The Canadian Journal of Statistics
Vol. 50, No. 3, 2022, Pages 795–819
La revue canadienne de statistique

Dummy endogenous treatment effect estimation using high-dimensional instrumental variables

Wei ZHONG[1], Wei ZHOU[1,2], Qingliang FAN[3*], and Yang GAO[1]

[1]*MOE Key Lab of Econometrics, Wang Yanan Institute for Studies in Economics, Department of Statistics & Data Science, School of Economics, and Fujian Key Lab of Statistics, Xiamen University, Xiamen, China*
[2]*School of Data Science, City University of Hong Kong, Kowloon, Hong Kong SAR*
[3]*Department of Economics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR*

Key words and phrases: Endogeneity; high-dimensional model; instrumental variable; logistic regression; treatment effect; variable selection.

MSC 2020: 62F10; 62J07; 62P20.

Abstract: We develop a two-stage approach to estimate the treatment effects of dummy endogenous variables using high-dimensional instrumental variables (IVs). In the first stage, instead of using a conventional linear reduced-form regression to approximate the optimal instrument, we propose a penalized logistic reduced-form model to accommodate both the binary nature of the endogenous treatment variable and the high dimensionality of the IVs. In the second stage, we replace the original treatment variable with its estimated propensity score and run a least-squares regression to obtain a penalized logistic regression instrumental variables estimator (LIVE). We show theoretically that the proposed LIVE is root-n consistent with the true treatment effect and asymptotically normal. Monte Carlo simulations demonstrate that LIVE is more efficient than existing IV estimators for endogenous treatment effects. In applications, we use LIVE to investigate whether the Olympic Games facilitate the host nation's economic growth and whether home visits from teachers enhance students' academic performance. In addition, the R functions for the proposed algorithms have been developed in an R package naivereg. *The Canadian Journal of Statistics* 50: 795–819; 2022 © 2021 Statistical Society of Canada

Résumé: de traitment de variables endogènes factices, les auteurs de ce travail élaborent une approche en deux étapes qui fait usage de variables instrumentales (IV) de grandes dimensions. En effet, afin d'accommoder le caractère dichotomique de la variable de traitement endogène et le fait que les variables instrumentales soient de grandes dimensions, la procédure commence par approximer l'instrument optimal en utilisant la forme réduite d'une régression logistique pénalisée. Ensuite, en remplaçant la variable de traitement originale par son score de propension, un nouvel estimateur est construit grâce à l'ajustement par moindres carrés d'un modèle de régression logistique pénalisé à variables instrumentales (LIVE). Les auteurs montrent que l'estimateur LIVE, ainsi construit, est asymptitquement normal et converge vers le véritable effet de traitement au taux habituel de racine de n lorsque $n \to \infty$. Une étude Monte Carlo leur a, également, permis de constater que pour estimer les effets d'un traitement endogène, LIVE est plus efficace que d'autres estimateurs existants. En guise d'applications, les auteurs ont utilisé LIVE pour examiner l'effet des Jeux Olympiques sur la croissance économique du pays hôte et pour vérifier si les visites à domicile des enseignants améliorent les résultats scolaires des élèves. Enfin, les algorithmes et procédures dévélopés dans le cadre de ce travail ont été réunis sous forme d'un package R, appelé 'naivereg'. *La revue canadienne de statistique* 50: 795–819; 2022 © 2021 Société statistique du Canada

[*] *Corresponding author: michaelqfan@gmail.com*

1. INTRODUCTION

This article focuses on regression models with a dummy endogenous variable that takes on binary values of 1 or 0 to indicate the presence of some treatment effect that may be expected to affect the outcome. Endogenous treatments are commonly encountered in program evaluations using observational data for which the selection-on-observables assumption does not hold. Endogeneity induces unwelcome estimation bias and inconsistency issues, which could generate misguided policy implications. The instrumental variable (IV) method is a popular technique for addressing endogeneity problems. Angrist & Imbens (1995) introduced the two-stage least-squares (TSLS) estimator for average treatment effects (ATEs). Das (2005) considered a nonparametric version of TSLS in the case of a discrete endogenous regressor. Cai et al. (2006) studied functional coefficient IV models. Wooldridge (2014) proposed a control function approach for discrete endogenous explanatory variables. However, we find that the traditional TSLS-based estimators are no longer efficient and have large variances when estimating the treatment effects of dummy endogenous variables (see the simulation studies in Section 4). This is because the resulting predicted value of the dummy endogenous treatment variable using the optimal instrument and based on a linear reduced-form model could be outside the range $[0, 1]$. Wooldridge (2010) suggested using a probit model to estimate the propensity score for treatment assignment but did not study its theoretical properties. Heckman (1978) proposed estimating the inverse Mill's ratio with a probit model for dummy endogenous variables in the Heckman correction model.

In practice, a large set of potential IVs, including their functional transformations and interactions, may be introduced into the reduced-form model to approximate the optimal instrument and improve the precision of the IV estimators. However, if many irrelevant instruments are included in the reduced-form model, the resulting IV estimator becomes less efficient (Donald & Newey, 2001). The optimal instrument is the conditional expectation of the endogenous variable, given the valid IVs, that minimizes the asymptotic variance of the IV estimator (Amemiya, 1974). However, it is often not known a priori which IVs in the model are truly valid. The researcher cannot identify which instruments are weak using the rule-of-thumb F-statistic (Hansen & Kozbur, 2014). A desirable approach would be to utilize strong IVs while discarding irrelevant IVs to improve asymptotic efficiency as well as accuracy of the finite sample estimation. The early origins of high-dimensional IV models can be traced back to Kloek & Mennes (1960), who proposed using principal components as a dimension reduction device in the first stage, and to Newey (1990), who suggested the nonparametric series approach to estimate the optimal instrument. Of particular interest is the recent seminal work of Belloni et al. (2012), who proposed a post-LASSO approach to approximate the optimal instruments using high-dimensional linear IV models. Farrell (2015) studied treatment effects with possibly more covariates than observations. Fan & Zhong (2018) proposed a nonparametric additive IV estimator with adaptive group LASSO. Other related studies include Bai & Ng (2010), Carrasco (2012), Fan & Liao (2014), Caner & Fan (2015), Lin, Feng & Li (2015), Kang et al. (2016), Gautier & Tsybakov (2018), Windmeijer et al. (2019) and Zhong et al. (2021). In the statistical literature, in addition to LASSO (Tibshirani, 1996), many other regularization methods with attractive statistical properties have been introduced to address high-dimensional data, such as smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), group LASSO (Yuan & Lin, 2006), adaptive LASSO (Zou, 2006), adaptive group LASSO (Huang, Horowitz & Wei, 2010) and the Dantzig selector (Candes & Tao, 2007).

In this article, we develop an efficient two-stage estimation method for the treatment effects of dummy endogenous variables to accommodate both the binary nature of the endogenous variable and the high dimensionality of the IVs. In the first stage, we propose the use of a logistic regression reduced-form model for the dummy endogenous treatment variable to estimate the optimal instrument. In the high-dimensional IV case, a penalized logistic reduced-form model is considered for the selection of relevant IVs. In the second stage, we replace the original treatment variable with its estimated propensity score and run a least-squares regression to obtain the penalized logistic regression instrumental variables estimator (LIVE). We summarize the main contributions of LIVE as follows. First, we consider the probabilistic nature of the treatment effects of dummy endogenous variables by estimating the propensity score function. The better approximation of the optimal instrument using logistic regression with high-dimensional IVs produces more efficient IV estimators than the conventional TSLS-based estimators. Simulation results show that LIVE with the SCAD penalty is more efficient than other IV methods, including post-LASSO TSLS (Belloni et al., 2012). Second, we demonstrate theoretically that LIVE is root-n consistent and asymptotically normal. Third, we develop an R function, *LIVE*, in our R package naivereg[1] for empirical researchers to implement the method easily. Therefore, the proposed LIVE-based method should be useful for tackling the problem of treatment effect estimation for dummy endogenous variables with a large number of potential IVs.

