0.267 (0.764)	0.164 (0.145)	-0.017 (0.059)	-0.033 (0.111)
$D \times Black$	-0.072 (0.386)	-0.052 (0.084)	-0.053 (0.022)	0.333 (0.278)
$\hat{s}_2$	26	47	69	126
n	3971	4537	3374	1614

8. Discussion

This study proposes novel estimation and inference procedures for heterogeneous treatment effects using observational data with a binary instrumental variable and high-dimensional covariates. The initial estimator for regression coefficients in LARF is obtained by the restricted lasso program (3.5) that provides a consistent estimator under the non-convex objective function $\mathcal{L}_{n,\tilde{\gamma}}(\theta)$. Other methods could be applied to obtain the initial estimator. For the linear model and the program (3.6), following the idea of CoCoLasso (Datta et al., 2017) for error-in-variables regressions, we could first project $\hat{M}_{\tilde{\gamma}}$ to the space of positive semi-definite matrices, and use this projected matrix to replace $\hat{M}_{\tilde{\gamma}}$ in (3.6), which results in a convex lasso program to estimate θ_0. Another alternative is the Dantzig selector (Candes and Tao, 2007), which solves the constrained convex optimization

$$\tilde{\theta} = \arg\min |\theta|_1 \quad \text{such that} \quad |\nabla \mathcal{L}_{n,\tilde{\gamma}}(\theta)|_\infty < \lambda. \tag{8.1}$$

The advantage of the Dantzig estimator is that only one tuning parameter λ is needed, which could be chosen by cross-validation. The ℓ_1 norm bound R on θ in the restricted lasso program (3.5) is not required. The consistency of $\tilde{\theta}$ in (8.1) can be verified under the deviation inequality on $|\nabla \mathcal{L}_{n,\tilde{\gamma}}(\theta)|_\infty$ and the local restricted strong convexity inequality in (S.1.1) in the SM. The inference procedure can be similarly constructed based on the Dantzig estimator by following the steps in Section 4.

It is also interesting to explore doubly robust estimation approaches under high-dimensional covariates. Let $h_c(X;\theta)$ be the model for the local covariate-specific treatment effect LCSTE(X) in (2.1). Under Assumption 1, it is shown in Abadie (2003); Ogburn et al. (2015) that $\mathbb{E}[\kappa_{\text{diff}}\{Y - Dh_c(X;\theta)\}|X] = 0$ for $\kappa_{\text{diff}} = \kappa^{(1)} - \kappa^{(0)}$, where $\kappa^{(0)}$ and $\kappa^{(1)}$ are given after (2.5). Let $h_y(X, z; \psi_{yz})$ and $h_d(X, z; \psi_{dz})$ be the models for $\mathbb{E}(Y|X, Z = z)$ and $\mathbb{E}(D|X, Z = z)$ over the entire population for $z = 0, 1$, respectively, where ψ_{y1}, ψ_{y0}, ψ_{d1} and ψ_{d0} are unknown parameters. Inspired by the findings of Ogburn et al. (2015), the estimation equations

$$\mathbb{E}\big(\nabla_\theta h_c(X;\theta)\kappa_{\text{diff}}[Y - k_y(X;\psi_{y1},\psi_{y0}) - \{D - k_d(X;\psi_{d1},\psi_{d0})\}h_c(X;\theta)]\big) = 0 \tag{8.2}$$

are considered for the estimation of θ, where

$$\begin{aligned} k_y(X;\psi_{y1},\psi_{y0}) &= \{1 - \Lambda(X^{\mathrm{T}}\gamma)\}h_y(X,1;\psi_{y1}) + \Lambda(X^{\mathrm{T}}\gamma)h_y(X,0;\psi_{y0}), \\ k_d(X;\psi_{d1},\psi_{d0}) &= \{1 - \Lambda(X^{\mathrm{T}}\gamma)\}h_d(X,1;\psi_{d1}) + \Lambda(X^{\mathrm{T}}\gamma)h_d(X,0;\psi_{d0}), \end{aligned}$$

and $\Lambda(X^{\mathrm{T}}\gamma)$ is the model for $\mathbb{P}(Z = 1|X)$. Given that the model $h_c(X;\theta)$ is accurately specified for LCSTE(X), the estimation equation in (8.2) is doubly-robust in

the sense that it is valid if either the instrument propensity score model $\Lambda(X^{\mathrm{T}}\gamma)$ or the regression models $h_y(X, z; \psi_{yz})$ and $h_d(X, z; \psi_{dz})$ (not necessarily both) are specified correctly. This is shown in detail in the SM.

Let $\hat{\psi}_{y1}$, $\hat{\psi}_{y0}$, $\hat{\psi}_{d1}$ and $\hat{\psi}_{d0}$ be the regularized estimators for ψ_{y1}, ψ_{y0}, ψ_{d1} and ψ_{d0}, respectively. The regularized estimation for $\mathbb{E}(Y|X, Z = z)$ and $\mathbb{E}(D|X, Z = z)$ was considered in Belloni et al. (2017a) for estimating LATE under high-dimensional covariates. The doubly-robust estimation for θ and LCSTE(X) can be obtained using the Dantzig selector similar to (8.1), which solves the optimization

$$\tilde{\theta} = \arg\min |\theta|_1 \quad \text{such that}$$

$$\left|\mathbb{E}_n\big(\nabla_\theta h_c(X_i;\theta)\hat{\kappa}_{\mathrm{diff}}[Y_i - k_y(X_i;\hat{\psi}_{y1},\hat{\psi}_{y0}) - \{D_i - k_d(X_i;\hat{\psi}_{d1},\hat{\psi}_{d0})\}h_c(X_i;\theta)]\big)\right|_\infty < \lambda.$$

Following the method described in Section 4, the inference procedure for θ can be developed based on $\tilde{\theta}$ and the de-sparsified estimators for γ, ψ_{yz} and ψ_{dz}. Unlike the method in Ogburn et al. (2015), the approach proposed herein can be applied to data with high-dimensional covariates. Importantly, it avoids fitting the regression of $Y - Dh_c(X;\theta)$ on X and Z for every value of θ, which would create difficulties in high-dimensional estimation. It should also be noted that the weight $\kappa^{(1)} - \kappa^{(0)}$ has more negative values than the weight κ in (2.5) used in the proposed procedure, because $\mathbb{E}\{\kappa^{(1)} - \kappa^{(0)}\} = 0$, but $\mathbb{E}(\kappa) = \mathbb{P}\{D(1) > D(0)\} > 0$. The negative weights may create numerical instability as well as computational challenges for the Dantzig estimator. These issues will be investigated further in future works.

The proposed approach can also be extended to multiple-level and continuous instrument variables. For a multiple-level instrument variable Z, consider its two levels z and z', and the complier group $\{D(z') > D(z)\}$. Let $\Pi_{z,z'} = \{Z = z \text{ or } z'\}$. Similar to the identification equality (2.6) under Assumption 1, if the independent instrument and monotonicity conditions (Tan, 2006) are satisfied for each pair of levels of Z, it can be shown that

$$\mathbb{E}\{g(Y, D, X)|D(z') > D(z), \Pi_{z,z'}\} = \mathbb{E}\{\kappa_{z,z'}g(Y, D, X)\}/\mathbb{P}(D(z') > D(z), \Pi_{z,z'})$$

for any measurable function $g(\cdot)$ of (Y, D, X), where $\kappa_{z,z'} = \mathbb{I}(\Pi_{z,z'}) - \frac{D\mathbb{I}(Z=z)}{\mathbb{P}(Z=z|X)} - \frac{(1-D)\mathbb{I}(Z=z')}{\mathbb{P}(Z=z'|X)}$. Estimation and inference procedures similar to the proposed approach can be established for LARF and LCSTE(X). This can be achieved based on the above identification equality, the model for $\mathbb{P}(Z = z|X)$, and the models of Y on X and D on each complier group $\{D(z') > D(z)\}$. A study of such models can be found in Tan (2006). Additionally, Imbens and Angrist (1994) provided an identification and two-stage estimation approach for LATE under a model with a multiple-level instrument variable but no covariates.

For causal inference with a continuous instrument Z, Kennedy et al. (2019) considered the extended monotonicity condition that $D(z) = \mathbb{I}(z \geq T)$ for an unobserved random threshold T, and proposed to estimate the local instrumental variable (LIV) curve, defined as $\mathbb{E}\{Y(1) - Y(0)|X, T = t\}$, under a fixed-dimensional setting. By modeling the LIV curve, the conditional expectations $\mathbb{E}(Y|X, Z = z)$ and $\mathbb{E}(D|X, Z = z)$ and the conditional distribution of Z given X, Kennedy et al. (2019) estimated the parameters in the LIV model by solving a system of estimating equations. Although the target parameters and models differ from those proposed herein, this is a two-stage estimation approach similar to the proposed methods and the aforementioned extensions. Using the Dantzig selector (8.1) on the estimating equations, this approach can be extended to the case of high-dimensional covariates. A full investigation of this problem will be conducted in future studies.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263.

Angrist, J. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Baicker, K., Finkelstein, A., Song, J., and Taubman, S. (2014). The impact of medicaid on labor market activity and program participation: evidence from the oregon health insurance experiment. *American Economic Review*, 104(5):322–328.

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., and Finkelstein, A. N. (2013). The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017a). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

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

Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2017b). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956.

Bonetti, M. and Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481.

Bühlmann, P. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

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

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies*, 86(2):1095–1122.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.

Datta, A., Zou, H., et al. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106.

Finkelstein, A. N., Taubman, S. L., Allen, H. L., Wright, B. J., and Baicker, K. (2016). Effect of medicaid coverage on ed use—further evidence from oregon's experiment. *New England Journal of Medicine*, 375(16):1505–1507.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734.

Hong, H. and Nekipelov, D. (2010). Semiparametric efficiency in nonlinear late models. *Quantitative Economics*, 1(2):279–304.

Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Janková, J. and van de Geer, S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv:1610.01353*.

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

Kennedy, E. H., Lorch, S., and Small, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143.

Lee, S., Okui, R., and Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32(7):1207–1225.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Ma, Y. and Zhou, X.-H. (2017). Treatment selection in a randomized clinical trial via covariate-specific treatment effect curves. *Statistical Methods in Medical Research*, 26(1):124–141.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:21–25.

Ogburn, E. L., Rotnitzky, A., and Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396.

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

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: A Review Journal of the Institute of Mathematical Statistics*, 29(4):640.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Tian, L. and Tibshirani, R. (2010). Adaptive index models for marker-based risk stratification. *Biostatistics*, 12(1):68–86.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Zhang, C.-H. and Zhang, S. (2014). 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.

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

Supplementary Material to "Inference of Heterogeneous Treatment Effects Using Observational Data with High-Dimensional Covariates"

Yumou Qiu, Jing Tao and Xiao-Hua Zhou

Iowa State University, University of Washington and Peking University

Recall that $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ is the data matrix of the covariates, where $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. Let $\mathcal{X}_j = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}}$ be the jth column of $\mathbf{X}$. Let $\mathcal{X}_{-j}$ be the $n \times (p-1)$ sub-matrix of $\mathbf{X}$ without the jth variable. Let $X_{i,-j}$ be the ith observation without the jth variable. Recall that $W_i = (D_i, X_i^{\mathrm{T}}, D_i X_i^{\mathrm{T}})^{\mathrm{T}}$ and $W_i = (D_i, X_i^{\mathrm{T}})^{\mathrm{T}}$ for models with and without interactions between D_i and X_i, respectively, and p_0 is the dimensionality of W_i and θ_0 in the LARF. Recall that C is a positive constant which may change from case to case.

1 Proof of Theorem 1

Recall that $z_{\theta,i} = -\frac{\partial}{\partial\theta} g(Y_i, D_i, X_i; \theta)$. Note from (**??**), the jth coordinate of the ith summation term in $\nabla \mathcal{L}_{n,\gamma_0}(\theta_0)$ is $-\kappa_i(\gamma_0)\big\{Y_i - \dot{b}(W_i^{\mathrm{T}}\theta_0)\big\}W_{ij}/\phi = -\kappa_i(\gamma_0) z_{\theta_0,ij}$. Since

$$\mathbb{E}\big\{\kappa_i(\gamma_0)\nabla \log f(Y_i|X_i, D_i; \theta_0)\big\} = \mathbb{E}_{D_1 > D_0}\big\{\nabla \log f(Y_i|X_i, D_i; \theta_0)\big\} = 0,$$

$\mathbb{E}\{\kappa_i(\gamma_0) z_{\theta_0,ij}\} = 0$ for all $j = 1, \ldots, p_0$.

Following Theorem 1 of Loh (2017), we first consider $|\nabla \mathcal{L}_{n,\tilde{\gamma}}(\theta_0)|_\infty$. Let $\tilde{Y}_i = \kappa_i(\gamma_0)\{Y_i - \dot{b}(W_i^{\mathrm{T}}\theta_0)\}$, which is sub-Exponential distributed with zero mean under Assumption 6. We will show $\max_{1 \le j \le p_0} |\mathbb{E}_n(\tilde{Y}_i W_{ij})| \le C\{\log(p_0)/n\}^{1/2}$ for a positive constant C.

Let $\check{Y}_i = \tilde{Y}_i \mathbb{I}\{|\tilde{Y}_i| \le C_0 \log(p_0)\}$. For any positive constant c_0, since $\check{Y}_i$ is sub-Exponential distributed, $\mathbb{P}\{\max_{1\le i\le n} |\tilde{Y}_i| > C_0 \log(p_0)\} \le n p_0^{-\tilde{c}_0} \le p_0^{-c_0}$ by choosing C_0 large enough. For any $1 \le j \le p_0$, we have

$$\mathbb{P}\big(|\mathbb{E}_n(\tilde{Y}_i W_{ij})| > C\{\log(p_0)/n\}^{1/2}\big) \le \mathbb{P}\big(|\mathbb{E}_n(\check{Y}_i W_{ij})| > C\{\log(p_0)/n\}^{1/2}\big) + p_0^{-c_0},$$

where by Chernoff bound,

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{E}_n(\check{Y}_i W_{ij}) - \mathbb{E}(\check{Y}_i W_{ij}) > C\{\log(p_0)/n\}^{1/2}\big) \\
&\qquad\qquad \le \ \exp(-Ct\{n\log(p_0)\}^{1/2}) \prod_{i=1}^{n} \mathbb{E}\exp\{t(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))\}.
\end{aligned}$$

By a Taylor expansion of the function $\exp(x)$ at 0, $\mathbb{E}\exp\{t(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))\} = 1 + t\mathbb{E}(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij})) + \mathbb{E}\big\{\sum_{k=2}^{\infty} t^k (\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))^k / k!\big\}$, which is bounded by $1 + \mathbb{E}\big[t^2(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))^2 \exp\{t|\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij})|\}\big]$. For any positive t at the order $\{\log(p_0)/n\}^{1/2}$, we have

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{E}_n(\check{Y}_i W_{ij}) - \mathbb{E}(\check{Y}_i W_{ij}) > C\{\log(p_0)/n\}^{1/2}\big) \\
\le\ &\exp(-Ct\{n\log(p_0)\}^{1/2}) \prod_{i=1}^{n} \Big(1 + \mathbb{E}\big[t^2(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))^2 \exp\{t|\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij})|\}\big]\Big) \\
\le\ &\exp\Big(-Ct\{n\log(p_0)\}^{1/2} + 2\sum_{i=1}^{n} \mathbb{E}\big[t^2(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))^2 \exp\{t|\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij})|\}\big]\Big),
\end{aligned}$$

where the last inequality is from the Taylor expansion of $\log(1+a)$ at 1 for small a. Since $\mathbb{E}(\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij}))^2 \le \mathbb{E}(\check{Y}_i W_{ij})^2 \le \mathbb{E}^{1/2} Y_i^4 \mathbb{E}^{1/2} W_{ij}^4 < C_1 < \infty$, and $\mathbb{E}\exp\{t|\check{Y}_i W_{ij} - \mathbb{E}(\check{Y}_i W_{ij})|\} \le \mathbb{E}\exp\{tC_0\log(p_0)(|W_{ij}| + \mathbb{E}|W_{ij}|)\} \le C_1$ by the fact $\log^3(p_0)/n \to 0$ and W_{ij} is from a sub-Gaussian distribution. Therefore,

$$\mathbb{P}\big(\mathbb{E}_n(\check{Y}_i W_{ij}) - \mathbb{E}(\check{Y}_i W_{ij}) > C\{\log(p_0)/n\}^{1/2}\big) \le \exp\big\{-CC_2^{1/2}\log(p_0) + 2C_2C_1^{3/2}\log(p_0)\big\}$$

for $t = \{C_2\log(p_0)/n\}^{1/2}$ for a positive constant C_2. It can be similarly shown that

$$\mathbb{P}\big(\mathbb{E}_n(\check{Y}_i W_{ij}) - \mathbb{E}(\check{Y}_i W_{ij}) < -C\{\log(p_0)/n\}^{1/2}\big) \le \exp\big\{-CC_2^{1/2}\log(p_0) + 2C_2C_1^{3/2}\log(p_0)\big\}.$$

From the above probability inequalities, by choosing C large enough, it follows that

$$\mathbb{P}\big(\max_{1\le j\le p_0}|\mathbb{E}_n(\check{Y}_i W_{ij})-\mathbb{E}(\check{Y}_i W_{ij})|>C\{\log(p_0)/n\}^{1/2}\big)\le p_0^{-c_0}$$

for a positive constant c_0. Since $\mathbb{E}(\check{Y}_i W_{ij})=-\mathbb{E}\big[W_{ij}\tilde{Y}_i\mathbb{I}\{|\tilde{Y}_i|>C_0\log(p_0)\}\big]$, by Cauchy Schwarz inequality, $|\mathbb{E}(\check{Y}_i W_{ij})|\le C\mathbb{E}^{1/2}\big[\tilde{Y}_i^2\mathbb{I}\{|\tilde{Y}_i|>C_0\log(p_0)\}\big]\le Cp^{-c_0}$ for all $j=1,\ldots,p_0$. Those results imply that for a positive constant C, $\max_{1\le j\le p_0}|\mathbb{E}_n(\tilde{Y}_i W_{ij})|\le C\{\log(p_0)/n\}^{1/2}$ and $|\nabla\mathcal{L}_{n,\gamma_0}(\theta_0)|_\infty\le C\{\log(p_0)/n\}^{1/2}$ with probability converging to 1 as $n,p\to\infty$.