The rest of the article is organized as follows. Section 2 describes the methodology and presents the penalized LIVE. Section 3 presents the theoretical results. In Section 4, we conduct simulation studies to assess the finite-sample performance. Section 5 demonstrates the use of the proposed methods in real data studies. Section 6 concludes the article. Technical proofs are contained in the Appendix.

2. METHODOLOGY

We consider a linear structural equation with an endogenous treatment variable

$$y_i = D_i\beta_0 + \mathbf{x}_i'\boldsymbol{\theta}_0 + \varepsilon_i, \tag{1}$$

where y_i is the response variable for individual i, D_i is a dummy endogenous treatment variable, β_0 denotes the true coefficient on D_i, $\mathbf{x}_i$ is an $m \times 1$ vector of other exogenous control variables, $\boldsymbol{\theta}_0$ is an $m \times 1$ vector of the true parameters associated with $\mathbf{x}_i$ and ε_i is the ith random error term for $i = 1, 2, \ldots, n$, where n is the sample size. We consider the case in which D_i is endogenous such that $\mathrm{E}\left(\varepsilon_i|D_i\right) \neq 0$, which leads to inconsistency in ordinary least-squares estimators. We focus on the binary treatment case, i.e., $D_i = 1$ when the ith individual belongs to the treatment group and $D_i = 0$ when it belongs to the control group. Under the assumption $\mathbb{E}[\varepsilon_i|D_i = 1, \mathbf{x}_i] = \mathbb{E}[\varepsilon_i|D_i = 0, \mathbf{x}_i]$, which is weaker than the conventional unconfoundedness assumption, we have $\beta_0 = \mathbb{E}[y_i|D_i = 1, \mathbf{x}_i] - \mathbb{E}[y_i|D_i = 0, \mathbf{x}_i]$, which is the ATE conditional on the value of $\mathbf{x}_i$. It is worth noting that when the assumption $\mathbb{E}[\varepsilon_i|D_i = 1, \mathbf{x}_i] = \mathbb{E}[\varepsilon_i|D_i = 0, \mathbf{x}_i]$ does not hold, it is also of interest to estimate β_0 as the true regression coefficient for the binary endogenous regressor in the structural equation. Our main goal is to estimate β_0 consistently using observational data and study its theoretical properties.

IV techniques are commonly used to obtain a consistent estimator for the regression coefficients of endogenous variables. We assume that there is a $p_n \times 1$ vector of IVs, denoted by $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip_n})'$, which are exogenous, i.e., $\mathrm{E}\left(\varepsilon_i|\mathbf{z}_i\right) = 0$ for all $i = 1, \ldots, n$, and correlated with the endogenous treatment variable. The optimal instrument is the conditional expectation of the endogenous variable, given the valid IVs, that minimizes the asymptotic variance of the IV estimator (Amemiya, 1974; Newey, 1990). Since the endogenous treatment variable in (1) is $D_i \in \{0, 1\}$, the optimal instrument is $\mathrm{E}(D_i|\mathbf{z}_i) = \mathrm{P}(D_i = 1|\mathbf{z}_i)$, which is the propensity score for treatment assignment given $\mathbf{z}_i$. However, the conventional TSLS estimator uses the first-stage linear reduced-form model of D_i against $\mathbf{z}_i$ to approximate the optimal instrument. The resulting predicted value of D_i could be outside the range of the optimal instrument, $[0, 1]$. As a result, the traditional TSLS estimator is no longer efficient and has large variances (refer to the simulation

[1] https://cran.r-project.org/web/packages/naivereg/index.html

studies in Section 4). To produce a better approximation of the optimal instrument, we consider a logistic regression as the reduced-form model used to estimate the propensity score in the first stage of the IV estimation. The resulting IV estimator is expected to be more efficient than the linear TSLS estimator. This conjecture is confirmed in later sections.

Without loss of generality, we include all exogenous explanatory variables $\mathbf{x}_i$ as IVs in the vector $\mathbf{z}_i$ in this section. We assume that the logistic regression reduced-form model estimates the relationship between the dummy endogenous treatment variable D_i and IVs $\mathbf{z}_i$, i.e.

$$\operatorname{logit}(p(\mathbf{z}_i, \boldsymbol{\gamma}_n)) = \log\left(\frac{p(\mathbf{z}_i, \boldsymbol{\gamma}_n)}{1 - p(\mathbf{z}_i, \boldsymbol{\gamma}_n)}\right) = \gamma_{n0} + \sum_{j=1}^{p_n} \gamma_{nj} z_{ij}, \tag{2}$$

where we denote the optimal instrument as $p(\mathbf{z}_i, \boldsymbol{\gamma}_n) = \mathrm{P}(D_i = 1|\mathbf{z}_i)$, $i = 1, \dots, n$. We further denote $\tilde{\mathbf{z}}_i = (1, \mathbf{z}_i')'$. Model (2) can also be represented as

$$p(\mathbf{z}_i, \boldsymbol{\gamma}_n) = \frac{\exp(\gamma_{n0} + \sum_{j=1}^{p_n} \gamma_{nj} z_{ij})}{1 + \exp(\gamma_{n0} + \sum_{j=1}^{p_n} \gamma_{nj} z_{ij})} =: \frac{\exp(\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i)}{1 + \exp(\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i)}. \tag{3}$$

The motivation behind the high-dimensional IV method is natural: when the ignorability assumption is violated, we need to search for valid IVs from a large pool of potential IVs. As mentioned in Belloni et al. (2012), it is necessary to consider many candidate IVs $\mathbf{z}_i$ such that we have either a large set of original instruments or many constructed polynomials and interactions of original instruments to substantially improve the precision of the IV estimators. However, irrelevant instruments in the reduced-form model could lead to finite sample bias and asymptotic inefficiency of the IV estimator. Regularization methods are commonly used to select relevant instruments and construct parsimonious predictive reduced-form models that can achieve better approximations of the optimal instruments. For example, Belloni et al. (2012) considered the post-LASSO method for estimating the first-stage linear reduced-form model with high-dimensional instruments and their series. The LASSO-based IV estimator is demonstrated to be root-n consistent and asymptotically normal under the approximate sparsity assumption. Here, approximate sparsity means that the conditional expectation of the endogenous variables, given many instruments, can be approximated well by a relatively small set of instruments.

To accommodate the binary nature of the dummy endogenous treatment variable, we propose a penalized logistic regression reduced-form model to select instruments and approximate the optimal instrument. The penalized likelihood function for the logistic regression reduced-form model is

$$\begin{aligned} Q_{n1}(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n, \lambda_{n1}) &= \sum_{i=1}^{n} \log f_n(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n) - n \sum_{j=0}^{p_n} p_{\lambda_{n1}}(|\gamma_{nj}|), \\ &= \sum_{i=1}^{n} \left[D_i \boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i - \log\left(1 + e^{\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i}\right)\right] - n \sum_{j=0}^{p_n} p_{\lambda_{n1}}(|\gamma_{nj}|), \end{aligned} \tag{4}$$

where $f_n(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n)$ is the probability density function of the logistic distribution, and $p_{\lambda_{n1}}(\cdot)$ is the penalty function. We consider the SCAD penalty (Fan & Li, 2001), the first derivative of which satisfies

$$p'_{\lambda_{n1}}(\gamma) = \lambda_{n1}\left[\mathbb{1}(\gamma \leq \lambda_{n1}) + \frac{(a\lambda_{n1} - \gamma)_+}{(a-1)\lambda_{n1}} \mathbb{1}(\gamma > \lambda_{n1})\right] \tag{5}$$

for $\gamma > 0$, where $a > 2$ is a pre-specified constant, $\mathbb{1}(\cdot)$ is the indicator function, and λ_{n1} is the tuning parameter to control the selected model size. In practice, Fan & Li (2001) suggested that $a = 3.7$ and λ_{n1} can be chosen with the cross-validation (CV) method or the Bayesian information criterion (Wang, Li & Tsai, 2007). The penalized likelihood estimator $\widehat{\boldsymbol{\gamma}}_n$ is obtained by maximizing the objective function in (4). That is

$$\widehat{\boldsymbol{\gamma}}_n = (\widehat{\gamma}_{n0}, \widehat{\gamma}_{n1}, \ldots, \widehat{\gamma}_{np_n})' = \underset{\boldsymbol{\gamma}_n}{\operatorname{argmax}}\, Q_{n1}(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n, \lambda_{n1}). \tag{6}$$

It implies that the conditional probability $p(\mathbf{z}_i, \boldsymbol{\gamma}_n)$ can be estimated accordingly by $\widehat{p}(\mathbf{z}_i, \widehat{\boldsymbol{\gamma}}_n)$. For ease of notation, we denote the optimal instrument as $D_i^* = \mathrm{P}(D_i = 1|\mathbf{z}_i) = p(\mathbf{z}_i, \boldsymbol{\gamma}^*)$, where $\boldsymbol{\gamma}^* = (\gamma_0^*, \gamma_1^*, \ldots, \gamma_{p_n}^*)'$ is the underlying true parameter in (2), and $\widehat{D}_i^* = \widehat{p}(\mathbf{z}_i, \widehat{\boldsymbol{\gamma}}_n)$ is the estimator of D_i^*. Without loss of generality, suppose that the first q_n elements of the true parameter vector $\boldsymbol{\gamma}^*$ are nonzero and the remaining $p_n - q_n$ elements are zero.

Next, we illustrate how to estimate the true endogenous treatment effect β_0. By taking the conditional expectation of both sides of (1), given IVs $\mathbf{z}_i$, we have

$$E(y_i|\mathbf{z}_i) = D_i^*\beta_0 + \mathbf{x}_i'\boldsymbol{\theta}_0, \tag{7}$$

where we note that $E(\mathbf{x}_i|\mathbf{z}_i) = \mathbf{x}_i$ because all exogenous variables $\mathbf{x}_i$ are included in the IVs. Adding $v_i = y_i - \mathrm{E}(y_i|\mathbf{z}_i)$ to both sides of (7) implies that

$$y_i = D_i^*\beta_0 + \mathbf{x}_i'\boldsymbol{\theta}_0 + v_i. \tag{8}$$

It is straightforward to show that $\mathrm{E}(v_i) = \mathrm{E}[\mathrm{E}(v_i|\mathbf{z}_i)] = 0$ and $\mathrm{cov}(D_i^*, v_i) = \mathrm{E}(D_i^* v_i) = \mathrm{E}[D_i^*\,\mathrm{E}(v_i|\mathbf{z}_i)] = 0$. Thus, v_i can be regarded as the independent random error with mean zero in the linear model (8), and D_i^* is an exogenous variable. It is worth noting that the coefficient on the optimal instrument D_i^* in the model (8) remains the same as that in the structural equation (1). If D_i^* is known, the treatment effect β_0 can be easily obtained through ordinary least squares. However, the optimal instrument D_i^* is unobservable. In practice, we replace D_i^* with its estimator $\widehat{D}_i^* = \widehat{p}(\mathbf{z}_i, \widehat{\boldsymbol{\gamma}}_n)$ and run a simple linear regression to obtain the final IV estimator for β_0. In particular, let $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{D} = (D_1, \ldots, D_n)'$, $\mathbf{D}^* = (D_1^*, \ldots, D_n^*)'$, $\widehat{\mathbf{D}}^* = (\widehat{D}_1^*, \ldots, \widehat{D}_n^*)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and $\boldsymbol{v} = (v_1, \ldots, v_n)'$. Let $\mathcal{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{Q} = \mathbf{I}_n - \mathcal{P}_{\mathbf{X}}$. The resulting IV estimator for β_0 takes the form

$$\widehat{\beta} = \left(\widehat{\mathbf{D}}^{*\prime}\mathbf{Q}\widehat{\mathbf{D}}^*\right)^{-1}\left(\widehat{\mathbf{D}}^{*\prime}\mathbf{Q}\mathbf{Y}\right). \tag{9}$$

We call $\widehat{\beta}$ the penalized LIVE. In summary, we provide the following algorithm to obtain LIVE for β_0.

Algorithm 1 Penalized logistic regression instrumental variable estimation (LIVE)

```
Step 1. Obtain the penalized likelihood estimator $\widehat{\boldsymbol{\gamma}}_n$ by maximizing $Q_{n1}(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n, \lambda_{n1})$ in (4), employing the cross-validation method to choose $\lambda_{n1}$.
Step 2. Estimate the conditional expectation of the endogenous treatment effect $\widehat{D}_i^* = \widehat{p}(\mathbf{z}_i, \widehat{\boldsymbol{\gamma}}_n)$ for $i = 1, \ldots, n$ according to (3).
Step 3. Compute the penalized LIVE for $\beta_0$, $\widehat{\beta} = \left(\widehat{\mathbf{D}}^{*\prime}\mathbf{Q}\widehat{\mathbf{D}}^*\right)^{-1}\left(\widehat{\mathbf{D}}^{*\prime}\mathbf{Q}\mathbf{Y}\right)$, in (9).
```

3. THEORETICAL RESULTS

In this section, we study the theoretical properties of LIVE with the SCAD penalty. We show that it is root-n consistent and asymptotically normal. Throughout the article, $\| \cdot \|_0$, $\| \cdot \|$ and $\| \cdot \|_\infty$ denote the ℓ_0-norm, the ℓ_2-norm and the infinity norm, respectively.

First, we assume the following regularity conditions:

(A) $\max_{0 \le j \le p_n} \{p'_{\lambda_n}(|\gamma^*_{nj}|), \gamma^*_{nj} \neq 0\} = O(n^{-1/2})$ and $\max_{0 \le j \le p_n} \{p''_{\lambda_n}|\gamma^*_{nj}|, \gamma^*_{nj} \neq 0\} \to 0$ hold as $n \to \infty$.
(B) There are constants a and b such that $\left|p''_{\lambda_{n1}}(c_1) - p''_{\lambda_{n1}}(c_2)\right| \le b\left|c_1 - c_2\right|$ for $c_1, c_2 > a\lambda_{n1}$.
(C) $\left\{\mathbf{z}_i\right\}_{i=1}^n$ are i.i.d., mean-zero, and bounded random vectors.
(D) There exists a constant $c_4 > 0$ such that $\lambda_{\min}\left(\mathbf{X}'\mathbf{X}\right)/n \ge c_3$, where $\lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalues of a matrix $\mathbf{A}$.
(E) The second moment for v_i exists, i.e., $\mathrm{E}(v_i^2) < \infty$.

Conditions (A) and (B) are imposed on the penalty function proposed by Fan & Peng (2004), which guarantee Lemma 1. Condition (C) is a special case of condition SE(P) in Belloni, Chernozhukov & Hansen (2014). Condition (D) is a mild condition derived from Zhang & Huang (2008). Condition (E) imposes the existence of a second moment for the error term v_i.

Lemma 1. *Under conditions (A) and (B) and (A1)–(C1) in the Appendix, if $p_n^4/n \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\gamma}}_n$ for $Q_n(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n, \lambda_{n1})$ such that*

$$\|\hat{p}(\mathbf{z}_i, \hat{\boldsymbol{\gamma}}_n) - p(\mathbf{z}_i, \boldsymbol{\gamma}^*_n)\| = O_p(\sqrt{p_n^2/n}). \tag{10}$$

It is worth noting that Fan & Peng (2004) proved that $\hat{\boldsymbol{\gamma}}_n$ is the root-n/p_n-consistent estimator in Theorem 1 if conditions (A) and (B) and (A1)–(C1) in the Appendix hold as $n \to \infty$, that is, if $\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^*_n\| = O_p(\sqrt{p_n/n})$ when $p_n^4/n \to 0$. Based on this seminal work, Lemma 1 establishes the consistency of the estimated propensity score of the dummy endogenous treatment variable given the IVs. In practice, if the dimension of the IVs p_n is much larger than n, one might apply certain screening methods (Fan & Song, 2010; Mai & Zou, 2013) to reduce the dimensionality before applying the regularization methods.