It suffices to focus on $\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)-\mathcal{L}_{n,\gamma_0}(\theta_0)$. Notice that

$$\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)-\nabla\mathcal{L}_{n,\gamma_0}(\theta_0) \;=\; -\frac{1}{n}\sum_{i=1}^n\{\kappa_i(\tilde{\gamma})-\kappa_i(\gamma_0)\}\nabla\big\{\log f(Y_i|X_i,D_i;\theta_0)\big\}.$$

By Cauchy Schwarz inequality and $|\kappa_i(\tilde{\gamma})-\kappa_i(\gamma_0)|\le CX_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)$ under Assumption 4,

$$|\nabla_j\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)-\nabla_j\mathcal{L}_{n,\gamma_0}(\theta_0)| \;\le\; C\mathbb{E}_n^{1/2}\big[X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\big]^2\mathbb{E}_n^{1/2}\big[\nabla_j\log f(Y_i|X_i,D_i;\theta_0)\big]^2$$

for all $j=1,\ldots,p$, where $\nabla_j\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)$, $\nabla_j\mathcal{L}_{n,\gamma_0}(\theta_0)$ and $\nabla_j\log f(Y_i|X_i,D_i;\theta_0)$ denote the jth component of $\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)$, $\nabla\mathcal{L}_{n,\gamma_0}(\theta_0)$ and $\nabla\log f(Y_i|X_i,D_i;\theta_0)$, respectively.

Notice that $z_{\theta_0,ij}=\nabla_j\log f(Y_i|X_i,D_i;\theta_0)=\big\{Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)\big\}W_{ij}/\phi$. Let $\check{z}_{\theta_0,ij}=z_{\theta_0,ij}\mathbb{I}\{|Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)|\le C_0\log(p_0)\}$ for a large positive constant C_0. Since $z_{\theta_0,ij}^2=(\check{z}_{\theta_0,ij}^2-\mathbb{E}\check{z}_{\theta_0,ij}^2)+\mathbb{E}(\check{z}_{\theta_0,ij})^2$ on the set $\{|Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)|\le C_0\log(p_0)\}$, and $\mathbb{E}(\check{z}_{\theta_0,ij}^2)\le\mathbb{E}(z_{\theta_0,ij}^2)\le C_1$ for a positive constant C_1, we have, for any $j=1,\ldots,p_0$,

$$\begin{aligned}\mathbb{P}(\mathbb{E}_n z_{\theta_0,ij}^2>2C_1) &\le \mathbb{P}\big\{\mathbb{E}_n(\check{z}_{\theta_0,ij}^2-\mathbb{E}\check{z}_{\theta_0,ij}^2)>C_1\big\}+\sum_{i=1}^n\mathbb{P}\big\{|Y_i-\dot{b}(W_i^{\mathrm{T}}\theta_0)|>C_0\log(p_0)\big\}\\ &\le \mathbb{P}\big\{\mathbb{E}_n(\check{z}_{\theta_0,ij}^2-\mathbb{E}\check{z}_{\theta_0,ij}^2)>C_1\big\}+p_0^{-c_0}\end{aligned}$$

for a positive constant c_0 related to C_0, and all $j=1,\ldots,p_0$. Following the proof for the deviation bound $\max_{1\le j\le p_0}\big\{\mathbb{E}_n(\check{Y}_i W_{ij})-\mathbb{E}(\check{Y}_i W_{ij})\big\}\le C\{\log(p_0)/n\}^{1/2}$, it can be similarly shown that

$$\mathbb{P}\big\{\max_{1\le j\le p_0}\mathbb{E}_n(\check{z}_{\theta_0,ij}^2-\mathbb{E}\check{z}_{\theta_0,ij}^2)>C\{\log(p_0)/n\}^{1/2}\big\}\le p_0^{-c_0}$$

for a large positive constant C, given $\log^5(p_0)/n \to 0$ as $n, p \to \infty$. Those results imply $\max_j \mathbb{E}_n(z^2_{\theta_0,ij}) = O_p(1)$. It follows that $|\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_0) - \nabla\mathcal{L}_{n,\gamma_0}(\theta_0)|_\infty \le Cs_1^{1/2}\lambda_1$ by Lemma 1, and $|\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_0)|_\infty \le C\{s_1\log(p_0)/n\}^{1/2}$ for a positive constant C, with probability converging to 1 as $n, p \to \infty$.

Next, we show the local restricted strong convexity (Loh and Wainwright, 2012) of the objection function $\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta)$ such that

$$\langle\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_1) - \nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_2), \theta_1 - \theta_2\rangle \ge \tau_{c,1}|\theta_1 - \theta_2|_2^2 - \tau_{c,2}|\theta_1 - \theta_2|_1^2\log(p_0)/n \qquad \text{(S.1.1)}$$

for two positive constants $\tau_{c,1}$ and $\tau_{c,2}$, and all $\theta_1, \theta_2 \in \mathbb{B}_{r_0}(\theta_0)$. Note that by Taylor expansion,

$$\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_1) - \nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_2) = -\frac{1}{n}\sum_{i=1}^n \kappa_i(\tilde{\gamma})\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y_i|X_i, D_i; \theta_3)(\theta_1 - \theta_2),$$

where θ_3 is between θ_1 and θ_2, and hence, $\theta_3 \in \mathbb{B}_{r_0}(\theta_0)$. It follows

$$\langle\nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_1) - \nabla\mathcal{L}_{n,\tilde{\gamma}}(\theta_2), \theta_1 - \theta_2\rangle = (\theta_1 - \theta_2)^{\mathrm{T}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})\nabla_i^2(\theta_3)\big](\theta_1 - \theta_2),$$

where $\nabla_i^2(\theta) = -\nabla^2\log f(Y_i|X_i, D_i; \theta) = -\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y_i|X_i, D_i; \theta) = \phi^{-1}\ddot{b}(W_i^{\mathrm{T}}\theta)W_iW_i^{\mathrm{T}}$ given in (**??**).

Since from (5.3), $\lambda_{\min}\big[\mathbb{E}_{D_1>D_0}\big\{-\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y_i|X_i, D_i; \theta)\big\}\big] > h_0$ for $\theta \in \mathbb{B}_{r_0}(\theta_0)$, by Lemma 13 of Loh and Wainwright (2012), to show (S.1.1), it suffices to prove for all $\theta \in \mathbb{B}_{r_0}(\theta_0)$,

$$\big|v^{\mathrm{T}}\{\mathbb{E}_n[\kappa_i(\tilde{\gamma})\nabla_i^2(\theta)] - \mathbb{E}[\kappa_i(\gamma_0)\nabla_i^2(\theta)]\}v\big| \le c_1h_0 \ \text{ for any } \ v \in \mathbb{K}(2\tau_c^{-1}) \qquad \text{(S.1.2)}$$

with probability converging to 1, where $\tau_c = c_2\log(p_0)/n$ for some proper positive constants c_1 and c_2, and

$$\mathbb{K}(2\tau_c^{-1}) = \{v \in \mathbb{R}^{p_0} : |v|_0 \le 2\tau_c^{-1} \ \text{ and } \ |v|_2 \le 1\}$$

is the set of the p_0-dimensional vectors within the unit ℓ_2 ball and with at most $2\tau_c^{-1}$ nonzero elements.

Notice that $\mathbb{E}_n[\kappa_i(\gamma)\nabla_i^2(\theta)] = \phi^{-1}\mathbb{E}_n[\kappa_i(\gamma)\ddot{b}(W_i^{\mathrm{T}}\theta)W_iW_i^{\mathrm{T}}]$, to show (S.1.2), it suffices to prove for all $\theta \in \mathbb{B}_{r_0}(\theta_0)$,

$$\big|\mathbb{E}_n\big[\{\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)\}\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big]\big| \;\leq\; c_1h_0\phi/2 \;\text{ and} \tag{S.1.3}$$

$$\big|\mathbb{E}_n\big[\kappa_i(\gamma_0)\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big] - \mathbb{E}\big[\kappa_i(\gamma_0)\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big]\big| \;\leq\; c_1h_0\phi/2 \tag{S.1.4}$$

for any $v \in \mathbb{K}(2\tau_c^{-1})$ with probability converging to 1 as $n, p \to \infty$. For (S.1.3), by Cauchy Schwarz inequality and Assumption 6, we have

$$\big|\mathbb{E}_n\big[\{\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)\}\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big]\big| \leq C\mathbb{E}_n\big[|\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)|(v^{\mathrm{T}}W_i)^2\big].$$

Since $\max_{1\leq i\leq n}|\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)| \leq C\{\log(p)\}^{1/2}|\tilde{\gamma} - \gamma_0|_1 = O_p\{s_1\log(p)n^{-1/2}\} = o_p(1)$ by Lemma 1 and Assumption 3, we only need to show $\max_{v\in\mathbb{K}(2\tau_c^{-1})}\mathbb{E}_n(v^{\mathrm{T}}W_i)^2$ is bounded. Noting that $v^{\mathrm{T}}W_i$ is sub-Gaussian distributed by Assumption 2, it follows $\mathbb{E}(v^{\mathrm{T}}W_i)^2 \leq C$ for any v with $|v|_2 \leq 1$. Note that $\tau_c^{-1}\log(p_0) = c_2^{-1}n$. By Lemma 15 of Loh and Wainwright (2012), for any positive constant t_0,

$$\mathbb{P}\Big\{\max_{v\in\mathbb{K}(2\tau_c^{-1})}\big|\mathbb{E}_n(v^{\mathrm{T}}W_i)^2 - \mathbb{E}(v^{\mathrm{T}}W_i)^2\big| \geq t_0\Big\} \leq C_2\exp\big\{-C_3nt_0 + 2\tau_c^{-1}\log(p_0)\big\}$$

for postive constants C_2, C_3, which converges to 0 at the rate $\exp(-c_3n)$ for $c_3 > 0$ by choosing c_2 large. This implies $\mathbb{E}_n(v^{\mathrm{T}}W_i)^2$ is bounded over $\mathbb{K}(2\tau_c^{-1})$, and

$$\max_{v\in\mathbb{K}(2\tau_c^{-1})}\mathbb{E}_n\big[\{\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)\}\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big] \to 0$$

in probability as $n, p \to \infty$. Hence, (S.1.3) holds with overwhelming probability when n and p are large enough.

For (S.1.4), since $v^{\mathrm{T}}W_i$ is sub-Gaussian distributed for $|v|_2 \leq 1$, and $\kappa_i(\gamma_0)$ and $\ddot{b}(W_i^{\mathrm{T}}\theta)$ are bounded for $\theta \in \mathbb{B}_{r_0}(\theta_0)$ under Assumptions 4 and 6, $\kappa_i(\gamma_0)\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2$ is sub-Exponential distributed for all $v \in \mathbb{K}(2\tau_c^{-1})$. Following the proof of Lemma 15 in Loh and Wainwright (2012), it can be shown that

$$\mathbb{P}\Big\{\max_{v\in\mathbb{K}(2\tau_c^{-1})}\big|\mathbb{E}_n\big[\kappa_i(\gamma_0)\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big] - \mathbb{E}\big[\kappa_i(\gamma_0)\ddot{b}(W_i^{\mathrm{T}}\theta)(v^{\mathrm{T}}W_i)^2\big]\big| \geq t_0\Big\}$$
$$\leq\; C_2\exp\big\{-C_3nt_0 + 2\tau_c^{-1}\log(p_0)\big\}.$$

Following the above argument, (S.1.4) is satisfied with probability converging to 1 by choosing c_2 large enough.

Given (S.1.2) hold, we have for any $\theta \in \mathbb{B}_{r_0}(\theta_0)$,

$$\left|v^{\mathrm{T}}\{\mathbb{E}_n[\kappa_i(\tilde{\gamma})\nabla_i^2(\theta)] - \mathbb{E}[\kappa_i(\gamma_0)\nabla_i^2(\theta)]\}v\right| \leq 27c_1h_0(|v|_2^2 + \tau_c|v|_1^2) \text{ for any } v \in \mathbb{R}^{p_0},$$

which leads to

$$\begin{aligned} v^{\mathrm{T}}\mathbb{E}_n[\kappa_i(\tilde{\gamma})\nabla_i^2(\theta)]v &\geq v^{\mathrm{T}}\mathbb{E}[\kappa_i(\gamma_0)\nabla_i^2(\theta)]v - 27c_1h_0(|v|_2^2 + \tau_c|v|_1^2) \\ &\geq h_0|v|_2^2 - 27c_1h_0(|v|_2^2 + \tau_c|v|_1^2) = \tau_{c,1}|v|_2^2 - \tau_{c,2}|v|_1^2\log(p_0)/n \end{aligned}$$

by choosing $c_1 = 1/54$, where $\tau_{c,1} = h_0/2$ and $\tau_{c,2} = h_0c_2/2$. This proves the local restricted strong convexity inequality (S.1.1) by letting $v = \theta_1 - \theta_2$ for all $\theta_1, \theta_2 \in \mathbb{B}_{r_0}(\theta_0)$. Applying the proof of Theorem 1 in Loh (2017) on the case of bounded $|\mathcal{A}|$, this theorem follows by choosing $\lambda_2 \asymp \{s_1\log(p_0)/n\}^{1/2}$ and $|\theta_0|_1 \leq R \leq cs_2$ for some positive constant c. Also note that

$$v^{\mathrm{T}}\mathbb{E}_n[\kappa_i(\tilde{\gamma})\nabla_i^2(\theta)]v \leq \lambda_{\max}\big\{\mathbb{E}[\kappa_i(\gamma_0)\nabla_i^2(\theta)]\big\}|v|_2^2 + h_0(|v|_2^2 + \tau_c|v|_1^2)/2,$$

where $\nabla_i^2(\theta) = \phi^{-1}\ddot{b}(W_i^{\mathrm{T}}\theta)W_iW_i^{\mathrm{T}}$. Since $\lambda_{\max}\big[\mathbb{E}(W_iW_i^{\mathrm{T}})\big] \leq 1/h_1$, $\lambda_{\max}\big\{\mathbb{E}[\kappa_i(\gamma_0)\nabla_i^2(\theta)]\big\} \leq C_4$ for a positive constant C_4. Letting $v = \tilde{\theta} - \theta_0$, we have $\mathbb{E}_n\big[W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)\big]^2 = O_p(s_2\lambda_2^2)$. □

2 Proof of Theorem 2

By the mean-value theorem on $\kappa_i(\tilde{\gamma})$ and Taylor expansion of $\frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \theta)$ at $\tilde{\theta}$, $\nabla\mathcal{L}_{n,\tilde{\gamma}}(\tilde{\theta})$ can be written as

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \kappa_i(\tilde{\gamma})\frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \theta_0) + \frac{1}{n}\sum_{i=1}^n \kappa_i(\tilde{\gamma})\left\{\frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \tilde{\theta}) - \frac{\partial}{\partial\theta}g(Y_i, D_i, X_i; \theta_0)\right\} \\ = \; &-\frac{1}{n}\sum_{i=1}^n \kappa_i(\gamma_0)z_{\theta_0,i} + \hat{N}_{\theta_0,\tilde{\gamma}}\big(\tilde{\gamma} - \gamma_0\big) + \hat{M}_{\tilde{\theta},\tilde{\gamma}}(\tilde{\theta} - \theta_0) + O_p\{s_1s_2\log^{3/2}(p_0)/n\}, \qquad \text{(S.2.1)} \end{aligned}$$

where $\bar{\gamma}$ is between $\tilde{\gamma}$ and γ_0. Using $\check{N}_{q_0}$ given in (4.4) as an estimator of N_{θ_0,γ_0} for $s \in (0,1]$, it is shown that $|\check{N}_{q_0} - N_{\theta_0,\gamma_0}|_\infty = o_p(1)$, and $\sqrt{n}\big|(\hat{N}_{\theta_0,\bar{\gamma}} - \check{N}_{q_0})(\tilde{\gamma}-\gamma_0)\big|_\infty = o_p(1)$. Together with (S.2.1), this leads to

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i} \;=\; -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\gamma_0)z_{\theta_0,i} + \check{N}_{q_0}\sqrt{n}(\tilde{\gamma}-\gamma_0) + \hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta}-\theta_0) + o_p(1), \tag{S.2.2}$$

given $s_1 \max\{s_2, \log(p_0)\}\log^{3/2}(p_0)/\sqrt{n} = o(1)$ and $s_1^{3/2}\max(\sqrt{s_2}, \log p_0)\log(p_0)/\sqrt{n} = o(1)$. Since the maximal difference between $\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}}$ and I_{p_0} are controlled by λ_4 in (4.6), we have that

$$\begin{aligned}
&\sqrt{n}(\tilde{\theta}-\theta_0) + \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big] \\
&\qquad = \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] - \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\gamma_0) + o_p(1).
\end{aligned} \tag{S.2.3}$$

We provide the following lemma on the consistency of $\hat{M}_{\tilde{\theta},\tilde{\gamma}}$ and $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$.

Lemma 2. Suppose that assumptions in Theorem 2 hold. Set $\lambda_4 \geq \|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty$. Then,

$$\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty \;=\; O_p\big(\{s_1 s_2 \log(p_0)/n\}^{1/2}\big) \tag{S.2.4}$$

$$\|\hat{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty \;=\; O_p\big(4s_5\lambda_4\big) \;\text{ and} \tag{S.2.5}$$

$$\big|\hat{B}_{\tilde{\theta},\tilde{\gamma},j} - B_{\theta_0,\gamma_0,j}\big|_1 \;\leq\; 2c_q c_2(p)(4s_5\lambda_4)^{1-q_2} \tag{S.2.6}$$

for each $j = 1, \ldots, p_0$, where $c_q = 1 + 2^{1-q_2} + 3^{1-q_2}$, and $\hat{B}_{\tilde{\theta},\tilde{\gamma},j}$ and $B_{\theta_0,\gamma_0,j}$ are the jth row of $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ and B_{θ_0,γ_0}, respectively.