Next, we establish the asymptotic properties of the proposed LIVE in the following theorem.

Theorem 1. *Suppose conditions (A)–(E) hold; then, LIVE with the SCAD penalty is root-n consistent and asymptotically normal. That is*

$$\sigma_n^{-1}\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, 1), \tag{11}$$

where $\sigma_n^2 = (\mathrm{E}(\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^*))^{-1}\,\mathrm{E}(\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^* v_i^2)(\mathrm{E}(\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^*))^{-1}$. If $\mathrm{E}(v_i^2) = \sigma_v^2$ almost surely for all $1 \le i \le n$, then $\sigma_n^2 = (\mathrm{E}(\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^*))^{-1}\sigma_v^2$.

Theorem 1 demonstrates that the proposed LIVE with the SCAD penalty is root-n consistent and asymptotically normal. This result is analogous to that in Theorem 3 in Belloni et al. (2012) and Theorem 3.1 in Fan & Zhong (2018). For the purposes of statistical inference, the asymptotic variance can be estimated with the plug-in method, and the corresponding asymptotic confidence interval for the true dummy treatment effect β_0 can be obtained by $(\hat{\beta} - z_{\tau/2}\hat{\sigma}_n/\sqrt{n}, \hat{\beta} + z_{\tau/2}\hat{\sigma}_n/\sqrt{n})$, where $z_{\tau/2}$ denotes the $\tau/2$ upper-tailed critical value of the standard normal distribution.

4. SIMULATIONS

In this section, we study the finite-sample performance of LIVE in both low-dimensional and high-dimensional IV cases.

Example 1. We first consider a simple structural equation with a dummy endogenous treatment variable only:

$$y_i = D_i\beta_0 + \varepsilon_i,$$

where we set the true treatment effect $\beta_0 = 1$. The dummy endogenous variable D_i is generated by the Bernoulli distribution with probability $\exp(\eta_i)/(1 + \exp(\eta_i))$, where $\eta_i = 0.6z_{i1} + 0.8z_{i2} + z_{i3} + z_{i4} + \xi_i$ and $\mathbf{z}_i = \{z_{ij}\}$ is randomly generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = (\rho_{j_1j_2})_{p\times p}$, $\rho_{j_1j_2} = 0.5^{|j_1-j_2|}$ for $j_1, j_2 = 1, \dots, p$, where p is specified in the following discussion. Furthermore, we generate the error terms $(\varepsilon_i, \xi_i)'$ in both the structural model and the reduced-form model from a bivariate normal distribution with mean zero and covariance matrix $\Sigma_{\varepsilon,\xi}$. To study the endogeneity of the treatment variable D_i, we let

$$(\varepsilon_i, \xi_i) \sim N(0, \Sigma_{\varepsilon,\xi}), \quad \text{with} \quad \Sigma_{\varepsilon,\xi} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

Example 2. In this example, we consider a structural equation with both an endogenous treatment variable and m exogenous control variables,

$$y_i = D_i\beta_0 + \mathbf{x}_i'\boldsymbol{\theta}_0 + \varepsilon_i.$$

We set $\beta_0 = 1$, and the endogenous treatment variable D_i is generated by the Bernoulli distribution with probability $\exp(\eta_i)/(1 + \exp(\eta_i))$, where $\eta_i = 0.5z_{i1} + 0.4z_{i3} + 0.9z_{i4} + 0.8z_{i5} + \xi_i$. The data generating processes for $\mathbf{z}_i$ and the two random errors $(\varepsilon_i, \xi_i)'$ are the same as in Example 1. Since $\mathbf{x}_i$ are assumed to be exogenous, without loss of generality, we set the first m columns of $\mathbf{z}_i$ as $\mathbf{x}_i$; here, $m = 2$ and $\boldsymbol{\theta}_0 = (0.1, 0.2)'$.

To study the finite sample performance of the estimators, we set the sample size $n = \{100, 200, 500, 1000\}$ and consider two cases for the number of potential instruments: $p = 5$ and $p = 200$. For the low-dimensional instrument case where $p = 5$, we include all five instruments in the reduced-form model and compare only LIVE without any penalty (denoted by "LIVE") with the conventional OLS and TSLS estimators for β_0. For the high-dimensional instrument case where $p = 200$, regularization methods are used to select the relevant instruments for estimating the endogenous effect. Then, we estimate the logistic reduced-form models based on the selected set of instruments by LASSO or SCAD (denoted by "LIVE-LASSO" and "LIVE-SCAD," respectively). The default 10-fold CV method is employed to find the optimal tuning parameter λ_{n1} for both LASSO and SCAD. The proposed methods are implemented with the R function *LIVE* in the R function package naivereg that we developed. We compare two existing methods: OLS and TSLS with LASSO (denoted by "TSLS-LASSO"). To evaluate the finite sample performance of each IV estimation method, we compute the average of the estimated bias (denoted as "Bias"), $N^{-1}\sum_{k=1}^{N}(\widehat{\beta}_k - \beta_0)$, with its empirical standard deviation and the estimated mean-squared errors (denoted as "MSE"), $N^{-1}\sum_{k=1}^{N}(\widehat{\beta}_k - \beta_0)^2$, where $\widehat{\beta}_k$ denotes the IV estimator of the true treatment effect β_0 in the kth simulation. All simulation results are based on 1000 repetitions. We summarize the results in Tables 1–4. Figure 1 also shows the boxplots of the different estimators for β_0.

The simulation results show that the OLS estimators are seriously biased and have large MSE values because of the endogeneity problem. In the low-dimensional IV case without

TABLE 1: Average biases with standard errors (in parentheses) and the mean-squared errors (MSEs) of OLS and the different instrumental variables estimators for the endogenous treatment effect ($\beta_0 = 1$) for $p = 5$ from Example 1.

n	Method	OLS	TSLS	LIVE
$n = 100$	Bias	0.2688 (0.0886)	0.0423 (0.0731)	0.0301 (0.1679)
	MSE	0.1280	0.2671	0.1509
$n = 200$	Bias	0.2690 (0.0799)	0.0254 (0.0358)	0.0159 (0.0115)
	MSE	0.0867	0.1876	0.1060
$n = 500$	Bias	0.2699 (0.0760)	0.0132 (0.0146)	0.0079 (0.0049)
	MSE	0.0564	0.1202	0.0699
$n = 1000$	Bias	0.2703 (0.0746)	0.0066 (0.0076)	0.0031 (0.0022)
	MSE	0.0388	0.0869	0.0469

TABLE 2: Average biases with standard errors (in parentheses) and the mean-squared errors (MSEs) of OLS and the different instrumental variables estimators for the endogenous treatment effect ($\beta_0 = 1$) for $p = 5$ from Example 2.

n	Method	OLS	TSLS	LIVE
$n = 100$	Bias	0.3624 (0.1502)	0.0172 (0.1021)	0.0250 (0.0266)
	MSE	0.1374	0.3193	0.1611
$n = 200$	Bias	0.3664 (0.1440)	0.0168 (0.0596)	0.0139 (0.0140)
	MSE	0.0989	0.2437	0.1177
$n = 500$	Bias	0.3670 (0.1388)	0.0049 (0.0215)	0.0058 (0.0056)
	MSE	0.0639	0.1466	0.0749
$n = 1000$	Bias	0.3667 (0.1363)	0.0076 (0.0117)	0.0038 (0.0026)
	MSE	0.0434	0.1079	0.0510

selecting IVs, the conventional TSLS approach reduces the bias and produces a consistent estimator, but the MSEs are substantially larger than those of the proposed LIVE. This is because the predicted value of the dummy treatment variable in the linear reduced-form model could be outside the range $[0, 1]$. Because logistic regression is able to capture the relationship between the dummy endogenous variable and the IVs and because of the probabilistic nature of the optimal instrument, it can provide better predictions of the dummy endogenous treatment variable. Tables 1 and 2 show that our proposed LIVE performs significantly better with a much smaller bias and MSE in the low-dimensional IV setting. In the high-dimensional instrument case, shown in Tables 3 and 4, LIVE-based estimators are significantly more efficient than the other estimators. The advantage of LIVE-LASSO over TSLS-LASSO shows that it is necessary to consider the logistic regression reduced-form model instead of the linear model for dummy endogenous treatment variables. The LIVE-SCAD estimator is slightly better than LIVE-LASSO because LASSO generally tends to select more irrelevant instruments than SCAD. As the sample size n increases, the biases and the MSEs of the proposed LIVE

TABLE 3: Average biases with standard errors (in parentheses) and the mean-squared errors (MSEs) of OLS and the different instrumental variables estimators for the endogenous treatment effect ($\beta_0 = 1$) for $p = 200$ from Example 1.

n	Method	OLS	TSLS-LASSO	LIVE-LASSO	LIVE-SCAD
$n = 100$	Bias	0.2679 (0.0942)	0.2455 (0.1166)	0.2172 (0.0709)	0.1292 (0.0459)
	MSE	0.1324	0.2375	0.1541	0.1712
$n = 200$	Bias	0.2691 (0.0822)	0.1669 (0.0632)	0.1355 (0.0342)	0.0754 (0.0185)
	MSE	0.0989	0.1882	0.1258	0.1134
$n = 500$	Bias	0.2698 (0.0766)	0.0859 (0.0226)	0.0606 (0.0096)	0.0221 (0.0058)
	MSE	0.0617	0.1235	0.0773	0.0730
$n = 1000$	Bias	0.2686 (0.0740)	0.0423 (0.0095)	0.0278 (0.0035)	0.0046 (0.0025)
	MSE	0.0437	0.0878	0.0524	0.0503

TABLE 4: Average biases with standard errors (in parentheses) and the mean-squared errors (MSEs) of OLS and the different instrumental variables estimators for the endogenous treatment effect ($\beta_0 = 1$) for $p = 200$ from Example 2.

n	Method	OLS	TSLS-LASSO	LIVE-LASSO	LIVE-SCAD
$n = 100$	Bias	0.2748 (0.0932)	0.2436 (0.1168)	0.2130 (0.0708)	0.1170 (0.0416)
	MSE	0.1331	0.2399	0.1596	0.1670
$n = 200$	Bias	0.2712 (0.0831)	0.1699 (0.0655)	0.1440 (0.0370)	0.0803 (0.0208)
	MSE	0.0976	0.1916	0.1319	0.1197
$n = 500$	Bias	0.2708 (0.0771)	0.0843 (0.0211)	0.0627 (0.0098)	0.0265 (0.0059)
	MSE	0.0610	0.1184	0.0767	0.0721
$n = 1000$	Bias	0.2664 (0.0729)	0.0446 (0.0101)	0.0275 (0.0034)	0.0046 (0.0025)
	MSE	0.0434	0.0899	0.0519	0.0497

estimators shrink, which validates its consistency in Theorem 3.1. Overall, the simulation results demonstrate that our proposed LIVE with a SCAD penalty is necessarily useful in estimating binary endogenous treatment effects, especially when many potential instruments are considered.

5. APPLICATIONS

5.1. Application to the Olympics

The Olympic Games are the world's largest sports event. On one hand, hosting mega-events such as the Olympics might boost the host country's economic development through broadcasting and ticket revenue, commercial sponsorships, tourism, enhanced public infrastructure, new job creation, and increased exports and foreign investment. On the other hand, the positive impact of hosting these sports extravaganzas on economic growth is challenged by many in both academia and the general public. Zimbalist (2015) laid out the economic

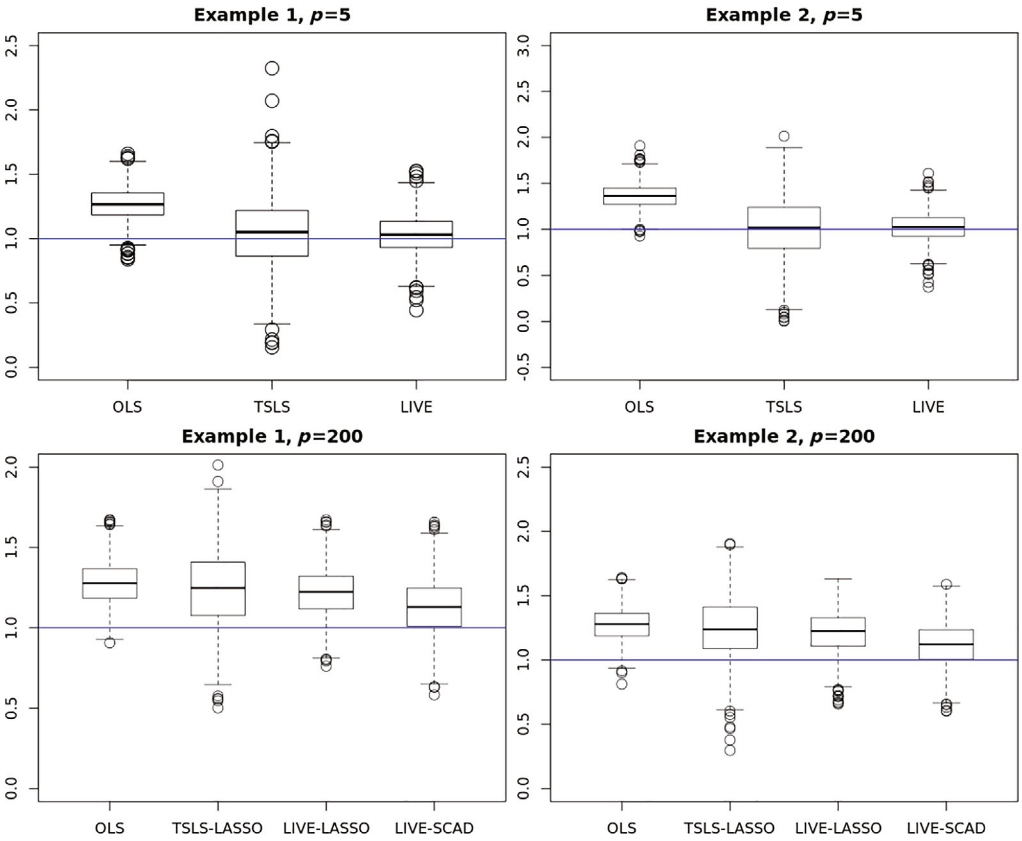

FIGURE 1: Boxplots of OLS and IV estimators for the endogenous treatment effect ($\beta_0 = 1$) using low-dimensional ($p = 5$) and high-dimensional ($p = 200$) IVs when $n = 100$ in Examples 1 and 2.

case against hosting mega sports events. Critiques often cite the extreme costs and lingering local taxes, overshooting of the government budget, unjust welfare distribution, and small foreseeable income (Owen, 2005; Coates, 2007), among other negative effects. Whether the Games have a positive or negative impact on the economy is still a hot topic in public debates.

In the literature, Rose & Spiegel (2011) found that hosting the Olympic Games has a positive effect on the economy through export channels. Furthermore, they argued that the act of bidding itself, regardless of whether it is eventually successful or not, has a similar economic impact. Bruckner & Pappa (2015) argued that bidding for the Olympics represents a news shock that predicts increases in future government investment. It is the anticipation effect that could induce positive output, investment, consumption, etc. However, the aforementioned empirical studies do not consider the endogeneity of the treatment variable (Baade & Matheson, 2016). One should take into account the fact that host cities are by no means randomly chosen by the International Olympic Committee; rather, only cities with bright prospects for the future are considered.