Proof of Lemma 2. It suffices to prove the results for the solution $\tilde{B}_{\tilde{\theta},\tilde{\gamma}} = (\tilde{b}_{j_1 j_2})$ of (4.6). Note that the matrix optimization problem (4.6) can be decomposed to p_0 vector minimization problems. Let e_j be a unit vector in $\mathbb{R}^{p_0}$ with 1 in the jth coordinate and 0 in all other coordinates. For $1 \leq j \leq p_0$, the jth row $\tilde{B}_{\tilde{\theta},\tilde{\gamma},j}$ of $\tilde{B}_{\tilde{\theta},\tilde{\gamma}}$ is the solution to the following minimization program:

$$\min_{b_j \in \mathbb{R}^{p_0}} |b_j|_1 \text{ subject to } \|\hat{M}_{\tilde{\theta},\tilde{\gamma}} b_j - e_j\|_\infty \leq \lambda_4. \tag{S.2.7}$$

Part (i): Note that $\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0} = (\hat{M}_{\tilde{\theta},\tilde{\gamma}} - \hat{M}_{\theta_0,\tilde{\gamma}}) + (\hat{M}_{\theta_0,\tilde{\gamma}} - \hat{M}_{\theta_0,\gamma_0}) + (\hat{M}_{\theta_0,\gamma_0} - M_{\theta_0,\gamma_0})$, where

$$\begin{aligned}
\hat{M}_{\tilde{\theta},\tilde{\gamma}} - \hat{M}_{\theta_0,\tilde{\gamma}} &= \phi^{-1}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})\{\ddot{b}(W_i^{\mathrm{T}}\tilde{\theta}) - \ddot{b}(W_i^{\mathrm{T}}\theta_0)\}W_iW_i^{\mathrm{T}}\big], \text{ and} \\
\hat{M}_{\theta_0,\tilde{\gamma}} - \hat{M}_{\theta_0,\gamma_0} &= \phi^{-1}\mathbb{E}_n\big[\{\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)\}\ddot{b}(W_i^{\mathrm{T}}\theta_0)W_iW_i^{\mathrm{T}}\big].
\end{aligned}$$

We have shown $|\ddot{b}(W_i^{\mathrm{T}}\tilde{\theta}) - \ddot{b}(W_i^{\mathrm{T}}\theta_0)| \le C|W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)|$ and $|\kappa_i(\tilde{\gamma}) - \kappa_i(\gamma_0)| \le C|X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)|$ for all $i = 1, \ldots, n$. By Cauchy Schwarz inequality, it follows that $\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - \hat{M}_{\theta_0,\tilde{\gamma}}\|_\infty \le C\sqrt{s_2}\lambda_2 \le C\{s_1 s_2 \log(p_0)/n\}^{1/2}$ and $\|\hat{M}_{\theta_0,\tilde{\gamma}} - \hat{M}_{\theta_0,\gamma_0}\|_\infty \le C\sqrt{s_1}\lambda_1 = C\{s_1 \log(p)/n\}^{1/2}$ for a positive constant C with probability converging to 1 as $n, p \to \infty$. By large deviation results, we also have $\|\hat{M}_{\theta_0,\gamma_0} - M_{\theta_0,\gamma_0}\|_\infty \le C\{\log(p_0)/n\}^{1/2}$. Those results imply that $\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty = O_p\big(\{s_1 s_2 \log(p_0)/n\}^{1/2}\big)$.

Part (ii): We provide the upper bound of $\|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty$. By Assumption 8, $\big\|B_{\theta_0,\gamma_0}\big\|_{\ell_1} \le s_5$. Note that

$$\begin{aligned}
\|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty &\le \|(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0})M_{\theta_0,\gamma_0}\|_\infty \|B_{\theta_0,\gamma_0}\|_{\ell_1} \\
&\le s_5\|(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0})M_{\theta_0,\gamma_0}\|_\infty, \qquad \text{(S.2.8)}
\end{aligned}$$

and $\|(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0})M_{\theta_0,\gamma_0}\|_\infty$ is bounded by

$$\begin{aligned}
&\big\|\big(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\big)\big(\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\big)\big\|_\infty + \big\|\big(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\big)\hat{M}_{\tilde{\theta},\tilde{\gamma}}\big\|_\infty \\
\le\ & \|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty \|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_{\ell_1} + \|\tilde{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I\|_\infty + \|I - B_{\theta_0,\gamma_0}\hat{M}_{\tilde{\theta},\tilde{\gamma}}\|_\infty \\
\le\ & 2\|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty + \|\tilde{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I\|_\infty + \|I - B_{\theta_0,\gamma_0}\hat{M}_{\tilde{\theta},\tilde{\gamma}}\|_\infty,
\end{aligned}$$

where the last inequality follows from the fact that B_{θ_0,γ_0} is a feasible point for Program (S.2.7) so we have $\|\tilde{B}_{\tilde{\theta},\tilde{\gamma}}\|_{\ell_1} \le \|B_{\theta_0,\gamma_0}\|_{\ell_1}$. Set $\lambda_4 \ge \|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty$, we have

$$\|(\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0})M_{\theta_0,\gamma_0}\|_\infty \le 2\|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty + 2\lambda_4 \le 4\lambda_4. \qquad \text{(S.2.9)}$$

Thus, (S.2.8) and (S.2.9) imply that

$$\|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty \le 4s_5\lambda_4. \qquad \text{(S.2.10)}$$

Part (iii): We next provide the upper bound of $\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j} - B_{\theta_0,\gamma_0,j}\big|_1$. Let $t_n = \|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty$. Note that

$$\begin{aligned}
&\sum_{j_2=1}^{p_0} \big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} - B_{\theta_0,\gamma_0,j_1j_2}\big| \leq t_n \sum_{j_2=1}^{p_0} \mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} \\
&\quad + \sum_{j_2=1}^{p_0} \big|B_{\theta_0,\gamma_0,j_1j_2}\big|\big|\mathbb{I}\{|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| \geq 2t_n\} - \mathbb{I}\{|B_{\theta_0,\gamma_0,j_1j_2}| \geq 2t_n\}\big| + (2t_n)^{1-q_2}c_2(p) \\
\leq\ & (2t_n)^{1-q_2}c_2(p) + (t_n)^{1-q_2}c_2(p) + (3t_n)^{1-q_2}c_2(p) \\
\leq\ & \big(1 + 2^{1-q_2} + 3^{1-q_2}\big)t_n^{1-q_2}c_2(p) \qquad \text{(S.2.11)}
\end{aligned}$$

by a similar argument in the proof of Theorem 6 in Cai et al. (2011). By the definition of $\tilde{B}_{\tilde{\theta},\tilde{\gamma}}$, we have $\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j}\big|_1 \leq \big|B_{\theta_0,\gamma_0,j}\big|_1$ for all $j = 1,\ldots,p_0$. Notice that $\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1}\big|_1 = \sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big|$ equals to

$$\begin{aligned}
&\sum_{j_2=1}^{p_0}\Big(\big|B_{\theta_0,\gamma_0,j_1j_2} + \tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} - B_{\theta_0,\gamma_0,j_1j_2}\big| \\
&\qquad\qquad + \big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2} - \tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\}\big|\Big) \\
\geq\ & \big|B_{\theta_0,\gamma_0,j_1}\big|_1 - \sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} - B_{\theta_0,\gamma_0,j_1j_2}\big| \\
&\qquad\qquad + \sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2} - \tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\}\big|,
\end{aligned}$$

which implies that

$$\sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2} - \tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\}\big| \leq \sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} - B_{\theta_0,\gamma_0,j_1j_2}\big|,$$

and hence,

$$\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1} - B_{\theta_0,\gamma_0,j_1}\big|_1 \leq 2\sum_{j_2=1}^{p_0}\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{\big|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\big| \geq 2t_n\big\} - B_{\theta_0,\gamma_0,j_1j_2}\big|. \qquad \text{(S.2.12)}$$

Because of (S.2.10), (S.2.11) and (S.2.12), we have

$$\begin{aligned}
|\tilde{B}_{\tilde{\theta},\tilde{\gamma},j_1} - B_{\theta_0,\gamma_0,j_1}|_1 &\leq 2c_2(p)\big(1 + 2^{1-q_2} + 3^{1-q_2}\big)\|\tilde{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_\infty^{1-q_2} \\
&\leq 2c_2(p)\big(1 + 2^{1-q_2} + 3^{1-q_2}\big)(4s_5\lambda_4)^{1-q_2}
\end{aligned}$$

for all $j_1 = 1, \ldots, p_0$. $\square$

We now verify (S.2.1)–(S.2.3) that serve as the foundation to prove Theorem 2.

Proof of (S.2.1). Notice that $\frac{\partial}{\partial\theta} g(Y_i, D_i, X_i; \theta) = -\nabla \log f(Y_i|X_i, D_i; \theta)$, and

$$\nabla_i^2(\theta) = -\nabla^2 \log f(Y_i|X_i, D_i; \theta) = \phi^{-1}\ddot{b}(W_i^{\mathrm{T}}\theta)W_i W_i^{\mathrm{T}}$$

is the second derivative of $g(Y_i, D_i, X_i; \theta)$ with respect to θ. We can express $M_{\theta,\gamma}$ and $\hat{M}_{\theta,\gamma}$ in (4.7) as $M_{\theta,\gamma} = \mathbb{E}\big\{\kappa_i(\gamma)\nabla_i^2(\theta)\big\} = \phi^{-1}\mathbb{E}\big\{\kappa_i(\gamma)\ddot{b}(W_i^{\mathrm{T}}\theta)W_i W_i^{\mathrm{T}}\big\}$ and $\hat{M}_{\theta,\gamma} = \mathbb{E}_n\big\{\kappa_i(\gamma)\nabla_i^2(\theta)\big\}$, respective.

For any γ, by the first order Taylor expansion of $-\mathbb{E}_n\big[\kappa_i(\gamma)\nabla \log f(Y_i|X_i, D_i; \theta)\big]$ at $\tilde{\theta}$, we have

$$-\mathbb{E}_n\big[\kappa_i(\gamma)\nabla \log f(Y_i|X_i, D_i; \theta)\big] + \mathbb{E}_n\big[\kappa_i(\gamma)\nabla \log f(Y_i|X_i, D_i; \tilde{\theta})\big] = \mathbb{E}_n\big[\kappa_i(\gamma)\nabla_i^2(\bar{\theta})\big](\theta - \tilde{\theta}),$$

where $\bar{\theta}$ is between θ and $\tilde{\theta}$. Therefore,

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \kappa_i(\tilde{\gamma})\bigg\{\frac{\partial}{\partial\theta} g(Y_i, D_i, X_i; \tilde{\theta}) - \frac{\partial}{\partial\theta} g(Y_i, D_i, X_i; \theta_0)\bigg\} \\ = \;& \mathbb{E}_n\big[\kappa_i(\tilde{\gamma})\nabla_i^2(\bar{\theta})\big](\tilde{\theta} - \theta_0) \;=\; \hat{M}_{\tilde{\theta},\tilde{\gamma}}(\tilde{\theta} - \theta_0) + (\hat{M}_{\bar{\theta},\tilde{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\theta} - \theta_0), \end{aligned}$$

where $\bar{\theta}$ is between θ_0 and $\tilde{\theta}$. Let $\mathrm{Rem}_1 = (\hat{M}_{\bar{\theta},\tilde{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\theta} - \theta_0)$. We have

$$\hat{M}_{\bar{\theta},\tilde{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}} = \mathbb{E}_n\big[\kappa_i(\tilde{\gamma})\{\nabla_i^2(\bar{\theta}) - \nabla_i^2(\tilde{\theta})\}\big] = \phi^{-1}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})\{\ddot{b}(W_i^{\mathrm{T}}\bar{\theta}) - \ddot{b}(W_i^{\mathrm{T}}\tilde{\theta})\}W_i W_i^{\mathrm{T}}\big].$$

Since $|\ddot{b}(W_i^{\mathrm{T}}\bar{\theta}) - \ddot{b}(W_i^{\mathrm{T}}\tilde{\theta})| \le C|W_i^{\mathrm{T}}(\bar{\theta} - \tilde{\theta})| \le C|W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)|$ by the Lipschitz continuity of $\ddot{b}(\cdot)$. It follows $|\mathrm{Rem}_1|_\infty \le C\{\log(p_0)\}^{1/2}\mathbb{E}_n\big[W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)\big]^2 \le C\{\log(p_0)\}^{1/2}s_2\lambda_2^2$ by Theorem 1. The expansion (S.2.1) follows by noticing that $\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\theta_0,i}\big] = \mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] + \mathbb{E}_n\big[z_{\theta_0,i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})X_i^{\mathrm{T}}\big](\tilde{\gamma} - \gamma_0)$, where $\bar{\gamma}$ is between $\tilde{\gamma}$ and γ_0. $\square$

Proof of (S.2.2). From (S.2.1), up to an additive small order term $O_p\{s_1 s_2 \log^{3/2}(p_0)/\sqrt{n}\}$, we have

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n \kappa_i(\tilde{\gamma})z_{\tilde{\theta},i} \;=\; -\frac{1}{\sqrt{n}}\sum_{i=1}^n \kappa_i(\gamma_0)z_{\theta_0,i} + \hat{N}_{\theta_0,\tilde{\gamma}}\sqrt{n}(\tilde{\gamma} - \gamma_0) + \hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta} - \theta_0),$$

where $\hat{N}_{\theta_0,\bar{\gamma}} = -\mathbb{E}_n\big[z_{\theta_0,i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})X_i^{\mathrm{T}}\big]$, and $\bar{\gamma}$ is between $\tilde{\gamma}$ and γ_0. To show (S.2.2), it suffices to prove $\sqrt{n}\big|(\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\gamma}-\gamma_0)\big|_\infty = o_p(1)$.

Note that $\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}} = (\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\theta_0,\tilde{\gamma}}) + (\hat{N}_{\theta_0,\tilde{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})$, where

$$\begin{aligned}
\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\theta_0,\tilde{\gamma}} &= \mathbb{E}_n\big[z_{\theta_0,i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}\big] - \mathbb{E}_n\big[z_{\theta_0,i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})X_i^{\mathrm{T}}\big] \text{ and} \\
\hat{N}_{\theta_0,\tilde{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}} &= \mathbb{E}_n\big[z_{\tilde{\theta},i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}\big] - \mathbb{E}_n\big[z_{\theta_0,i}\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}\big] \\
&= \mathbb{E}_n\big[(z_{\tilde{\theta},i} - z_{\theta_0,i})\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}\big].
\end{aligned}$$

By the Lipschitz continuity of $\dot{\kappa}_0(\cdot)$ and $\dot{\Lambda}(\cdot)$, we have

$$|\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})| \le C|X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)|$$

for all $i = 1,\ldots,n$. Based on this inequality, $|(\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\theta_0,\tilde{\gamma}})(\tilde{\gamma}-\gamma_0)|_\infty$ is bounded by

$$\begin{aligned}
&\big|\mathbb{E}_n\big[z_{\theta_0,i}\big\{\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})\big\}X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\big]\big|_\infty \\
\le\ & C\mathbb{E}_n\big[|z_{\theta_0,i}|\{X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\}^2\big] \ \le\ C\log^{3/2}(p_0)s_1\lambda_1^2
\end{aligned}$$

with probability converging to 1. Since $z_{\tilde{\theta},i} - z_{\theta_0,i} = \phi^{-1}\big\{\dot{b}(W_i^{\mathrm{T}}\theta_0) - \dot{b}(W_i^{\mathrm{T}}\tilde{\theta})\big\}W_i$, similarly, $|(\hat{N}_{\theta_0,\tilde{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\gamma}-\gamma_0)|_\infty$ is bounded by

$$\begin{aligned}
&\big|\mathbb{E}_n\big[(z_{\tilde{\theta},i} - z_{\theta_0,i})\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\big]\big|_\infty \\
\le\ & C\big|\mathbb{E}_n\big[|W_i||W_i^{\mathrm{T}}(\tilde{\theta}-\theta_0)||X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)|\big]\big|_\infty \\
\le\ & C\{\log(p_0)\}^{1/2}\mathbb{E}_n^{1/2}\big[W_i^{\mathrm{T}}(\tilde{\theta}-\theta_0)\big]^2\mathbb{E}_n^{1/2}\big[X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\big]^2 \ \le\ C\{s_1s_2\log(p_0)\}^{1/2}\lambda_1\lambda_2
\end{aligned}$$

with probability converging to 1. Therefore, we can write

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n \kappa_i(\tilde{\gamma})z_{\tilde{\theta},i} \ =\ -\frac{1}{\sqrt{n}}\sum_{i=1}^n \kappa_i(\gamma_0)z_{\theta_0,i} + \hat{N}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\gamma}-\gamma_0) + \hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta}-\theta_0) + \mathrm{Rem}_2,$$

where $\mathrm{Rem}_2 = \sqrt{n}\mathrm{Rem}_1 + \sqrt{n}(\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\gamma}-\gamma_0)$, and $\mathrm{Rem}_1 = (\hat{M}_{\bar{\theta},\tilde{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\theta}-\theta_0)$. Through the above derivation, we have

$$|\mathrm{Rem}_2|_\infty = O_p\{s_1s_2\log^{3/2}(p_0)/\sqrt{n}\} + O_p\{s_1\max(\sqrt{s_2},\log p_0)\log^{3/2}(p_0)/\sqrt{n}\}.$$

we can further express (S.2.1) as

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i} \tag{S.2.13}$$
$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i(\gamma_0)z_{\theta_0,i} + \check{N}_{q_0}\sqrt{n}(\tilde{\gamma}-\gamma_0) + \hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta}-\theta_0) + \mathrm{Rem}_3,$$

where $\mathrm{Rem}_3 = \mathrm{Rem}_2 + \sqrt{n}(\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \check{N}_{q_0})(\tilde{\gamma}-\gamma_0)$, and $|\sqrt{n}(\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \check{N}_{q_0})(\tilde{\gamma}-\gamma_0)|_\infty \le \sqrt{n}q_0 s_1\lambda_3\lambda_1 \asymp s_1^{3/2}\max(\sqrt{s_2},\log p_0)\log(p_0)/\sqrt{n}$. Combining with the bound on $|\mathrm{Rem}_2|_\infty$, we have $|\mathrm{Rem}_3|_\infty \le Cs_1\sqrt{s_2}\max(\sqrt{s_1},\sqrt{s_2\log p_0})\log(p_0)/\sqrt{n}$ if $s_2 \ge C\log^2(p)$.