Specifically, we consider the following benchmark cross-sectional regression model:

$$y_i = \alpha + \beta Oly_i + \gamma' X_i + \epsilon_i,$$

where y_i is the output variable, which we consider to be GDP per capita[2] in 2010. Oly_i is the dummy variable indicating whether country i has ever hosted the summer Olympics post World War II. β captures the marginal effect of hosting mega-events, conditional on the cross-country heterogeneity in factors of production. We also use the winter Olympics as a robustness check (as an alternative event study). Moreover, we use a dummy variable for summer Olympics bidding to evaluate the signalling effect of bidding, as discussed in Rose & Spiegel (2011). X_i is a vector of explanatory variables, including the capital stock, employment, and total factor production (TFP), which are employed to control for production frontier elements. Granted that a panel data model that explores the dynamic effects of hosting the Olympics would also be useful, the simple form of the model we use in this article provides a benchmark result for whether hosting the Olympic Games makes any difference to the output variables. We collected data from 186 countries/regions, of which a final sample of 163 was used in the regressions. The summary statistics and data source are presented in Table 5.

The ideal instruments here would be those that determine the likelihood of a city hosting the Olympics but are uncorrelated with the unobserved factors in the structural equation. For those reasons, we consider the host city's geographic variables, including the area of the host city (the capital city if the country has never hosted), its elevation, and its meteorological variables (precipitation, temperature, etc.). Since the Olympics often involve the resources of the whole country instead of just those of the host city, we also consider country-level IVs such as total land area, total water area, land boundaries, coastline, a dummy for being landlocked (landlocked = 1), elevation, arable land as a percentage of total land.[3] In total, we include 14 original IVs. These geographic variables satisfy the exogeneity condition for IVs because they are pre-determined and exogenous to the economic variables. However, it is not clear to practitioners which of those variables are truly important in determining eventual Olympic host cities.

First, the OLS regression results are reported in Table 6. For the mega-events, we consider the summer Olympics, winter Olympics and summer Olympics bidders that did not receive the right to host the games.[4] The summer Olympics seem to be associated with higher GDP per capita, conditional on the logarithm of the capital stock and TFP, while other mega-events do not seem to have strong impacts. However, as we discussed, the OLS results are prone to bias as a result of the endogeneity of the dummy variable. Hence, we now turn to the proposed LIVE with SCAD.

The IV regression results are reported in Table 7. We can see that the magnitude of the estimated effect of hosting the summer Olympics is almost three times larger (and more statistically significant) than that of the OLS estimate, which implies that the OLS bias is large and tends to underestimate the effect of hosting the Olympics. One possible source of bias comes from underestimating the spillover effects to other cities that are not directly captured by the Olympic event variable itself. The winter Olympics LIVE estimate is almost twice as large in its effects as the OLS result and is significant at the 5% level. The effect of hosting the summer Olympics is much stronger than that of hosting the winter Olympics. Contrary to the studies of Rose & Spiegel (2011) and Bruckner & Pappa (2015), we find that the bidding signal effect is not strong. Cities that won the bid are put in the spotlight for 7–8 years before the game starts,

[2] As a robustness check, we also consider other outcome variables, such as private consumption, investment, government expenditures, consumer price level, nominal exchange rate, and openness to trade. The results are available upon request to the authors.

[3] In Bazzi & Clemens (2013), the validity of some common IVs (such as legal origin, population size, etc.) used in the growth literature is discussed. We did not use any of those instruments, which could possibly be invalid. Instead, we use only geographic and meteorological variables, which yield a P-value of 0.1567 (for the summer Olympics) and 0.2217 (for the winter Olympics) for Hansen's J-test.

[4] In our sample, there are 30 bidding countries, including 12 developing countries and 18 developed countries, according to the International Monetary Fund's World Economic Outlook Report in April 2015.

TABLE 5: Summary statistics.

	Mean	SD	Median	Min.	Max.	Sample size
Outcome variable						
GDP per capita	1.3E4	1.7E4	7.2E3	321.60	1.4E5	186
Treatment variable						
Olympics	0.08	0.26	0	0	1	186
Other controls						
Population	3.6E4	1.3E5	7.3E3	20.88	1.3E6	186
Consumption	2.0E11	8.4E11	1.6E10	1.1E8	1.0E13	186
Government expenditure	6.0E10	2.3E11	3.5E9	2.9E7	2.5E12	186
Investment	8.4E10	3.2E11	6.1E9	3.2E7	2.9E12	186
Capital stock	1.7E12	5.5E12	1.6E11	1.6E9	4.9E13	169
Labour	16.80	70.24	3.73	0.01	781.38	179
TFP	0.57	0.30	0.57	0.07	1.81	165
Openness	92.08	47.68	85.32	1.98	392.09	186
IV						
Area_city	1.0E3	3.0E3	239.65	0.7	2.5E4	186
Elevation_city	392.12	646.24	61.5	−28	3.6E3	186
Max_temp_city	83.14	10.35	84.36	50.61	115.59	186
Min_temp_city	65.47	11.26	66.34	32.88	85.75	186
Ave_temp_city	74.34	10.45	76.023	45.76	103.34	186
Precipitation_city	4.472	6.33	2.48	0	42.2	186
Land	6.8E5	1.9E6	1.3E5	54	1.6E7	186
Water	3.6E4	1.9E5	1.5E3	0	2.3E6	186
Land boundaries	2.8E3	3.5E3	1.8E3	0	2.2E4	186
Coastline	3.9E3	1.6E4	500	0	2.0E5	186
Landlocked	0.21	0.41	0	0	1	186
Elevation	2.7E3	2.0E3	12.5E3	2.4	8.9E3	186
% Arable land	14.43	13.21	10.20	0.02	57.99	186
% Permanent crops	4.01	7.01	1.15	0	44.44	186

Note: GDP and the national account variables are in USD, the population is in thousands, trade openness is in percentages, and the labour force is in millions of persons. The geographic variables of the host city (or capital city if the country never bid for the Olympics) have the subscript "_city"; otherwise, all variables are country-level variables. Area is in km^2, the coastline and land boundaries are measured in kilometres, elevation is in metres, temperature is in degrees Fahrenheit and precipitation is in inches. Source: Penn World Table 9.0, the World Bank, United Nations Industrial Development Organization, OECD National Accounts, Demographic Yearbook of United Nations Statistics Division, National Oceanic and Atmospheric Administration, International Olympic Committee and the World Factbook. Year of the sample: 2010.

TABLE 6: Results based on OLS.

	Summer Olympics	Winter Olympics	Summer bidding
Constant	5.303	5.048	5.003
	(0.431)	(0.402)	(0.389)
Mega-event	0.466*	0.253	0.153
	(0.269)	(0.325)	(0.206)
Capital	0.142***	0.168***	0.171***
	(0.038)	(0.035)	(0.034)
TFP	2.866***	2.811***	2.817***
	(0.244)	(0.248)	(0.247)
Sample size	163	163	163
R^2	0.626	0.621	0.621

Note: Standard errors are reported in parentheses. Significance levels of 0.1, 0.05 and 0.01 are noted with *, ** and ***, respectively. Intercept significance levels are not reported. The final sample size of 163 includes all countries with no missing observations.

TABLE 7: Results based on the penalized logistic regression instrumental variables estimator with smoothly clipped absolute deviation.

	Summer Olympics	Winter Olympics	Summer bidding
Constant	7.697	5.326	5.103
	(1.165)	(0.408)	(0.426)
Mega-event	1.332**	0.535**	0.523
	(0.674)	(0.263)	(0.595)
Capital	0.108***	0.149***	0.167***
	(0.010)	(0.040)	(0.039)
TFP	3.926***	2.633***	2.767***
	(0.718)	(0.260)	(0.258)
Sample size	163	163	163
R^2	0.622	0.632	0.621

Note: Standard errors are reported in parentheses. Significance levels of 0.1, 0.05 and 0.01 are noted with *, ** and ***, respectively. Intercept significance levels are not reported.

and there are frequent news updates about the progress of venues, preparations and the new events to be included in the Olympic Games.