Notice that $\hat{N}_{\tilde{\theta},\tilde{\gamma}} - N_{\theta_0,\gamma_0} = (\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \hat{N}_{\theta_0,\tilde{\gamma}}) + (\hat{N}_{\theta_0,\tilde{\gamma}} - \hat{N}_{\theta_0,\gamma_0}) + (\hat{N}_{\theta_0,\gamma_0} - N_{\theta_0,\gamma_0})$, where $|\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \hat{N}_{\theta_0,\tilde{\gamma}}|_\infty \le C\sqrt{s_2}\lambda_2$, $|\hat{N}_{\theta_0,\tilde{\gamma}} - \hat{N}_{\theta_0,\gamma_0}|_\infty \le C\log(p_0)\sqrt{s_1}\lambda_1$ and $|\hat{N}_{\theta_0,\gamma_0} - N_{\theta_0,\gamma_0}|_\infty \le C\{\log(p_0)/n\}^{1/2}$ with probability converging to 1 as $n,p\to\infty$. Those results imply that $|\hat{N}_{\tilde{\theta},\tilde{\gamma}} - N_{\theta_0,\gamma_0}|_\infty = O_p(\lambda_3)$ for $\lambda_3 = \sqrt{s_1}\max(\sqrt{s_2},\log p_0)\{\log(p_0)/n\}^{1/2}$. Similar as $\check{N}_{q_0,j_1j_2} = \hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}\mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3\big\}$, let $\tilde{N}_{q_0,j_1j_2} = N_{\theta_0,\gamma_0,j_1j_2}\mathbb{I}\big\{|N_{\theta_0,\gamma_0,j_1j_2}| > q_0\lambda_3\big\}$ based on the true influence matrix N_{θ_0,γ_0}. Let $\tilde{N}_{q_0} = (\tilde{N}_{q_0,j_1j_2})$. For the matrix ℓ_∞ norm, under Assumption 7, we have

$$\begin{aligned}
\|\check{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} \;\le\; & \|\check{N}_{q_0} - \tilde{N}_{q_0}\|_{\ell_\infty} + \|\tilde{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} \;\le\; C\lambda_3^{1-q_1}c_1(p) \\
& + \max_{j_1}\sum_{j_2=1}^{p}|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}|\mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| \le q_0\lambda_3\big\} \\
& + \max_{j_1}\sum_{j_2=1}^{p}|N_{\theta_0,\gamma_0,j_1j_2}|\mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| \le q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| > q_0\lambda_3\big\} \\
& + \max_{j_1}\sum_{j_2=1}^{p}|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2} - N_{\theta_0,\gamma_0,j_1j_2}|\mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| > q_0\lambda_3\big\},
\end{aligned}$$

where the last term in the above inequality is bounded by $C\lambda_3^{1-q_1}c_1(p)$. For the second last term, it is bounded by

$$\begin{aligned}
& \max_{j_1}\sum_{j_2=1}^{p}|N_{\theta_0,\gamma_0,j_1j_2} - \hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}|\mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| \le q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| > q_0\lambda_3\big\} \\
+ \;\; & q_0\lambda_3\max_{j_1}\sum_{j_2=1}^{p}\mathbb{I}\big\{|N_{\theta_0,\gamma_0,j_1j_2}| > q_0\lambda_3\big\} \;\le\; C\lambda_3^{1-q_1}c_1(p).
\end{aligned}$$

Since $|\hat{N}_{\tilde{\theta},\tilde{\gamma}} - N_{\theta_0,\gamma_0}|_\infty \le C\lambda_3$ with probability $1 - p^{-c_0}$ for a constant $c_0 > 0$, for a large $q_0 > 0$ and a small $c_1 > 0$, we have

$$\begin{aligned} & \max_{j_1} \sum_{j_2=1}^{p} |\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| \mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| \le q_0\lambda_3\big\} \\ \le & \max_{j_1} \sum_{j_2=1}^{p} |\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2} - N_{\theta_0,\gamma_0,j_1j_2}| \mathbb{I}\big\{|\hat{N}_{\tilde{\theta},\tilde{\gamma},j_1j_2}| > q_0\lambda_3, |N_{\theta_0,\gamma_0,j_1j_2}| > c_1q_0\lambda_3\big\} + C\lambda_3^{1-q_1}c_1(p), \end{aligned}$$

which is bounded by $C\lambda_3^{1-q_1}c_1(p)$. It follows that $\|\check{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} = O_p\{\lambda_3^{1-q_1}c_1(p)\}$, and $\|\check{N}_{q_0}\|_{\ell_\infty} \le s_4$ given $\lambda_3^{1-q_1}c_1(p) \to 0$ as $n, p \to \infty$. □

Proof of (S.2.3). Multiplying $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ on both sides of (S.2.13), we have

$$\begin{aligned} & -\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big] \\ = & -\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] + \hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}\sqrt{n}(\tilde{\gamma} - \gamma_0) + \hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta} - \theta_0) + \hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathrm{Rem}_3, \end{aligned}$$

where $\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}}\sqrt{n}(\tilde{\theta} - \theta_0) = \sqrt{n}(\tilde{\theta} - \theta_0) + (\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I_{p_0})\sqrt{n}(\tilde{\theta} - \theta_0)$. From (4.6),

$$\big|(\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I_{p_0})\sqrt{n}(\tilde{\theta} - \theta_0)\big|_\infty \le \sqrt{n}\|\hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{M}_{\tilde{\theta},\tilde{\gamma}} - I_{p_0}\|_\infty|\tilde{\theta} - \theta_0|_1 = O_p(\sqrt{n}s_2\lambda_2\lambda_4),$$

which converges to 0 if $s_2\{s_1\log(p_0)\}^{1/2}\lambda_4 \to 0$ as $n, p \to \infty$.

From the derivation of (S.2.1) and (S.2.2), we have $\mathrm{Rem}_3 = \mathrm{Rem}_2 + \sqrt{n}(\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \check{N}_{q_0})(\tilde{\gamma} - \gamma_0)$, and $\mathrm{Rem}_2 = \sqrt{n}(\hat{M}_{\bar{\theta},\bar{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\theta} - \theta_0) + \sqrt{n}(\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\gamma} - \gamma_0)$, where

$$\begin{aligned} (\hat{M}_{\bar{\theta},\bar{\gamma}} - \hat{M}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\theta} - \theta_0) &= \phi^{-1}\mathbb{E}_n\big[W_i\kappa_i(\tilde{\gamma})\{\ddot{b}(W_i^{\mathrm{T}}\bar{\theta}) - \ddot{b}(W_i^{\mathrm{T}}\tilde{\theta})\}W_i^{\mathrm{T}}(\tilde{\theta} - \theta_0)\big] \text{ and} \\ (\hat{N}_{\theta_0,\bar{\gamma}} - \hat{N}_{\tilde{\theta},\tilde{\gamma}})(\tilde{\gamma} - \gamma_0) &= \mathbb{E}_n\big[z_{\theta_0,i}\{\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\bar{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\bar{\gamma})\}X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)\big] \\ &+ \mathbb{E}_n\big[(z_{\tilde{\theta},i} - z_{\theta_0,i})\dot{\kappa}_0(\Lambda(X_i^{\mathrm{T}}\tilde{\gamma}))\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)\big]. \end{aligned}$$

Recall that $\hat{B}_{\tilde{\theta},\tilde{\gamma},j}$ is the jth row of $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$. Given $|\hat{B}_{\tilde{\theta},\tilde{\gamma},j} - B_{\theta_0,\gamma_0,j}|_2 \le |\hat{B}_{\tilde{\theta},\tilde{\gamma},j} - B_{\theta_0,\gamma_0,j}|_1 = o(1)$ for any $j = 1, \ldots, p_0$ from Lemma 2, and $\max_{1\le j\le p_0}|B_{\theta_0,\gamma_0,j}|_2 \le C$, we have $\max_{1\le j\le p_0}|\hat{B}_{\tilde{\theta},\tilde{\gamma},j}|_2 \le C$, and $\hat{B}_{\tilde{\theta},\tilde{\gamma},j}^{\mathrm{T}}W_i$ is sub-Gaussian distributed for any $j = 1, \ldots, p_0$. Similar to the proofs of (S.2.1) and (S.2.2), it can be shown that

$$|\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathrm{Rem}_2|_\infty \le Cs_1\max\{s_2, \log(p_0)\}\log^{3/2}(p_0)/\sqrt{n}$$

with probability converging to 1. Also notice that

$$\big|\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}(\hat{N}_{\tilde{\theta},\tilde{\gamma}} - \check{N}_{q_0})(\tilde{\gamma}-\gamma_0)\big|_\infty \le s_1^{3/2} s_5 \max(\sqrt{s_2}, \log p_0)\log(p_0)/\sqrt{n}.$$

Therefore, it follows that

$$\begin{aligned}
&\sqrt{n}(\tilde{\theta}-\theta_0) + \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\big] \\
&\qquad = \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] - \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\gamma_0) + o_p(1)
\end{aligned}$$

given $\lambda_4 \asymp \|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{M}_{\tilde{\theta},\tilde{\gamma}} - M_{\theta_0,\gamma_0}\|_\infty \asymp s_5\{s_1 s_2 \log(p_0)/n\}^{1/2}$ and $s_0^{7/2}\log^{5/2}(p_0)/\sqrt{n} = o(1)$, where $s_0 = \max_{1\le k\le 5}\{s_k\}$. □

Proof of (4.9). From (S.2.3) and (S.5.13), we have

$$\begin{aligned}
\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\gamma_0) &= \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\hat{\gamma}) + \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\hat{\gamma}-\gamma_0) \\
&= \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}(\tilde{\gamma}-\hat{\gamma}) + \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}\big\{\Xi_{\gamma_0}\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big] + \mathrm{Rem}_a\big\},
\end{aligned}$$

where $\mathrm{Rem}_a = \hat{\Xi}_{\tilde{\gamma}}R_1 + R_2 + R_3$, $R_1 = \mathbb{E}_n\big[\{\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\Lambda}(\tilde{a}_i)\}X_iX_i^{\mathrm{T}}\big](\tilde{\gamma}-\gamma_0)$ from (S.5.3), $R_2 = (I - \hat{\Xi}_{\tilde{\gamma}}\mathbf{X}^{\mathrm{T}}G_{\tilde{\gamma}}^2\mathbf{X}/n)(\tilde{\gamma}-\gamma_0)$ from (S.5.5) and $R_3 = (\hat{\Xi}_{\tilde{\gamma}} - \Xi_{\gamma_0})\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big]$ from (S.5.13) in the proof of Lemma 1. We have shown that with probability converging to 1, $|\hat{\Xi}_{\tilde{\gamma}}R_1|_\infty \le Cs_1\log^{3/2}(p)/n$, $|R_2|_\infty \le Cs_1\log^{3/2}(p)/n$, and $|R_3|_\infty \le C(s_3+s_1)\log^{3/2}(p)/n$ in Lemma 1. It follows that $|\sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}\mathrm{Rem}_a|_\infty \le \sqrt{n}\|\hat{B}_{\tilde{\theta},\tilde{\gamma}}\|_{\ell_\infty}\|\check{N}_{q_0}\|_{\ell_\infty}|\mathrm{Rem}_a|_\infty \le C(s_1+s_3)s_4s_5\log^{3/2}(p)/\sqrt{n}$, which leads to (4.9). □

Proof of Theorem 2. From (4.9), we have

$$\sqrt{n}(\hat{\theta}-\theta_0) \;=\; \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] - \sqrt{n}\hat{B}_{\tilde{\theta},\tilde{\gamma}}\check{N}_{q_0}\Xi_{\gamma_0}\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big] + o_p(1).$$

Note that $\mathbb{E}\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big] = \mathbb{E}_{D_1>D_0}(z_{\theta_0,i}) = 0$. From the proof of Theorem 1, it has been shown that $\big|\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big]\big|_\infty = |\nabla\mathcal{L}_{n,\gamma_0}(\theta_0)|_\infty \le C\{\log(p_0)/n\}^{1/2}$ for a positive constant C with probability converging to 1 as $n,p\to\infty$. By large deviation results, we also have

$\big|\mathbb{E}_n\big[\Xi_{\gamma_0}\dot\rho_{\gamma_0}(X_i,Z_i)\big]\big|_\infty \le C\{\log(p)/n\}^{1/2}$ with probability converging to 1 as $n,p\to\infty$. It follows that

$$\begin{aligned}\big|\sqrt{n}(\hat B_{\tilde\theta,\tilde\gamma}-B_{\theta_0,\gamma_0})\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big]\big|_\infty &\le \|\hat B_{\tilde\theta,\tilde\gamma}-B_{\theta_0,\gamma_0}\|_{\ell_\infty}\big|\sqrt{n}\mathbb{E}_n\big[\kappa_i(\gamma_0)z_{\theta_0,i}\big]\big|_\infty\\ &\le C\{\log(p_0)\}^{1/2}\|\hat B_{\tilde\theta,\tilde\gamma}-B_{\theta_0,\gamma_0}\|_{\ell_\infty},\\ &\le C\{\log(p_0)\}^{1/2}c_2(p)(s_5\lambda_4)^{1-q_2}\end{aligned}$$

for $\lambda_4\asymp s_5\{s_1s_2\log(p_0)/n\}^{1/2}$ by Lemma 2, which converges to 0 given

$$\{\log(p_0)\}^{1/2}c_2(p)\{s_0^6\log(p_0)/n\}^{(1-q_2)/2}\to 0$$

as $n,p\to\infty$. Also notice that $\sqrt{n}\big|\Xi_{\gamma_0}\mathbb{E}_n\big[\dot\rho_{\gamma_0}(X_i,Z_i)\big]\big|_\infty\le C\{\log(p)\}^{1/2}$, and

$$\begin{aligned}\|\hat B_{\tilde\theta,\tilde\gamma}\check N_{q_0}-B_{\theta_0,\gamma_0}N_{\theta_0,\gamma_0}\|_{\ell_\infty} &\le \|\hat B_{\tilde\theta,\tilde\gamma}\|_{\ell_\infty}\|\check N_{q_0}-N_{\theta_0,\gamma_0}\|_{\ell_\infty}+\|\hat B_{\tilde\theta,\tilde\gamma}-B_{\theta_0,\gamma_0}\|_{\ell_\infty}\|N_{\theta_0,\gamma_0}\|_{\ell_\infty}\\ &\le s_5\lambda_3^{1-q_1}c_1(p)+s_4\|\hat B_{\tilde\theta,\tilde\gamma}-B_{\theta_0,\gamma_0}\|_{\ell_\infty}.\\ &\le s_5\lambda_3^{1-q_1}c_1(p)+s_4c_2(p)\{s_0^6\log(p_0)/n\}^{(1-q_2)/2}.\end{aligned}$$

This implies

$$\begin{aligned}&\big|\sqrt{n}(\hat B_{\tilde\theta,\tilde\gamma}\check N_{q_0}-B_{\theta_0,\gamma_0}N_{\theta_0,\gamma_0})\Xi_{\gamma_0}\mathbb{E}_n\big[\dot\rho_{\gamma_0}(X_i,Z_i)\big]\big|_\infty\\ &\qquad\le C\{\log(p)\}^{1/2}\big[s_5\lambda_3^{1-q_1}c_1(p)+s_4c_2(p)\{s_0^6\log(p_0)/n\}^{(1-q_2)/2}\big],\end{aligned}$$

which converges to 0 as $n,p\to\infty$. The conclusion of Theorem 2 follows. □

3 Proof of Theorem 3.

From (5.5) in Theorem 2, $V=B_{\theta_0,\gamma_0}V_0B_{\theta_0,\gamma_0}^{\mathrm{T}}$ is the variance of the leading order term in the expansion (5.5), where V_0 is given in (4.11). We first derive the difference between V and $\hat V$. Let

$$\hat V_0=\frac{1}{n}\sum_{i=1}^n\big\{\kappa_i(\tilde\gamma)z_{\tilde\theta,i}-\check N_{q_0}\hat\Xi_{\tilde\gamma}\dot\rho_{\tilde\gamma}(X_i,Z_i)\big\}\big\{\kappa_i(\tilde\gamma)z_{\tilde\theta,i}-\check N_{q_0}\hat\Xi_{\tilde\gamma}\dot\rho_{\tilde\gamma}(X_i,Z_i)\big\}^{\mathrm{T}},$$

such that the estimated variance $\hat{V}$ in (4.12) can be written as $\hat{V} = \hat{B}_{\tilde{\theta},\tilde{\gamma}}\hat{V}_0\hat{B}^{\mathrm{T}}_{\tilde{\theta},\tilde{\gamma}}$. From (4.11), we have $V_0 = V_{0,1} + V_{0,2} - V_{0,3} - V_{0,3}^{\mathrm{T}}$ where $V_{0,1} = \mathbb{E}\{\kappa_i^2(\gamma_0)z_{\theta_0,i}z_{\theta_0,i}^{\mathrm{T}}\}$, $V_{0,2} = N_{\theta_0,\gamma_0}\Xi_{\gamma_0}\mathbb{E}\{\dot{\rho}_{\gamma_0}(X_i,Z_i)\dot{\rho}_{\gamma_0}^{\mathrm{T}}(X_i,Z_i)\}\Xi_{\gamma_0}N_{\theta_0,\gamma_0}^{\mathrm{T}}$ and $V_{0,3} = \mathbb{E}\{\kappa_i(\gamma_0)z_{\theta_0,i}\dot{\rho}_{\gamma_0}^{\mathrm{T}}(X_i,Z_i)\}\Xi_{\gamma_0}N_{\theta_0,\gamma_0}^{\mathrm{T}}$. Correspondingly, we decompose $\hat{V}_0 = \hat{V}_{0,1} + \hat{V}_{0,2} - \hat{V}_{0,3} - \hat{V}_{0,3}^{\mathrm{T}}$.

Note that $\hat{V}_{0,1} - V_{0,1} = \mathbb{E}_n\big[\{\kappa_i^2(\tilde{\gamma}) - \kappa_i^2(\gamma_0)\}z_{\tilde{\theta},i}z_{\tilde{\theta},i}^{\mathrm{T}}\big] + \mathbb{E}_n\big[\kappa_i^2(\gamma_0)(z_{\tilde{\theta},i}z_{\tilde{\theta},i}^{\mathrm{T}} - z_{\theta_0,i}z_{\theta_0,i}^{\mathrm{T}})\big] + \big\{\mathbb{E}_n\big[\kappa_i^2(\gamma_0)z_{\theta_0,i}z_{\theta_0,i}^{\mathrm{T}}\big] - \mathbb{E}\big[\kappa_i^2(\gamma_0)z_{\theta_0,i}z_{\theta_0,i}^{\mathrm{T}}\big]\big\}$, where by the Lipschitz continuity of $\kappa_0(\cdot, D_i, Z_i)$ and $\dot{b}(\cdot)$, the first term and the second term on the right side of this equality are at the orders $\log^2(p)\sqrt{s_1}\lambda_1$ and $\log(p)\sqrt{s_2}\lambda_2$, respectively. By large deviation results, the third term is bounded by $C\{\log(p)/n\}^{1/2}$. It follows that $\hat{V}_{0,1} - V_{0,1} = O_p\big\{\sqrt{s_1 s_2}\log^{5/2}(p)/\sqrt{n}\big\}$.

For $\hat{V}_{0,2} - V_{0,2}$, it can be decomposed as $\check{N}_{q_0}\big\{\hat{\Xi}_{\tilde{\gamma}}\mathbb{E}_n\big[\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\dot{\rho}_{\tilde{\gamma}}^{\mathrm{T}}(X_i,Z_i)\big]\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}} - \Xi_{\gamma_0}\big\}\check{N}_{q_0}^{\mathrm{T}} + (\check{N}_{q_0}\Xi_{\gamma_0}\check{N}_{q_0}^{\mathrm{T}} - N_{\theta_0,\gamma_0}\Xi_{\gamma_0}N_{\theta_0,\gamma_0}^{\mathrm{T}})$. From the proof of Lemma 1, $\|\hat{\mathcal{I}}_{\tilde{\gamma}} - \mathcal{I}_{\gamma_0}\|_\infty = O_p(\{s_1\log(p)/n\}^{1/2})$, where $\hat{\mathcal{I}}_{\tilde{\gamma}} = \mathbb{E}_n\big[\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\dot{\rho}_{\tilde{\gamma}}^{\mathrm{T}}(X_i,Z_i)\big]$. It implies $\|\hat{\Xi}_{\tilde{\gamma}}\hat{\mathcal{I}}_{\tilde{\gamma}}\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}} - \Xi_{\gamma_0}\|_\infty \le \|\hat{\Xi}_{\tilde{\gamma}}(\hat{\mathcal{I}}_{\tilde{\gamma}} - \mathcal{I}_{\gamma_0})\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}}\|_\infty + \|(\hat{\Xi}_{\tilde{\gamma}} - \Xi_{\gamma_0})\mathcal{I}_{\gamma_0}\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}}\|_\infty + \|\hat{\Xi}_{\tilde{\gamma}}\mathcal{I}_{\gamma_0}(\hat{\Xi}_{\tilde{\gamma}} - \Xi_{\gamma_0})^{\mathrm{T}}\|_\infty \le \|\Xi_{\gamma_0}\|_{\ell_1}^2\|\hat{\mathcal{I}}_{\tilde{\gamma}} - \mathcal{I}_{\gamma_0}\|_\infty = O_p(s_3^2\{s_1\log(p)/n\}^{1/2} + (s_1+s_3)\log(p)/\sqrt{n})$. Given $\|\check{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} \to 0$, we have

$$\begin{aligned}\|\hat{V}_{0,2} - V_{0,2}\|_\infty &\le \|N_{\theta_0,\gamma_0}\|_{\ell_\infty}^2\|\hat{\Xi}_{\tilde{\gamma}}\hat{\mathcal{I}}_{\tilde{\gamma}}\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}} - \Xi_{\gamma_0}\|_\infty + C\|N_{\theta_0,\gamma_0}\|_{\ell_\infty}\|\check{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} \\ &\le Cs_3^2s_4^2\{s_1\log(p)/n\}^{1/2} + C(s_1+s_3)s_4^2\log(p)/\sqrt{n} + Cs_4\lambda_3^{1-q_1}c_1(p)\end{aligned}$$

for a positive constant C, with probability converging to 1. Similarly,

$$\begin{aligned}\hat{V}_{0,3} - V_{0,3} &= \big[\mathbb{E}_n\{\kappa_i(\tilde{\gamma})z_{\tilde{\theta},i}\dot{\rho}_{\tilde{\gamma}}^{\mathrm{T}}(X_i,Z_i)\} - \mathbb{E}\{\kappa_i(\gamma_0)z_{\theta_0,i}\dot{\rho}_{\gamma_0}^{\mathrm{T}}(X_i,Z_i)\}\big]\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}}\check{N}_{q_0}^{\mathrm{T}} \\ &+ \mathbb{E}\{\kappa_i(\gamma_0)z_{\theta_0,i}\dot{\rho}_{\gamma_0}^{\mathrm{T}}(X_i,Z_i)\}(\hat{\Xi}_{\tilde{\gamma}}\check{N}_{q_0} - \Xi_{\gamma_0}N_{\theta_0,\gamma_0})^{\mathrm{T}},\end{aligned}$$

which is bounded by $C\|\Xi_{\gamma_0}\|_{\ell_1}\|N_{\theta_0,\gamma_0}\|_{\ell_\infty}\sqrt{s_1 s_2}\log^{3/2}(p)/\sqrt{n} + C\|\hat{\Xi}_{\tilde{\gamma}}\check{N}_{q_0} - \Xi_{\gamma_0}N_{\theta_0,\gamma_0}\|_{\ell_\infty}$ with probability converging to 1. Since $\|\hat{\Xi}_{\tilde{\gamma}}\check{N}_{q_0} - \Xi_{\gamma_0}N_{\theta_0,\gamma_0}\|_{\ell_\infty} \le \|\Xi_{\gamma_0}\|_{\ell_1}\|\check{N}_{q_0} - N_{\theta_0,\gamma_0}\|_{\ell_\infty} + \|\hat{\Xi}_{\tilde{\gamma}} - \Xi_{\gamma_0}\|_{\ell_1}\|N_{\theta_0,\gamma_0}\|_{\ell_\infty}$, from the proof of Lemma 1 and (S.2.2), it turns out that

$$\|\hat{V}_{0,3} - V_{0,3}\|_\infty \le C\sqrt{s_1 s_2}s_3 s_4\log^{3/2}(p)/\sqrt{n} + Cs_3\lambda_3^{1-q_1}c_1(p) + C(s_1+s_3)s_4\log(p)/\sqrt{n}.$$

Combining all the three parts together, we have $\|\hat{V}_0 - V_0\|_\infty \le s_0^{9/2}\log^{5/2}(p)/\sqrt{n} + s_0\lambda_3^{1-q_1}c_1(p)$. Note that $\hat{V} - V = \hat{B}_{\tilde{\theta},\tilde{\gamma}}(\hat{V}_0 - V_0)\hat{B}_{\tilde{\theta},\tilde{\gamma}}^{\mathrm{T}} + (\hat{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0})V_0\hat{B}_{\tilde{\theta},\tilde{\gamma}}^{\mathrm{T}} + B_{\theta_0,\gamma_0}V_0(\hat{B}_{\tilde{\theta},\tilde{\gamma}} -$

$B_{\theta_0,\gamma_0})^{\mathrm{T}}$. Since the singular values of N_{θ_0,γ_0} are bounded, it can be shown that the maximal eigenvalue of V_0 is also bounded. Therefore, with probability converging to 1,

$$\|\hat{V} - V\|_\infty \leq \|B_{\theta_0,\gamma_0}\|_{\ell_1}^2 \|\hat{V}_0 - V_0\|_\infty + \lambda_{\max}(V_0)\|B_{\theta_0,\gamma_0}\|_{\ell_1}\|\hat{B}_{\tilde{\theta},\tilde{\gamma}} - B_{\theta_0,\gamma_0}\|_{\ell_\infty},$$

which converges to 0 under the conditions of Theorem 3.

Based on the asymptotic expansion of $\hat{\theta}$ established in Theorem 2, since the eigenvalues of B_{θ_0,γ_0} and Ξ_{γ_0}, and the singular values of N_{θ_0,γ_0} are bounded, for all $j = 1, \ldots, p_0$, $B_{\theta_0,\gamma_0,j}^{\mathrm{T}} W_i$ and $B_{\theta_0,\gamma_0,j}^{\mathrm{T}} N_{\theta_0,\gamma_0} \Xi_{\gamma_0} X_i$ are sub-Gaussian distributed, where $B_{\theta_0,\gamma_0,j}$ is the jth row of B_{θ_0,γ_0}. Therefore, the second moment of $B_{\theta_0,\gamma_0,j}^{\mathrm{T}} \kappa_i(\gamma_0) z_{\theta_0,i} - B_{\theta_0,\gamma_0,j}^{\mathrm{T}} N_{\theta_0,\gamma_0} \Xi_{\gamma_0} \dot{\rho}_{\gamma_0}(X_i, Z_i)$ exists, and $\sqrt{n}(\hat{\theta}_j - \theta_{0,j}) V_{jj}^{-1/2} \to N(0,1)$ in distribution as $n, p \to \infty$. The claim of Theorem 3 follows by noticing $\hat{V}_{jj} \to V_{jj}$ in probability as $n, p \to \infty$. $\square$

4 Proof of Theorem 4 and Corollary 1.

Following the proofs of Theorems 2 and 3, it can be similarly shown that

$$\sqrt{n} w_c^{\mathrm{T}}(\hat{\theta} - \theta_0) \;=\; \sqrt{n} w_c^{\mathrm{T}} B_{\theta_0,\gamma_0} \mathbb{E}_n\big[\kappa_i(\gamma_0) z_{\theta_0,i}\big] - \sqrt{n} w_c^{\mathrm{T}} B_{\theta_0,\gamma_0} N_{\theta_0,\gamma_0} \Xi_{\gamma_0} \mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i, Z_i)\big] + o_p(1),$$

and $w_c^{\mathrm{T}}(\hat{V} - V) w_c \to 0$ as $n, p \to \infty$. Since $|w_c|_2 < C$, $w_c^{\mathrm{T}} B_{\theta_0,\gamma_0} W_i$ and $w_c^{\mathrm{T}} B_{\theta_0,\gamma_0} N_{\theta_0,\gamma_0} \Xi_{\gamma_0} X_i$ are sub-Gaussian distributed. This implies $\sqrt{n} w_c^{\mathrm{T}}(\hat{\theta} - \theta_0)(w_c^{\mathrm{T}} V w_c)^{-1/2} \to N(0,1)$ in distribution as $n, p \to \infty$, and the conclusion of Theorem 4 follows.

For the target parameter $h(w_c^{\mathrm{T}}\theta_0)$ in Corollary 1, by Taylor expansion, we have $h(w_c^{\mathrm{T}}\hat{\theta}) = h(w_c^{\mathrm{T}}\theta_0) + \dot{h}(w_c^{\mathrm{T}}\theta_0) w_c^{\mathrm{T}}(\hat{\theta} - \theta_0) + \ddot{h}(w_c^{\mathrm{T}}\theta_*)\{w_c^{\mathrm{T}}(\hat{\theta} - \theta_0)\}^2/2$, where θ_* is between $\hat{\theta}$ and θ_0. Since $w_c^{\mathrm{T}}(\hat{\theta} - \theta_0) = O_p(n^{-1/2})$ from Theorem 4, we have

$$\sqrt{n}\{h(w_c^{\mathrm{T}}\hat{\theta}) - h(w_c^{\mathrm{T}}\theta_0)\} = \sqrt{n}\dot{h}(w_c^{\mathrm{T}}\theta_0) w_c^{\mathrm{T}}(\hat{\theta} - \theta_0) + o_p(1).$$

Notice that $\dot{h}(w_c^{\mathrm{T}}\tilde{\theta}) \to \dot{h}(w_c^{\mathrm{T}}\theta_0)$ due to $|\tilde{\theta} - \theta_0|_1 \to 0$. Therefore, $\dot{h}(w_c^{\mathrm{T}}\tilde{\theta})^2 w_c^{\mathrm{T}} \hat{V} w_c \to \dot{h}(w_c^{\mathrm{T}}\theta_0)^2 w_c^{\mathrm{T}} V w_c$ in probability. It follows that $\sqrt{n}\{h(w_c^{\mathrm{T}}\hat{\theta}) - h(w_c^{\mathrm{T}}\theta_0)\}\{\dot{h}(w_c^{\mathrm{T}}\tilde{\theta})^2 w_c^{\mathrm{T}} \hat{V} w_c\}^{-1/2}$ converges to the standard normal distribution as $n, p \to \infty$.

For the second claim, similarly by Taylor expansion, we have

$$h(w_1^{\mathrm{T}}\hat{\theta})-h(w_0^{\mathrm{T}}\hat{\theta}) = h(w_1^{\mathrm{T}}\theta_0)-h(w_0^{\mathrm{T}}\theta_0)+\dot{h}(w_1^{\mathrm{T}}\theta_0)w_1^{\mathrm{T}}(\hat{\theta}-\theta_0)-\dot{h}(w_0^{\mathrm{T}}\theta_0)w_0^{\mathrm{T}}(\hat{\theta}-\theta_0)+o_p(n^{-1}).$$

Let $\tilde{w}_a = \dot{h}(w_1^{\mathrm{T}}\theta_0)w_1 - \dot{h}(w_0^{\mathrm{T}}\theta_0)w_0$. This implies that

$$\sqrt{n}\big[\{h(w_1^{\mathrm{T}}\hat{\theta}) - h(w_0^{\mathrm{T}}\hat{\theta})\} - \{h(w_1^{\mathrm{T}}\theta_0) - h(w_0^{\mathrm{T}}\theta_0)\}\big] = \sqrt{n}\tilde{w}_a^{\mathrm{T}}(\hat{\theta} - \theta_0) + o_p(1).$$

Since $\tilde{w}_a^{\mathrm{T}}V\tilde{w}_a = \dot{h}(w_1^{\mathrm{T}}\theta_0)^2 w_1^{\mathrm{T}}Vw_1 + \dot{h}(w_0^{\mathrm{T}}\theta_0)^2 w_0^{\mathrm{T}}Vw_0 - 2\dot{h}(w_1^{\mathrm{T}}\theta_0)\dot{h}(w_0^{\mathrm{T}}\theta_0)w_0^{\mathrm{T}}Vw_1$, it can be similarly shown that $w_a^{\mathrm{T}}\hat{V}w_a$ is a consistent estimator of $\tilde{w}_a^{\mathrm{T}}V\tilde{w}_a$ for $w_a = \dot{h}(w_1^{\mathrm{T}}\tilde{\theta})w_1 - \dot{h}(w_0^{\mathrm{T}}\tilde{\theta})w_0$. Following a similar proof of Theorem 4, this implies the second claim. □

5 Proof of Lemma 1

This derivation mainly follows the theoretical treatment for the de-sparsifying Lasso in van de Geer et al. (2014) under bounded covariates. For the completeness of the presentation, we provide a detail proof for the asymptotic expansion of the de-sparsifying Lasso estimator under Logistic regression with sub-Gaussian distributed covariates.

First, we show the theoretical properties of the Lasso estimator $\tilde{\gamma}$ in (3.1). From Assumption 2, each component of the covariates X_i is sub-Gaussian distributed. This leads to $\max_{i,j}|X_{ij}| \leq C\{\log(p)\}^{1/2}$ for a positive constant C. From Assumption 4, $c_0 \leq \dot{\Lambda}(X_i^{\mathrm{T}}\gamma_0) \leq 1-c_0$ for a small positive constant c_0, and all $i = 1, \ldots, n$. This implies $(1-c_0)\mathbb{E}(X_iX_i^{\mathrm{T}}) \succ \mathbb{E}[\dot{\Lambda}(X_i^{\mathrm{T}}\gamma_0)X_iX_i^{\mathrm{T}}]$, and $\lambda_{\min}(\mathbb{E}(X_iX_i^{\mathrm{T}})) \geq (1-c_0)^{-1}\lambda_{\min}(\mathcal{I}_{\gamma_0})$. Since $\lambda_{\min}(\mathcal{I}_{\gamma_0}) \geq 1/C$ by Assumption 5, the compatibility condition for Lasso is satisfied. Note that $\log(p)/\sqrt{n} = o(1)$ and $\{\log(p)\}^{1/2}\lambda_1 s_1 = o(1)$ for $\lambda_1 \asymp \{\log(p)/n\}^{1/2}$ by Assumption 3. By Lemma 6.8 and Corollary 6.6 of Bühlmann and Van de Geer (2011), we have

$$|\tilde{\gamma} - \gamma_0|_1 \leq Cs_1\lambda_1 \ \text{ and } \ \mathbb{E}\big\{\mathbb{E}_n[\rho_{\gamma_0}(X_i, Z_i)] - \mathbb{E}_n[\rho_{\tilde{\gamma}}(X_i, Z_i)]|\mathbf{X}\big\} \leq Cs_1\lambda_1^2 \qquad \text{(S.5.1)}$$

with probability $1-\exp(-c_0)$ for any arbitrarily large constant c_0 and $\lambda_1 \asymp \{\log(p)/n\}^{1/2}$.

Recall that $\rho(z,a) = z\log(\Lambda(a)) + (1-z)\log(1-\Lambda(a)) = za - \log(1+\exp(a))$. Since $-\ddot{\rho}(z,a) = \dot{\Lambda}(a) > c_1 > 0$ for any $|a| \le \max_{1\le i\le n}\{X_i^{\mathrm{T}}\gamma_0\}$ by Assumption 4, $-\rho(z,a)$ is strictly convex in a. Notice that $\dot{\rho}(z,a) = z - \Lambda(a)$. we have

$$-\rho(z,a_1) + \rho(z,a) \ge -(z-\Lambda(a))(a_1-a) + c_1(a_1-a)^2/2$$

for $|a_1|, |a| \le \max_{1\le i\le n}\{X_i^{\mathrm{T}}\gamma_0\}$. For any γ in a neighborhood of γ_0, from the above inequality, it follows that

$$\rho_{\gamma_0}(X_i,Z_i) - \rho_{\gamma}(X_i,Z_i) \ge -(Z_i - \Lambda(X_i^{\mathrm{T}}\gamma_0))X_i^{\mathrm{T}}(\gamma-\gamma_0) + c_1(X_i^{\mathrm{T}}(\gamma-\gamma_0))^2/2. \quad \text{(S.5.2)}$$

Taking the expectation with respect to Z_i given $\mathbf{X}$ leads to

$$\mathbb{E}\big\{\rho_{\gamma_0}(X_i,Z_i) - \rho_{\gamma}(X_i,Z_i)|\mathbf{X}\big\} \ge c_1(X_i^{\mathrm{T}}(\gamma-\gamma_0))^2/2.$$

Summing the above inequality over $i = 1,\ldots,n$, we have $\mathbb{E}_n[X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)]^2 \le Cs_1\lambda_1^2$. Let $\Sigma_x = \mathbb{E}(X_iX_i^{\mathrm{T}})$ and $\hat{\Sigma}_x = \mathbb{E}_n(X_iX_i^{\mathrm{T}})$. Since X_{ij} is sub-Gaussian distributed by Assumption 2 for $j = 1,\ldots,p$, by large deviation results, $\|\Sigma_x - \hat{\Sigma}_x\|_\infty \le C\{\log(p)/n\}^2$ for a positive constant C with probability converging to 1 as $n,p\to\infty$. It follows that

$$(\tilde{\gamma}-\gamma_0)^{\mathrm{T}}(\Sigma_x - \hat{\Sigma}_x)(\tilde{\gamma}-\gamma_0) \le \|\Sigma_x - \hat{\Sigma}_x\|_\infty |\tilde{\gamma}-\gamma_0|_1^2 \le Cs_1^2\lambda_1^3 = o(s_1\lambda_1^2)$$

by Assumption 3. It turns out $(\tilde{\gamma}-\gamma_0)^{\mathrm{T}}\Sigma_x(\tilde{\gamma}-\gamma_0) \le Cs_1\lambda_1^2$. Since $\lambda_{\min}(\Sigma_x) > 1/C > 0$, we have $|\tilde{\gamma}-\gamma_0|_2 \le C\sqrt{s_1}\lambda_1$.