5.2. Application to the Effect of Teachers' Home Visits

In this case study, we demonstrate the estimation of the treatment effect of teachers' home visits on students' academic performance. It is important to understand the influence of family on children's performance at school (Castro et al., 2015; Zhong et al., 2021). Unlike previous studies that mainly focus on income factors (Dohl & Lochner, 2012), we evaluate the interactions of teachers and parents on students' academic performance while controlling for a rich set of factors, including income. We use comprehensive survey data from the China Education Panel Survey to investigate the treatment effects of class advisers' home visits on student performance as measured by standardized exam grades.

The basic econometric model we consider is

$$score_i = \alpha_0 D_i + \mathbf{x}_i' \boldsymbol{\beta}_0 + \varepsilon_i,$$

where $score_i$ is the mathematics/Chinese/English score of the student; D_i is the treatment variable of receiving a home visit, for which $D_i = 1$ if the class adviser visited the student at home during that school year; $\mathbf{x}_i$ is a set of covariates that include the student's gender, his/her cognitive ability test score, his/her health status, his/her total hours of study after school, the mother's education, family income, the number of siblings, whether the student lives with his/her grandparents and the characteristics of the student's teacher for the subject in the response variable including age, gender and total teaching hours.

The decision to make a home visit is affected by certain common unobserved factors determining academic performance. For example, unobserved common factors such as "the student's real interest in the subject" could affect both the decision to make a home visit and the student's grades. Thus, endogeneity is an issue that is embedded in this study. We consider the exogenous variables of the non-adviser and of teachers of different subjects as instruments. Specifically, the candidate instruments include the age (of the non-adviser teacher of a different subject from that of the response variable, which is the same for the rest of the instruments), marital status, gender, the number of other classes being taught outside the sample class, total teaching hours last week, total teaching preparation hours last week, the number of hours spent grading homework last week and the number of minutes spent communicating with students after class. We also include the polynomial terms for the discrete and continuous variables up to order 3 and the interactions of all these variables. The total number of IVs is 210. To reduce the influence of outliers and potential endogenous IV concerns, we further trim the sample by deleting the lower 5th percentile and upper 95th percentile of the score data and instruments such as hours spent preparing lessons last week. The final sample size of our study is 7617. The percentage of observations with home visits is 42.93%.

The regression results are presented in Table 8. The OLS estimation in general reports no significant effect of home visits on any of the three subjects. It is well known that the OLS estimator is severely biased when there is endogeneity in the treatment variable. TSLS estimates show that for all three subjects, home visits can improve standardized exam grades by approximately 0.5 points (on a 100-point scale) on average, holding other factors constant.

TABLE 8: Estimated effects of home visit on school performance.

	Math		Chinese		English	
	Effect	Std. err.	Effect	Std. err.	Effect	Std. err.
OLS (univariate)	−0.150	0.198	−0.159	0.198	−0.379*	0.198
OLS (multivariate)	0.003	0.212	0.052	0.199	−0.055	0.201
TSLS	0.624**	0.297	0.556*	0.290	0.497*	0.292
TSLS-LASSO	0.730**	0.317	0.646*	0.355	0.538*	0.322
LIVE-SCAD	0.814***	0.306	0.805**	0.327	0.769**	0.318

Note: OLS (univariate) is the univariate OLS with only the treatment variable. OLS (multivariate) is the multivariate regression. Significance levels of 0.1, 0.05 and 0.01 are noted by *, ** and ***, respectively.

Compared to the IV method, it seems that OLS underestimates the effect of home visits. TSLS-LASSO results show a slightly larger treatment effect than the standard TSLS results for all three subjects. The proposed LIVE with SCAD provides the strongest evidence that home visits are able to improve students' performance in all three subjects.

6. EXTENSION TO HIGH-DIMENSIONAL CONTROLS

In empirical applications, it is often not clear which control variables to include in the structural equation when using a rich micro-dataset with many variables that could potentially affect the outcome variable. To estimate the treatment effect accurately, we are inclined to include as many confounding control covariates as possible in the model (1). That is, we now allow the dimensionality of $\mathbf{x}_i$, m, to be large and even to be greater than n. To reduce the omitted variable bias caused by high-dimensional control variables, we develop a double selection plus logistic regression IV estimator (DS-LIVE) for the dummy endogenous treatment effect parameter β_0 with both high-dimensional control variables and IVs. The DS-LIVE method can be considered a hybrid between the double selection (DS) method in Belloni, Chernozhukov & Hansen (2014) and our LIVE method.

For a better presentation of the new method, we separately write the IVs as two parts: the exogenous control variables $\mathbf{x}_i$ and the additional IVs $\tilde{\mathbf{z}}_i$. Then, the logistic reduced-form model (3) can be rewritten as

$$p(\mathbf{z}_i, \mathbf{x}_i; \boldsymbol{\gamma}_n, \boldsymbol{\omega}_n) =: \frac{\exp(\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i + \boldsymbol{\omega}_n' \mathbf{x}_i)}{1 + \exp(\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i + \boldsymbol{\omega}_n' \mathbf{x}_i)}. \tag{12}$$

Our proposed DS-LIVE algorithm proceeds with the following three steps.

In the first step, we select the relevant control variables that are helpful in predicting the outcome variable using regularization methods for the data $(y_i, \mathbf{x}_i)$. Specifically, we consider the penalized objective function as follows:

$$L_{n1}(y_i, \mathbf{x}_i; \boldsymbol{\theta}, \lambda_{n2}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\theta})^2 + \sum_{j=1}^{p} p_{\lambda_{n2}}(|\theta_{nj}|), \tag{13}$$

where λ_{n2} is a tuning parameter controlling model complexity. The penalized estimator $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the objective function $L_{n1}(y_i, \mathbf{x}_i; \boldsymbol{\theta}, \lambda_{n2})$ in (13).

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)' = \arg\min_{\boldsymbol{\theta}} L_{n1}(y_i, \mathbf{x}_i; \boldsymbol{\theta}, \lambda_{n2}), \tag{14}$$

and we denote the selected set of the control variables using the data $(y_i, \mathbf{x}_i)$ as $\hat{I}_1 = \{j : \hat{\theta}_j \neq 0, j = 1, 2, \dots, p\}$.

In the second step, we select both relevant control variables and the IVs for the endogenous treatment in the reduced-form equation (12). This step is crucial in the algorithm because it can estimate the optimal instrument using high-dimensional IVs and select additional relevant control variables that might be missed in the first step but are nonetheless important to the treatment variable. This step helps to mitigate omitted variable bias and enhance IV validity. The objective function is

$$\begin{aligned} &Q_{n2}(D_i, \mathbf{z}_i, \mathbf{x}_i; \boldsymbol{\gamma}_n, \boldsymbol{\omega}_n, \lambda_{n3}) \\ &\quad = \sum_{i=1}^{n} \log f_n(D_i, \mathbf{z}_i, \mathbf{x}_i; \boldsymbol{\gamma}_n, \boldsymbol{\omega}_n) - n \sum_{j=1}^{p} p_{\lambda_{n3}}(|\gamma_{nj}|) - n \sum_{k=1}^{m} p_{\lambda_{n3}}(|\omega_{nk}|), \end{aligned} \tag{15}$$

where the estimator $\hat{\boldsymbol{\gamma}}_n$ and $\hat{\boldsymbol{\omega}}_n$ can be obtained by maximizing the objective function

$$(\hat{\boldsymbol{\gamma}}_n', \hat{\boldsymbol{\omega}}_n')' = \underset{\boldsymbol{\gamma}_n, \boldsymbol{\omega}_n}{\operatorname{argmax}}\, Q_{n2}(D_i, \mathbf{z}_i, \mathbf{x}_i; \boldsymbol{\gamma}_n, \boldsymbol{\omega}_n, \lambda_{n3}). \tag{16}$$

We denote $\hat{I}_2=\{j : \hat{\omega}_j \neq 0, j = 1, 2, \dots, m\}$ as the set of the selected confounders $\mathbf{x}_i$ using data $(D_i, \mathbf{z}_i, \mathbf{x}_i)$. The optimal instrument is estimated by $\hat{p}(\mathbf{z}_i, \mathbf{x}_i; \hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\omega}}_n)$ by (12), which is denoted as $\hat{D}_i^* = \hat{p}(\mathbf{z}_i, \mathbf{x}_i; \hat{\boldsymbol{\gamma}}_n, \hat{\boldsymbol{\omega}}_n)$, as in Section 2 with a slight abuse of notation.