Next, for the de-sparsified Lasso estimator $\hat{\gamma}$ in (4.3), consider the derivative of the function $\rho_\gamma(X_i,Z_i)$ in (3.2). For any γ, by Taylor expansion of $\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)$ at γ_0, we have

$$\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i) = \big(Z_i - \Lambda(X_i^{\mathrm{T}}\gamma_0)\big)X_i - \dot{\Lambda}(\tilde{a}_i)X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)X_i,$$

where $\tilde{a}_i$ is between $X_i^{\mathrm{T}}\tilde{\gamma}$ and $X_i^{\mathrm{T}}\gamma_0$. Since $|\dot{\Lambda}(\tilde{a}_i)X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0) - \dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)| \le C\{X_i^{\mathrm{T}}(\tilde{\gamma}-\gamma_0)\}^2$ for a positive constant C independent of i by the Lipschitz continuity of $\dot{\Lambda}(a)$, it follows that

$$\mathbb{E}_n\big[\dot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\big] \;=\; \mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i,Z_i)\big] + \mathbb{E}_n\big[\ddot{\rho}_{\tilde{\gamma}}(X_i,Z_i)\big](\tilde{\gamma}-\gamma_0) + R_1, \quad \text{(S.5.3)}$$

where the reminder term $R_1 = \mathbb{E}_n\big[\{\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\Lambda}(\tilde{a}_i)\}X_i X_i^{\mathrm{T}}\big](\tilde{\gamma} - \gamma_0)$.

Note that $\mathbb{E}_n\big[\ddot{\rho}_{\tilde{\gamma}}(X_i, Z_i)\big] = -\mathbf{X}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X}/n$. Let $\hat{\phi}_{\tilde{\gamma},j,-j}$ be the $p-1$ dimensional subvector of $\hat{\phi}_{\tilde{\gamma},j}$ without the jth coordinate, where we set $\hat{\phi}_{\tilde{\gamma},j,j} = -1$. For the nodewise regression (4.1) and (4.2), by the Karush–Kuhn–Tucker (KKT) conditions, for all $j = 1, \ldots, p$, we have

$$\hat{\phi}_{\tilde{\gamma},j,-j}^{\mathrm{T}} \mathcal{X}_{-j}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X} \hat{\phi}_{\tilde{\gamma},j}/n + \lambda_{1,j}|\hat{\phi}_{\tilde{\gamma},j,-j}|_1 = 0 \ \text{ and } \ |-\mathcal{X}_{-j}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X} \hat{\phi}_{\tilde{\gamma},j}/n|_\infty \le \lambda_{1,j}. \quad \text{(S.5.4)}$$

From the equation in (S.5.4), it follows $\hat{\tau}_{\tilde{\gamma},j}^2 = -\mathcal{X}_j^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X} \hat{\phi}_{\tilde{\gamma},j}/n$, and $\mathcal{X}_j^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X} \hat{\Xi}_{\tilde{\gamma},j}^{\mathrm{T}}/n = 1$, where $\hat{\Xi}_{\tilde{\gamma},j} = -\hat{\phi}_{\tilde{\gamma},j}/\hat{\tau}_{\tilde{\gamma},j}^2$ is the jth row of $\hat{\Xi}_{\tilde{\gamma}}$. From the second inequality in (S.5.4), we have $|\mathcal{X}_{-j}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X} \hat{\Xi}_{\tilde{\gamma},j}^{\mathrm{T}}/n|_\infty \le \lambda_{1,j}/\hat{\tau}_{\tilde{\gamma},j}^2$. Recall that e_j is the p dimensional vector with the jth coordinate being 1 and other entries being 0. The above results indicate

$$|\hat{\Xi}_{\tilde{\gamma},j} \mathbf{X}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X}/n - e_j^{\mathrm{T}}|_\infty \le \lambda_{1,j}/\hat{\tau}_{\tilde{\gamma},j}^2$$

for all $j = 1, \ldots, p$. Let $\lambda_{1,*} = \max_{1 \le j \le p}\{\lambda_{1,j}/\hat{\tau}_{\tilde{\gamma},j}^2\}$. From (S.5.3), we have

$$\tilde{\gamma} - \gamma_0 + \hat{\Xi}_{\tilde{\gamma}} \mathbb{E}_n\big[\dot{\rho}_{\tilde{\gamma}}(X_i, Z_i)\big] = \hat{\Xi}_{\tilde{\gamma}} \mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i, Z_i)\big] + \hat{\Xi}_{\tilde{\gamma}} R_1 + R_2, \quad \text{(S.5.5)}$$

where $R_2 = (I - \hat{\Xi}_{\tilde{\gamma}} \mathbf{X}^{\mathrm{T}} G_{\tilde{\gamma}}^2 \mathbf{X}/n)(\tilde{\gamma} - \gamma_0)$, and $|R_2|_\infty \le \lambda_{1,*}|\tilde{\gamma} - \gamma_0|_1 \le C s_1 \lambda_1 \lambda_{1,*}$.

Let $\phi_{\gamma_0,j} = \arg\min_{\phi_{j,j}=-1} \mathbb{E}(g_{i,\gamma_0} X_i \phi_j)^2$, and $\epsilon_{\gamma_0,j} = -G_{\gamma_0}\mathbf{X}\phi_{\gamma_0,j} = G_{\gamma_0}\mathcal{X}_j - G_{\gamma_0}\mathcal{X}_{-j}\phi_{\gamma_0,j,-j}$ for $\epsilon_{\gamma_0,j} = (\epsilon_{\gamma_0,1,j}, \ldots, \epsilon_{\gamma_0,n,j})^{\mathrm{T}}$ be the error from projecting $\mathcal{X}_j$ on the space of $\mathcal{X}_{-j}$ by the FI matrix $\mathcal{I}_{\gamma_0}$. Let $\tau_{\gamma_0,j}^2 = \mathbb{E}(\epsilon_{\gamma_0,i,j})^2$. It can be shown that $\Xi_{\gamma_0,j} = -\phi_{\gamma_0,j}/\tau_{\gamma_0,j}^2$, $\tau_{\gamma_0,j}^2 = 1/\Xi_{\gamma_0,j,j}$, and $\mathbb{E}(g_{i,\gamma_0} X_{i,j_1} \epsilon_{\gamma_0,i,j}) = 0$ for all $j_1 \ne j$ and all $i = 1, \ldots, n$. Consider the regression

$$G_{\tilde{\gamma}} G_{\gamma_0}^{-1} \epsilon_{\gamma_0,j} = -G_{\tilde{\gamma}} \mathbf{X} \phi_{\gamma_0,j}. \quad \text{(S.5.6)}$$

Let I_n be the $n \times n$ identity matrix. From the definition of the Lasso estimator $\hat{\phi}_{\tilde{\gamma},j}$ in

(4.1), we have

$$
\begin{aligned}
& |G_{\tilde{\gamma}}\mathbf{X}\hat{\phi}_{\tilde{\gamma},j}|_2^2/n - |G_{\tilde{\gamma}}\mathbf{X}\phi_{\gamma_0,j}|_2^2/n + 2\lambda_{1,j}|\hat{\phi}_{\tilde{\gamma},j}|_1 \;\le\; 2\lambda_{1,j}|\phi_{\gamma_0,j}|_1 \\
\iff\; & \frac{1}{n}\sum_{i=1}^{n}\big\{(g_{i,\tilde{\gamma}}X_{ij} - g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}\hat{\phi}_{\tilde{\gamma},j,-j})^2 - (g_{i,\tilde{\gamma}}X_{ij} - g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}\phi_{\gamma_0,j,-j})^2\big\} \\
& \qquad\qquad +2\lambda_{1,j}|\hat{\phi}_{\tilde{\gamma},j}|_1 \;\le\; 2\lambda_{1,j}|\phi_{\gamma_0,j}|_1 \\
\iff\; & \mathbb{E}_n\big[g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 - 2\mathbb{E}_n\big[g_{i,\tilde{\gamma}}^2 g_{i,\gamma_0}^{-1}\epsilon_{\gamma_0,i,j}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big] \\
& \qquad\qquad +2\lambda_{1,j}|\hat{\phi}_{\tilde{\gamma},j}|_1 \;\le\; 2\lambda_{1,j}|\phi_{\gamma_0,j}|_1. \qquad \text{(S.5.7)}
\end{aligned}
$$

Note that $\|G_{\tilde{\gamma}}G_{\gamma_0}^{-1} - I_n\|_\infty = \max_{1\le i\le n}|g_{i,\tilde{\gamma}}/g_{i,\gamma_0} - 1|$. Since g_{i,γ_0} is bounded away from 0 and ∞, it follows $|g_{i,\tilde{\gamma}}^2 - g_{i,\gamma_0}^2| \le CX_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)$. We have

$$
\begin{aligned}
& 2\big|\mathbb{E}_n\big[g_{i,\tilde{\gamma}}^2 g_{i,\gamma_0}^{-1}\epsilon_{\gamma_0,i,j}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big] - \mathbb{E}_n\big[g_{i,\gamma_0}\epsilon_{\gamma_0,i,j}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]\big| \\
\le\; & 2\mathbb{E}_n^{1/2}\big[(g_{i,\tilde{\gamma}}^2 g_{i,\gamma_0}^{-2} - 1)\epsilon_{\gamma_0,i,j}\big]^2\mathbb{E}_n^{1/2}\big[g_{i,\gamma_0}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 \\
\le\; & \delta_0^{-1}\mathbb{E}_n\big[(g_{i,\tilde{\gamma}}^2 g_{i,\gamma_0}^{-2} - 1)\epsilon_{\gamma_0,i,j}\big]^2 + \delta_0\mathbb{E}_n\big[g_{i,\gamma_0}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2
\end{aligned}
$$

for any $0 < \delta_0 < 1$. Since $\epsilon_{\gamma_0,i,j}$ is sub-Gaussian distributed by Assumptions 2 and 5, $\max_{i,j}|\epsilon_{\gamma_0,i,j}| \le C\{\log(p)\}^{1/2}$ for a positive constant C with probability converging to 1. This leads to

$$
\mathbb{E}_n\big[(g_{i,\tilde{\gamma}}^2 g_{i,\gamma_0}^{-2} - 1)\epsilon_{\gamma_0,i,j}\big]^2 \le C\log(p)\mathbb{E}_n\big[X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)\big]^2 \le C\log(p)s_1\lambda_1^2.
$$

Since $\max_{1\le i\le n}|g_{i,\tilde{\gamma}} - g_{i,\gamma_0}| = o(1)$, $\mathbb{E}_n\big[g_{i,\gamma_0}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 \le 2\mathbb{E}_n\big[g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2$ as $n, p \to \infty$. Plugging those results into (S.5.7), for a small δ_0, we have

$$
\begin{aligned}
& (1-\delta_0)\mathbb{E}_n\big[g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 + 2\lambda_{1,j}|\hat{\phi}_{\tilde{\gamma},j}|_1 \\
\le\; & 2\mathbb{E}_n\big[g_{i,\gamma_0}\epsilon_{\gamma_0,i,j}X_{i,-j}^{\mathrm{T}}\big](\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j}) + 2\lambda_{1,j}|\phi_{\gamma_0,j}|_1 + C\log(p)s_1\lambda_1^2, \qquad \text{(S.5.8)}
\end{aligned}
$$

which serves as the basic inequality for Lasso; see Lemma 6.1 in Bühlmann and Van de Geer (2011). Since $\mathbb{E}(g_{i,\gamma_0}X_{i,j_1}\epsilon_{\gamma_0,i,j}) = 0$ for all $j_1 \ne j$, the concentration inequality can be applied to control the maximum of the term $\mathbb{E}_n\big[g_{i,\gamma_0}\epsilon_{\gamma_0,i,j}X_{i,-j}^{\mathrm{T}}\big]$ at the order

$\{\log(p)/n\}^{1/2}$. Also note that $|\phi_{\gamma_0,j}|_0 \le Cs_3$ for $\Xi_{\gamma_0} \in \mathcal{H}_0(s_3, M)$ in (5.1). Following the proof of Theorem 6.1 in Bühlmann and Van de Geer (2011), it can be shown that

$$|\hat{\phi}_{\tilde{\gamma},j} - \phi_{\gamma_0,j}|_1 \le Cs_3\lambda_{1,j} + C\log(p)s_1\lambda_1^2/\lambda_{1,j} \text{ and} \quad \text{(S.5.9)}$$

$$\mathbb{E}_n\big[g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 \le Cs_3\lambda_{1,j}^2 + C\log(p)s_1\lambda_1^2 \quad \text{(S.5.10)}$$

with probability converging to 1 as $n, p \to \infty$.

For $\hat{\tau}_{\tilde{\gamma},j}^2$, notice that $\hat{\tau}_{\tilde{\gamma},j}^2 = -\mathcal{X}_j^{\mathrm{T}}G_{\tilde{\gamma}}^2\mathbf{X}\hat{\phi}_{\tilde{\gamma},j}/n$ and $\epsilon_{\gamma_0,j} = G_{\gamma_0}\mathcal{X}_j - G_{\gamma_0}\mathcal{X}_{-j}\phi_{\gamma_0,j,-j}$. Since $G_{\tilde{\gamma}}\mathcal{X}_j = G_{\tilde{\gamma}}G_{\gamma_0}^{-1}(G_{\gamma_0}\mathcal{X}_{-j}\phi_{\gamma_0,j,-j} + \epsilon_{\gamma_0,j})$, and $-G_{\tilde{\gamma}}\mathbf{X}\hat{\phi}_{\tilde{\gamma},j} = -G_{\tilde{\gamma}}G_{\gamma_0}^{-1}G_{\gamma_0}\mathbf{X}(\phi_{\gamma_0,j} + \hat{\phi}_{\tilde{\gamma},j} - \phi_{\gamma_0,j}) = G_{\tilde{\gamma}}G_{\gamma_0}^{-1}\{\epsilon_{\gamma_0,j} - G_{\gamma_0}\mathbf{X}(\hat{\phi}_{\tilde{\gamma},j} - \phi_{\gamma_0,j})\}$, where the last equation is by (S.5.6), we have

$$\begin{aligned}
\hat{\tau}_{\tilde{\gamma},j}^2 - \tau_{\gamma_0,j}^2 &= (G_{\gamma_0}\mathcal{X}_{-j}\phi_{\gamma_0,j,-j} + \epsilon_{\gamma_0,j})^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-2}\{\epsilon_{\gamma_0,j} - G_{\gamma_0}\mathbf{X}(\hat{\phi}_{\tilde{\gamma},j} - \phi_{\gamma_0,j})\}/n - \tau_{\gamma_0,j}^2 \\
&= \big(\epsilon_{\gamma_0,j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-2}\epsilon_{\gamma_0,j}/n - \tau_{\gamma_0,j}^2\big) + \phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-1}\epsilon_{\gamma_0,j}/n \\
&\quad -\epsilon_{\gamma_0,j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-1}\mathcal{X}_{-j}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})/n \quad \text{(S.5.11)} \\
&\quad -\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2\mathcal{X}_{-j}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})/n.
\end{aligned}$$

For the first term in (S.5.11), $|\epsilon_{\gamma_0,j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-2}\epsilon_{\gamma_0,j}/n - \tau_{\gamma_0,j}^2| \le |\epsilon_{\gamma_0,j}^{\mathrm{T}}\epsilon_{\gamma_0,j}/n - \tau_{\gamma_0,j}^2| + |\epsilon_{\gamma_0,j}^{\mathrm{T}}(G_{\tilde{\gamma}}^2G_{\gamma_0}^{-2} - I_n)\epsilon_{\gamma_0,j}/n| \le C\{\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1$ for all $j = 1, \ldots, p$. For the second term, $|\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-1}\epsilon_{\gamma_0,j}/n| \le |\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\gamma_0}\epsilon_{\gamma_0,j}/n| + |\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\gamma_0}(G_{\tilde{\gamma}}^2G_{\gamma_0}^{-2} - I)\epsilon_{\gamma_0,j}/n| \le C|\phi_{\gamma_0,j,-j}|_1\{\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1 \le C\{s_3\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1$, since $|\phi_{\gamma_0,j,-j}|_2 \le C$ and $\phi_{\gamma_0,j,-j}^{\mathrm{T}}X_{i,-j}$ is sub-Gaussian distributed by Assumption 2. Similarly, the third term is at a smaller order as well. For the last term in (S.5.11), since $|\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2(\mathcal{X}_j - \mathcal{X}_{j-1}\hat{\phi}_{\tilde{\gamma},j,-j})/n|_\infty \le \lambda_{1,j}$ by (S.5.4) and $-\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2\mathcal{X}_{-j}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})/n = \mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2(\mathcal{X}_j - G_{\gamma_0}^{-1}\epsilon_{\gamma_0,j} - \mathcal{X}_{-j}\hat{\phi}_{\tilde{\gamma},j,-j})/n$, we have

$$|-\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2\mathcal{X}_{-j}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})/n|_\infty \le |\phi_{\gamma_0,j,-j}^{\mathrm{T}}|_1\lambda_{1,j} + |\phi_{\gamma_0,j,-j}^{\mathrm{T}}\mathcal{X}_{-j}^{\mathrm{T}}G_{\tilde{\gamma}}^2G_{\gamma_0}^{-1}\epsilon_{\gamma_0,j}/n|_\infty.$$

Therefore, the last term in (S.5.11) is bounded by $Cs_3^{1/2}\lambda_{1,j} + C\{s_3\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1$. Combining the bounds for all the four terms into (S.5.11) leads to

$$|\hat{\tau}_{\tilde{\gamma},j}^2 - \tau_{\gamma_0,j}^2| \le Cs_3^{1/2}\lambda_{1,j} + C\{s_3\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1 \quad \text{(S.5.12)}$$

with probability converging to 1 as $n, p \to \infty$.

Recall that $\hat{\Xi}_{\tilde{\gamma},j} = -\hat{\phi}_{\tilde{\gamma},j}/\hat{\tau}^2_{\tilde{\gamma},j}$ and $\Xi_{\gamma_0,j} = -\phi_{\gamma_0,j}/\tau^2_{\gamma_0,j}$. We can write

$$\begin{aligned}
|\hat{\Xi}_{\tilde{\gamma},j} - \Xi_{\gamma_0,j}|_1 &= |\phi_{\gamma_0,j}/\tau^2_{\gamma_0,j} - \hat{\phi}_{\tilde{\gamma},j}/\tau^2_{\gamma_0,j} + \hat{\phi}_{\tilde{\gamma},j}/\tau^2_{\gamma_0,j} - \hat{\phi}_{\tilde{\gamma},j}/\hat{\tau}^2_{\tilde{\gamma},j}|_1 \\
&\leq C|\phi_{\gamma_0,j} - \hat{\phi}_{\tilde{\gamma},j}|_1 + |\hat{\phi}_{\tilde{\gamma},j}|_1(\hat{\tau}^{-2}_{\tilde{\gamma},j} - \tau^{-2}_{\gamma_0,j}) \\
&\leq Cs_3[\lambda_{1,j} + \{\log(p)/n\}^{1/2}] + C\log(p)s_1\lambda_1^2/\lambda_{1,j} + C\{\log(p)s_1s_3\}^{1/2}\lambda_1
\end{aligned}$$

By choosing $\lambda_1 \asymp \{\log(p)/n\}^{1/2}$ and $\lambda_{1,j} \asymp \log(p)/\sqrt{n}$, it follows $\|\hat{\Xi}^{\mathrm{T}}_{\tilde{\gamma}} - \Xi^{\mathrm{T}}_{\gamma_0}\|_{\ell_1} \leq C(s_3 + s_1)\{\log(p)\}^{1/2}\{\log(p)/n\}^{1/2}$, and $\max_{1\leq j\leq p}|\hat{\tau}^2_{\tilde{\gamma},j} - \tau^2_{\gamma_0,j}| \leq C\log(p)(s_3/n)^{1/2} + C\log(p)(s_1/n)^{1/2} = o(1)$ under Assumptions 3 and 5. These results imply that $\lambda_{1,*} = \max_{1\leq j\leq p}\{\lambda_{1,j}/\hat{\tau}^2_{\tilde{\gamma},j}\} \asymp \log(p)/\sqrt{n}$. Note that

$$\tilde{\gamma} - \gamma_0 + \hat{\Xi}_{\tilde{\gamma}}\mathbb{E}_n\big[\dot{\rho}_{\tilde{\gamma}}(X_i, Z_i)\big] = \Xi_{\gamma_0}\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i, Z_i)\big] + \hat{\Xi}_{\tilde{\gamma}}R_1 + R_2 + R_3 \tag{S.5.13}$$

from (S.5.5), where $R_3 = (\hat{\Xi}_{\tilde{\gamma}} - \Xi_{\gamma_0})\mathbb{E}_n\big[\dot{\rho}_{\gamma_0}(X_i, Z_i)\big]$. We have shown that $R_1 = \mathbb{E}_n\big[\{\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\Lambda}(\tilde{a}_i)\}X_iX_i^{\mathrm{T}}\big](\tilde{\gamma} - \gamma_0)$ in (S.5.3), and $|R_2|_\infty \leq Cs_1\lambda_1\lambda_{1,*} \leq Cs_1\log^{3/2}(p)/n$ in (S.5.5). It follows that

$$\hat{\Xi}_{\tilde{\gamma}}R_1 = \mathbb{E}_n\big[\hat{\Xi}_{\tilde{\gamma}}X_i\{\dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma}) - \dot{\Lambda}(\tilde{a}_i)\}X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)\big].$$

Since $|\dot{\Lambda}(\tilde{a}_i)X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0) - \dot{\Lambda}(X_i^{\mathrm{T}}\tilde{\gamma})X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)| \leq C\{X_i^{\mathrm{T}}(\tilde{\gamma} - \gamma_0)\}^2$ for a positive constant C, $|\hat{\Xi}_{\tilde{\gamma}}R_1|_\infty \leq C\{\log(p)\}^{1/2}s_1\lambda_1^2 \leq C\{\log(p)\}^{3/2}s_1/n$. We also have $|R_3|_\infty \leq C\{\log(p)/n\}^{1/2}\|\hat{\Xi}^{\mathrm{T}}_{\tilde{\gamma}} - \Xi^{\mathrm{T}}_{\gamma_0}\|_{\ell_1} \leq C(s_3 + s_1)\log^{3/2}(p)/n$. Therefore,

$$\sqrt{n}\big(\hat{\gamma} - \gamma_0\big) = \Xi_{\gamma_0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\rho}_{\gamma_0}(X_i, Z_i) + O_p\big\{(s_3 + s_1)\log^{3/2}(p)/\sqrt{n}\big\},$$

which proves Lemma 1. □

In the following, we present the result of Lemma 1 under weakly sparsity condition that allows many small nonzero values for the entries of the inverse Fisher Information matrix Ξ_{γ_0}. We make the following assumption for the sparsity of Ξ_{γ_0} by its ℓ_q norm.

Assumption S5. The Fisher Information matrix $\mathcal{I}_{\gamma_0}$ is positive definite with bounded entries such that $\|\mathcal{I}_{\gamma_0}\|_\infty \le C$ and $\lambda_{\min}(\mathcal{I}_{\gamma_0}) \ge 1/C$, and $\max_{1\le j_1\le p}\sum_{j_2=1}^{p}|\Xi_{\gamma_0,j_1j_2}|^{q_3} \le c_3(p)$ for $0 \le q_3 < 1$ and $c_3(p)\{\log(p)/n\}^{(1-q_3)/2} = o(1)$.

Lemma S1. Suppose that Assumptions 2–4 and S5 hold. For the Lasso parameters $\lambda_1, \lambda_{1,j} \asymp \{\log(p)/n\}^{1/2}$, we have

$$\sqrt{n}\big(\hat{\gamma} - \gamma_0\big) = \Xi_{\gamma_0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\rho}_{\gamma_0}(X_i, Z_i) + O_p\big\{a_n \log^{1/2}(p)\big\} \tag{S.5.14}$$

for $a_n = c_3(p)\{\log(p)/n\}^{(1-q_3)/2} + \{c_3(p)\}^{\frac{2}{2-q_3}}\{\log(p)/n\}^{1/2} + s_1\{\log^3(p)/n\}^{1/2}$.

Proof of Lemma S1. Following the proof of (S.5.9) and (S.5.10) in Lemma 1 and the proof of Theorem 1.8.1 in Van de Geer (2016) for Lasso under weak sparsity, it can be shown that

$$\begin{aligned} |\hat{\phi}_{\tilde{\gamma},j} - \phi_{\gamma_0,j}|_1 &\le Cc_3(p)\lambda_{1,j}^{1-q_3} + C\log(p)s_1\lambda_1^2/\lambda_{1,j}, \quad \text{(S.5.15)} \\ \mathbb{E}_n\big[g_{i,\tilde{\gamma}}X_{i,-j}^{\mathrm{T}}(\hat{\phi}_{\tilde{\gamma},j,-j} - \phi_{\gamma_0,j,-j})\big]^2 &\le Cc_3(p)\lambda_{1,j}^{2-q_3} + C\log(p)s_1\lambda_1^2 \quad \text{(S.5.16)} \end{aligned}$$

for all $j = 1, \ldots, p$ with probability converging to 1 as $n \to \infty$. Notice that under Assumption S5, $\|\Xi_{\gamma_0}\|_{\ell_1} \le C\{c_3(p)\}^{\frac{1}{2-q_3}}$ for a positive constant C. Following the proof of (S.5.12), we have

$$|\hat{\tau}^2_{\tilde{\gamma},j} - \tau^2_{\gamma_0,j}| \le C\lambda_{1,j}\{c_3(p)\}^{\frac{1}{2-q_3}} + C\{c_3(p)\}^{\frac{1}{2-q_3}}\{\log(p)/n\}^{1/2} + C\{\log(p)s_1\}^{1/2}\lambda_1 \tag{S.5.17}$$

with probability converging to 1. Similarly, for all $j = 1, \ldots, p$,

$$\begin{aligned} |\hat{\Xi}_{\tilde{\gamma},j} - \Xi_{\gamma_0,j}|_1 &\le C|\phi_{\gamma_0,j} - \hat{\phi}_{\tilde{\gamma},j}|_1 + |\hat{\phi}_{\tilde{\gamma},j}|_1(\hat{\tau}^{-2}_{\tilde{\gamma},j} - \tau^{-2}_{\gamma_0,j}) \\ &\le Cc_3(p)\lambda_{1,j}^{1-q_3} + C\lambda_{1,j}\{c_3(p)\}^{\frac{2}{2-q_3}} + C\log(p)s_1\lambda_1^2/\lambda_{1,j}. \quad \text{(S.5.18)} \end{aligned}$$

By choosing $\lambda_1, \lambda_{1,j} \asymp \{\log(p)/n\}^{1/2}$, since $\lambda_{1,j}\{c_3(p)\}^{\frac{2}{2-q_3}} \le \{c_3(p)\}^{\frac{1}{1-q_3}}\{\log(p)/n\}^{1/2} = \big[c_3(p)\{\log(p)/n\}^{(1-q_3)/2}\big]^{\frac{1}{1-q_3}} = o(1)$, it follows $\|\hat{\Xi}_{\tilde{\gamma}}^{\mathrm{T}} - \Xi_{\gamma_0}^{\mathrm{T}}\|_{\ell_1} = o_p(1)$ and $\max_{1\le j\le p}|\hat{\tau}^2_{\tilde{\gamma},j} - \tau^2_{\gamma_0,j}| = o_p(1)$ under Assumptions 3 and S5.

From (S.5.5), the result of Lemma S1 follows by bounding $\hat{\Xi}_{\tilde{\gamma}} R_1$, R_2 and R_3 under (S.5.17) and (S.5.18). First, we still have $|\hat{\Xi}_{\tilde{\gamma}} R_1|_\infty \leq C\{\log(p)\}^{1/2} s_1 \lambda_1^2 \leq C\{\log(p)\}^{3/2} s_1/n$. Second, $|R_2|_\infty \leq C s_1 \lambda_1 \lambda_{1,*} \leq C s_1 \log(p)/n$ as $\lambda_{1,*} = \max_{1\leq j\leq p}\{\lambda_{1,j}/\hat{\tau}^2_{\tilde{\gamma},j}\} \asymp \{\log(p)/n\}^{1/2}$. Finally, $|R_3|_\infty \leq C\{\log(p)/n\}^{1/2} \|\hat{\Xi}^{\mathrm{T}}_{\tilde{\gamma}} - \Xi^{\mathrm{T}}_{\gamma_0}\|_{\ell_1} \leq C\{\log(p)/n\}^{1/2} a_n$, where

$$a_n = c_3(p)\{\log(p)/n\}^{(1-q_3)/2} + \{c_3(p)\}^{\frac{2}{2-q_3}}\{\log(p)/n\}^{1/2} + s_1\{\log^3(p)/n\}^{1/2}.$$

Therefore,

$$\sqrt{n}\big(\hat{\gamma} - \gamma_0\big) = \Xi_{\gamma_0} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\rho}_{\gamma_0}(X_i, Z_i) + O_p\big\{a_n \log^{1/2}(p)\big\}. \ \square$$

6 Proof of (8.2)

Under the correct model $h_c(X;\theta)$ for LCSTE(X), we need to show (8.2) is valid if either the instrument propensity score model $\Lambda(X^{\mathrm{T}}\gamma)$ or the regression models $h_y(X, z; \psi_{yz})$ and $h_d(X, z; \psi_{dz})$ are correctly specified, but not necessarily both.

First, if the model for $\mathbb{P}(Z = 1|X)$ is correctly specified, from Theorem 1 in Ogburn et al. (2015), we have

$$\mathbb{E}[\kappa_{\mathrm{diff}}\{Y - Dh_c(X;\theta)\}|X] = 0.$$

Since $\mathbb{E}(\kappa_{\mathrm{diff}}|X) = 0$, it also follows

$$\mathbb{E}[\kappa_{\mathrm{diff}}\{k_y(X;\psi_{y1},\psi_{y0}) - k_d(X;\psi_{d1},\psi_{d0})h_c(X;\theta)\}|X] = 0.$$

Those results together imply (8.2).

Second, consider the case that the regression models $h_y(X, z; \psi_{yz})$ and $h_d(X, z; \psi_{dz})$ are correctly specified. We can write $\mathbb{E}[\nabla_\theta h_c(X;\theta)\kappa_{\mathrm{diff}}\{Y - k_y(X;\psi_{y1},\psi_{y0})\}]$ as

$$\mathbb{E}[\nabla_\theta h_c(X;\theta)\kappa_{\mathrm{diff}}\{Y - h_y(X,Z;\psi_{yZ})\}] + \mathbb{E}[\nabla_\theta h_c(X;\theta)\kappa_{\mathrm{diff}}\{h_y(X,Z;\psi_{yZ}) - k_y(X;\psi_{y1},\psi_{y0})\}],$$

where the first term is equal to $\mathbb{E}[\nabla_\theta h_c(X;\theta)\kappa_{\text{diff}}\mathbb{E}\{Y - h_y(X,Z;\psi_{yZ})|X,Z\}] = 0$ as the model $h_y(X,z;\psi_{yz})$ for Y given X and Z is correctly specified. For the second term, notice that

$$\begin{aligned} h_y(X,1;\psi_{y1}) - k_y(X;\psi_{y1},\psi_{y0}) &= \Lambda(X^{\mathrm{T}}\gamma)\{h_y(X,1;\psi_{y1}) - h_y(X,0;\psi_{y0})\} \text{ and} \\ h_y(X,0;\psi_{y0}) - k_y(X;\psi_{y1},\psi_{y0}) &= -\{1-\Lambda(X^{\mathrm{T}}\gamma)\}\{h_y(X,1;\psi_{y1}) - h_y(X,0;\psi_{y0})\}. \end{aligned}$$

It can be written as

$$\begin{aligned} & \mathbb{E}\big(\nabla_\theta h_c(X;\theta)\mathbb{E}[\kappa_{\text{diff}}\{h_y(X,Z;\psi_{yZ}) - k_y(X;\psi_{y1},\psi_{y0})\}|X]\big) \\ = \; & \mathbb{E}\big(\nabla_\theta h_c(X;\theta)[\Lambda(X^{\mathrm{T}}\gamma)^{-1}\{h_y(X,1;\psi_{y1}) - k_y(X;\psi_{y1},\psi_{y0})\}\mathbb{P}(Z=1|X) \\ & \qquad -\{1-\Lambda(X^{\mathrm{T}}\gamma)\}^{-1}\{h_y(X,0;\psi_{y0}) - k_y(X;\psi_{y1},\psi_{y0})\}\mathbb{P}(Z=0|X)]\big) \\ = \; & \mathbb{E}\big(\nabla_\theta h_c(X;\theta)[\{h_y(X,1;\psi_{y1}) - h_y(X,0;\psi_{y0})\}\mathbb{P}(Z=1|X) \\ & \qquad +\{h_y(X,1;\psi_{y1}) - h_y(X,0;\psi_{y0})\}\mathbb{P}(Z=0|X)]\big) \\ = \; & \mathbb{E}[\nabla_\theta h_c(X;\theta)\{h_y(X,1;\psi_{y1}) - h_y(X,0;\psi_{y0})\}] \\ = \; & \mathbb{E}[\nabla_\theta h_c(X;\theta)\{\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)\}]. \end{aligned}$$

Therefore,

$$\mathbb{E}[\nabla_\theta h_c(X;\theta)\kappa_{\text{diff}}\{Y - k_y(X;\psi_{y1},\psi_{y0})\}] = \mathbb{E}[\nabla_\theta h_c(X;\theta)\{\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)\}]$$

Similarly, it can be shown that $\mathbb{E}[\nabla_\theta h_c(X;\theta)h_c(X;\theta)\kappa_{\text{diff}}\{D - h_d(X,Z;\psi_{dZ})\}] = 0$ as the model $h_d(X,z;\psi_{dz})$ for D given X and Z is correctly specified, and

$$\begin{aligned} & \mathbb{E}[\nabla_\theta h_c(X;\theta)h_c(X;\theta)\kappa_{\text{diff}}\{h_d(X,Z;\psi_{dZ}) - k_d(X;\psi_{d1},\psi_{d0})\}] \\ = \; & \mathbb{E}[\nabla_\theta h_c(X;\theta)h_c(X;\theta)\{\mathbb{E}(D|X,Z=1) - \mathbb{E}(D|X,Z=0)\}]. \end{aligned}$$

Those results imply that

$$\begin{aligned} & \mathbb{E}\big(\nabla_\theta h_c(X;\theta)\kappa_{\text{diff}}[Y - k_y(X;\psi_{y1},\psi_{y0}) - \{D - k_d(X;\psi_{d1},\psi_{d0})\}h_c(X;\theta)]\big) \\ = \; & \mathbb{E}\big(\nabla_\theta h_c(X;\theta)[\{\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)\} \\ & \qquad -h_c(X;\theta)\{\mathbb{E}(D|X,Z=1) - \mathbb{E}(D|X,Z=0)\}]\big), \end{aligned}$$

which is equal to 0, since $h_c(X;\theta) = \frac{\mathbb{E}(Y|X,Z=1)-\mathbb{E}(Y|X,Z=0)}{\mathbb{E}(D|X,Z=1)-\mathbb{E}(D|X,Z=0)}$. This proofs (8.2). □

References

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

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594–607.

Ogburn, E. L., Rotnitzky, A. and Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 373–396.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, **40**, 1637–1664.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, **45**, 866–896.

Van de Geer, S., Bühlmann, P., Ritov, Y. A. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.

Van de Geer, S. (2016). *Lecture notes on sparsity*. Online manuscript.

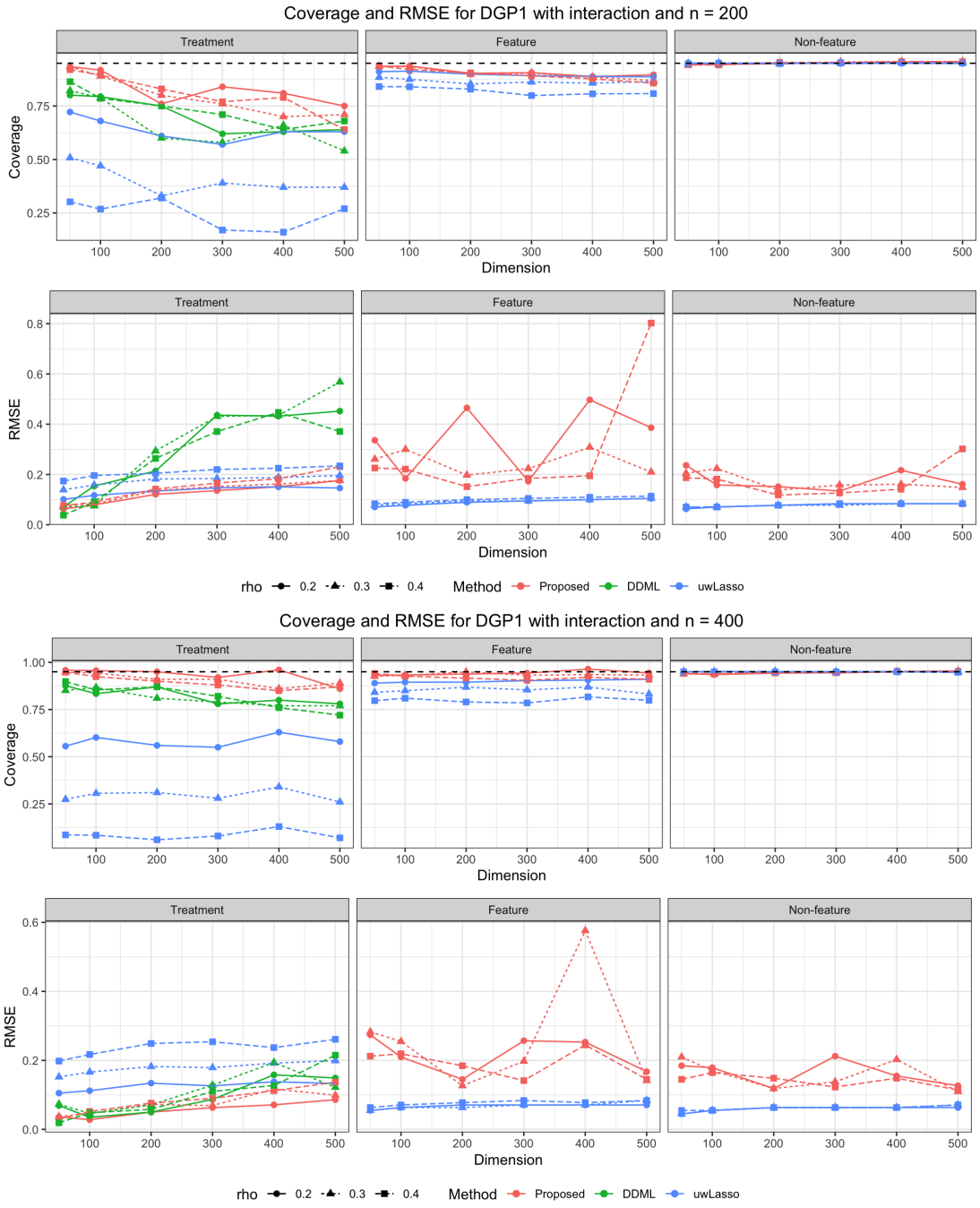

Figure S1: The empirical coverages of confidence intervals and the RMSEs of the estimated coefficients for the proposed method, DDML and uwLasso under DGP1 with interaction, $n = 200, 400$ and $\rho_\epsilon = 0.2, 0.3, 0.4$.

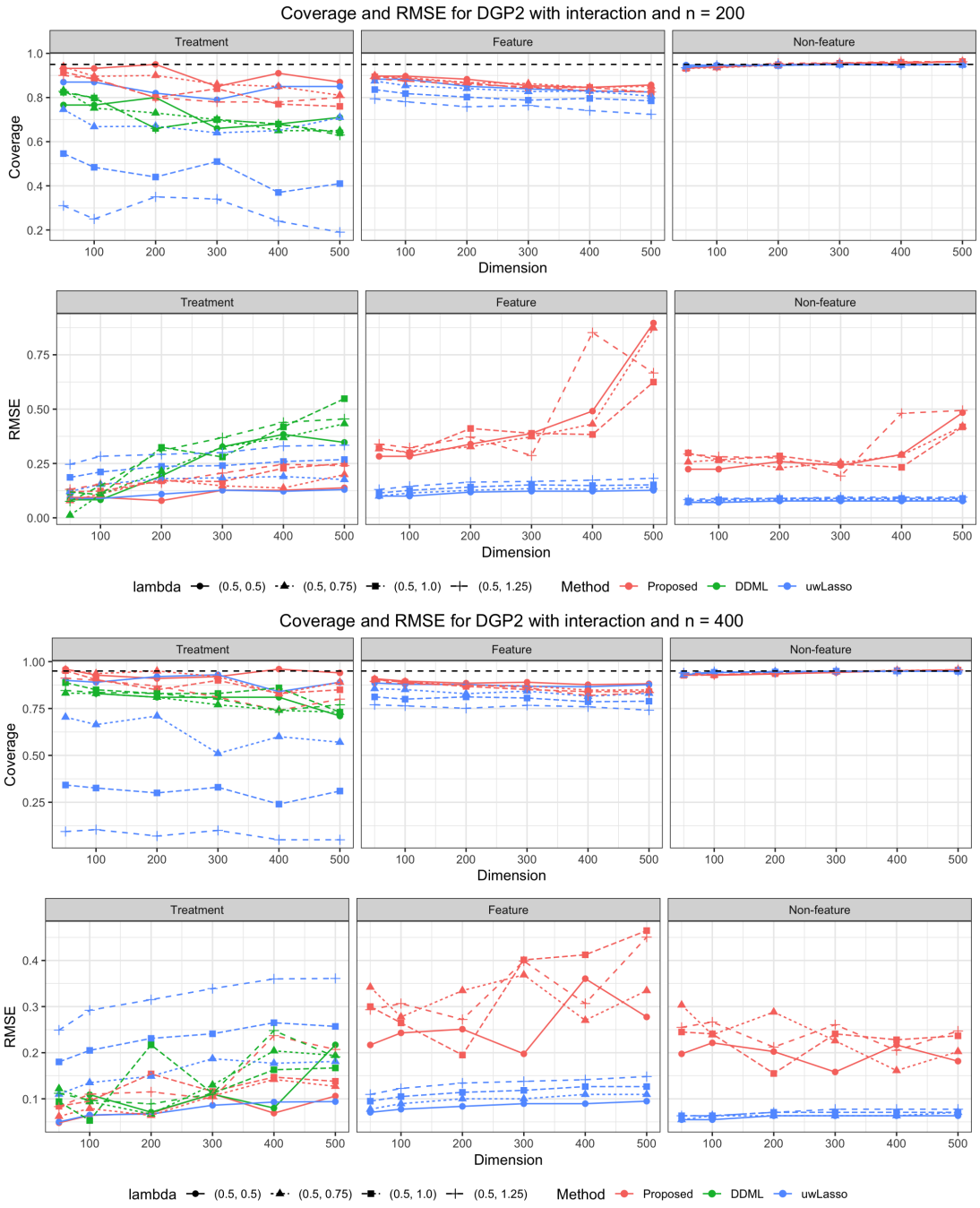

Figure S2: The empirical coverages of confidence intervals and the RMSEs of the estimated coefficients for the proposed method, DDML and uwLasso under DGP2 with interaction, $n = 200, 400$ and $(\lambda_{00}, \lambda_{11}) = (0.5, 0.5), (0.5, 0.75), (0.5, 1), (0.5, 1.25)$.

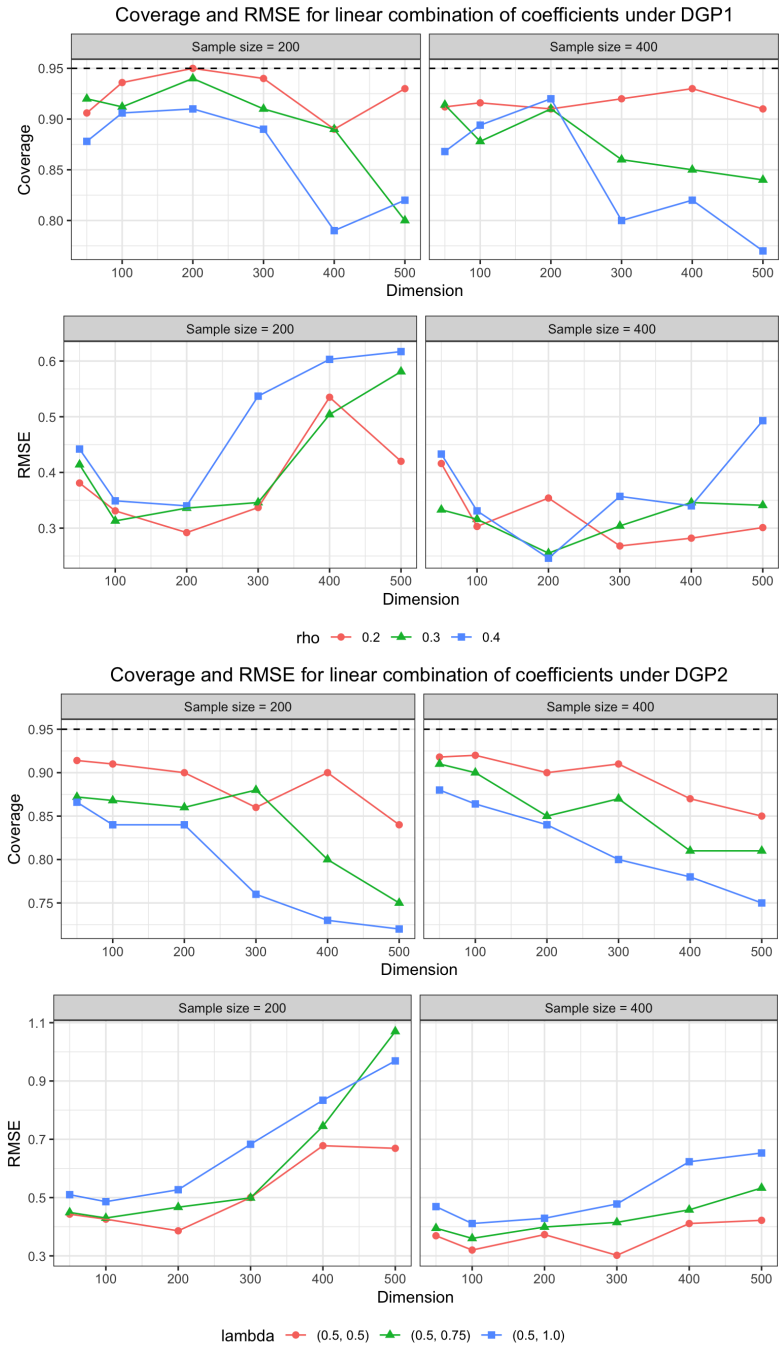

Figure S3: The empirical coverage of the confidence interval for $\alpha_0 + \sum_{j=1}^{10}(10-j)\beta_{0,j}/10$ and the RMSE of its estimate for the proposed method under DGP1 and DGP2 without interaction.

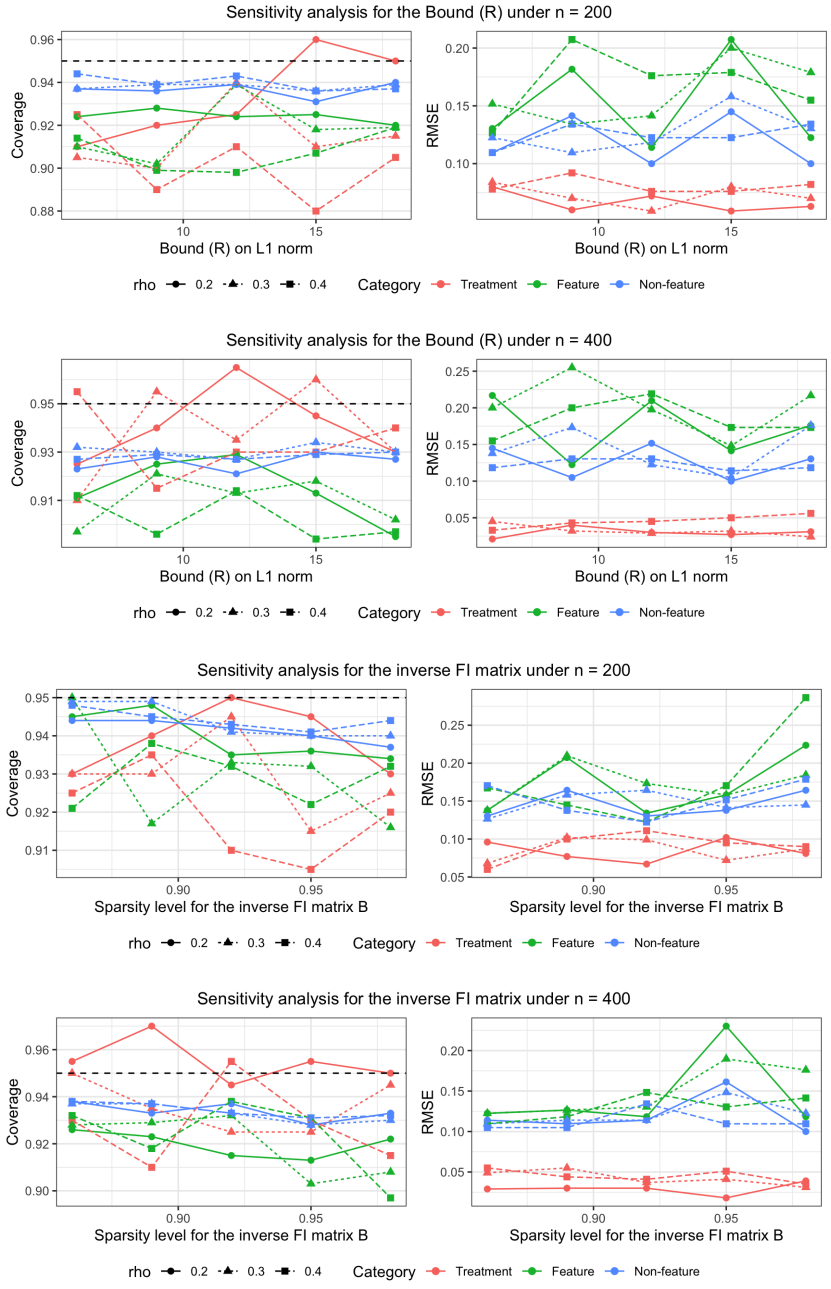

Figure S4: The empirical coverage of the confidence interval and the RMSE of the estimated coefficient for the proposed method with respect to R in (3.5) (with 90% zeros for $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$) and the sparsity of $\hat{B}_{\tilde{\theta},\tilde{\gamma}}$ (proportion of zeros) (with $R = 12$) under DGP1 without interaction, $p = 100$, and $\rho_\epsilon = 0.2, 0.3, 0.4$.

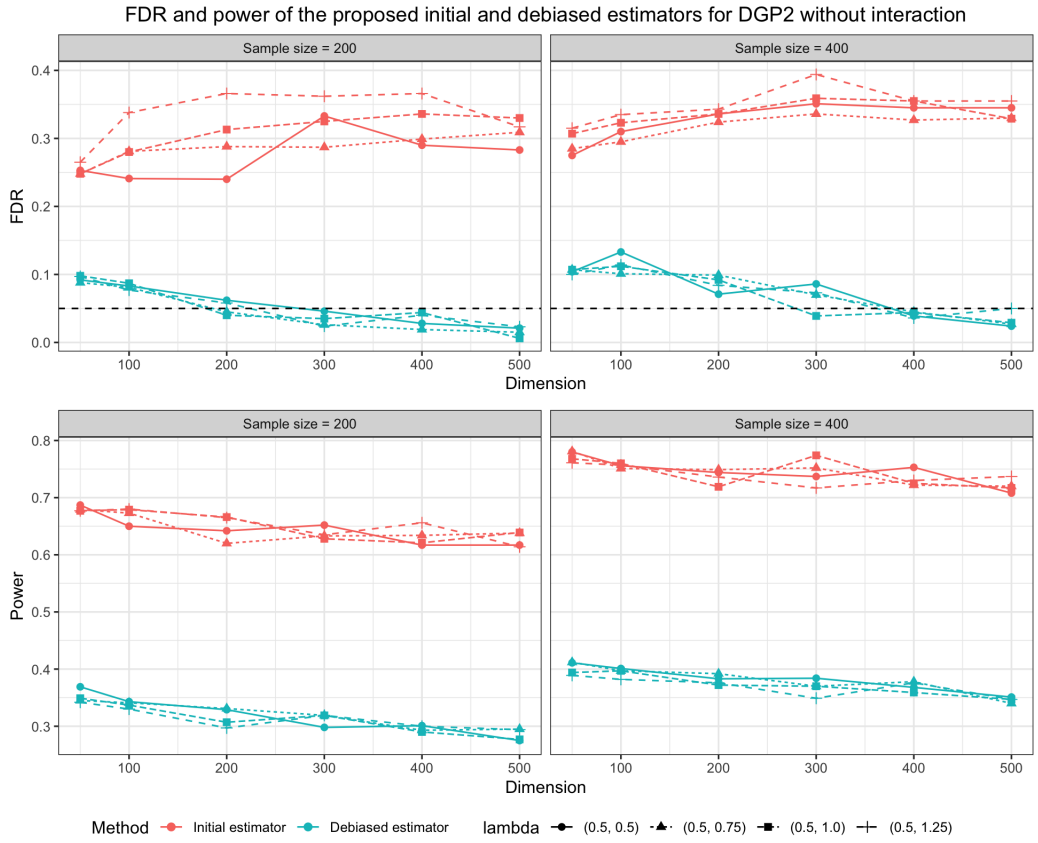

Figure S5: The false discovery rate (FDR = number of false positives/number of discoveries) and the power (number of true positives/number of true signals) of the nonzero components of $\tilde{\theta}$ (Initial estimator) and the Benjamini-Hochberg multiple testing procedure applied on the p-values from the proposed inference procedure (Debiased estimator) under DGP2 without interaction. The signals are regarded as the coefficients with value larger than 0.02. The nominal FDR level controlled by the BH procedure is 0.05.