In the third step, we define the post-double-selection LIVE $\hat{\beta}_{\text{DS}}$ for the dummy endogenous treatment effect based on the predicted treatment variable $\hat{D}_i^*$ and the union of the control variables selected in the first two variable selection steps denoted by $\hat{I} = \hat{I}_1 \cup \hat{I}_2$. Specifically, the DS-LIVE for β_0 is defined by

$$\hat{\beta}_{\text{DS}} = \left(\hat{\mathbf{D}}^{*\prime} \mathbf{Q}_{\hat{I}} \hat{\mathbf{D}}^*\right)^{-1} \left(\hat{\mathbf{D}}^{*\prime} \mathbf{Q}_{\hat{I}} \mathbf{Y}\right), \tag{17}$$

where $\mathbf{Q}_{\hat{I}} = \mathbf{I}_n - \mathcal{P}_{\hat{I}}$, $\mathcal{P}_{\hat{I}} = \mathbf{X}_{\hat{I}}(\mathbf{X}_{\hat{I}}'\mathbf{X}_{\hat{I}})^{-1}\mathbf{X}_{\hat{I}}'$, $\mathbf{X}_{\hat{I}}$ denotes the design matrix $\mathbf{X}$ corresponding to the selected control variables.

For illustration, we examine the finite-sample performance of the proposed DS-LIVE to show the importance of double selection when there are too many control variables in the structural equation. We consider the following structural model:

$$y_i = D_i\beta_0 + \mathbf{x}_i'\boldsymbol{\theta}_0 + \varepsilon_i, \tag{18}$$

where we set the true treatment effect parameter $\beta_0 = 0.75$, $\boldsymbol{\theta}_0 = (3, 0.15, 0.18, 1.5, 2, \mathbf{0}_{m-5})'$. Note that some true coefficients of key control variables are relatively small, which means they could be missed in the single-variable selection.

We randomly generate control variables $\mathbf{x}_i$ from a multivariate normal distribution $N(0, \Sigma_{\mathbf{X}})$ with $\Sigma_{\mathbf{X}} = (\rho_{ij})_{m\times m}$ with $\rho_{ij} = 0.5^{|i-j|}$ for $i, j = 1, 2, \dots, m$. We fix the sample size as $n = 100$ and the dimension of the control variables as $m = 200$, which is greater than n. The dummy endogenous treatment variable is generated by the Bernoulli distribution with $\exp(\eta_i)/(1 + \exp(\eta_i))$ for which $\eta_i = 0.8x_{i1} + 1.96x_{i2} + 1.85x_{i3} + 0.9x_{i4} + 0.7x_{i5} + 1.16z_{i1} + 0.7z_{i2} + 0.95z_{i3} + \xi_i$, and the IVs $\mathbf{z}_i$ are generated from another multivariate normal distribution $N(0, \Sigma_{\mathbf{Z}})$ with $\Sigma_{\mathbf{Z}} = (\rho_{ij})_{p\times p}$, $\rho_{ij} = 0.5^{|i-j|}$ for $i, j = 1, 2, \dots, p$ with $p = 20$. The generation of the two random errors $(\varepsilon_i, \xi_i)'$ is the same as that in Section 4 to guarantee endogeneity. We obtain the DS-LIVE estimators with two different penalties (LASSO and SCAD), denoted by "DS-LIVE-LASSO" and "DS-LIVE-SCAD," respectively. For comparison purposes, the following estimators are also considered. First, if we ignore the endogeneity in the treatment variable, we can consider the single-variable selection methods ("LASSO" and "SCAD") in the structural equation and the traditional double selection estimators (Belloni, Chernozhukov & Hansen, 2014) with LASSO and SCAD penalties ("DS-LASSO," "DS-SCAD," respectively). Second, if we ignore the binary nature of the treatment variable, we can consider the hybrid estimators combining the TSLS methods based on linear reduced-form models and those based on double selection, denoted by "DS-TSLS-LASSO" and "DS-TSLS-SCAD," respectively. Note that DS-TSLS-LASSO can also be considered a hybrid of two methods in Belloni et al. (2012) and Belloni, Chernozhukov & Hansen (2014).

Table 9 summarizes the average of the estimated biases (Bias) with standard deviations in parentheses and MSE based on 1000 simulations. Figure 2 displays the boxplots of different estimators for the endogenous treatment effect. Without considering the endogeneity of the

TABLE 9: Average biases (standard deviations) and mean-squared errors (MSEs) of different estimators for the endogenous treatment effect $\beta_0 = 0.75$ from Equation (18).

Method	Bias	MSE	Method	Bias	MSE
LASSO	0.1630 (0.1704)	0.0556	SCAD	0.1986 (0.1658)	0.0669
DS-LASSO	0.1248 (0.1717)	0.0450	DS-SCAD	0.1405 (0.1586)	0.0448
DS-TSLS-LASSO	0.0846 (1.3356)	1.7894	DS-TSLS-SCAD	0.0491 (0.6760)	0.4589
DS-LIVE-LASSO	0.0244 (0.1869)	0.0355	DS-LIVE-SCAD	0.0100 (0.1897)	0.0360

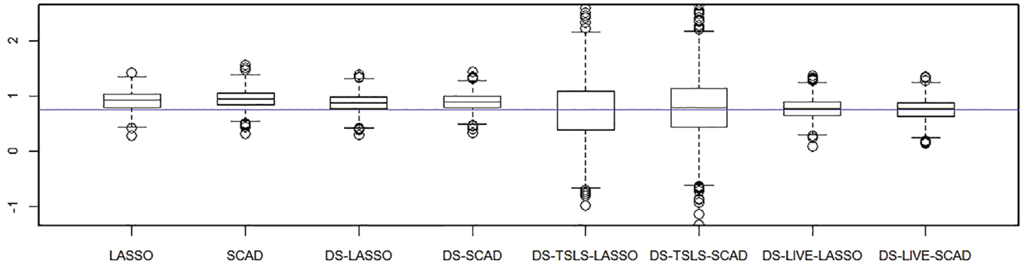

FIGURE 2: Boxplots of different estimators for the endogenous treatment effect ($\beta_0 = 0.75$) in the high-dimensional controls case.

treatment variable, the ordinary penalized methods (LASSO and SCAD) and the double selection methods (DS-LASSO and DS-SCAD) result in large biases in the estimators for the treatment effect. The latter DS estimators have smaller biases than the former penalized methods (LASSO or SCAD) because double selection can reduce omitted variable bias. Compared with the DS-TSLS-LASSO and DS-TSLS-SCAD estimators, DS-LIVE-LASSO and DS-LIVE-SCAD have smaller biases and much smaller MSEs. This is again because LIVE considers the binary nature of the dummy endogenous treatment variable and produces more efficient estimators.

Overall, our simulation results demonstrate that the proposed DS-LIVE is very useful in estimating dummy endogenous treatment effects, especially when there are many potential instruments and high-dimensional control variables.

7. CONCLUSION

In this article, we studied the dummy endogenous variable problem. We considered the penalized logistic regression reduced-form model to estimate the optimal instruments. The proposed penalized LIVE with the SCAD penalty for estimating a binary endogenous treatment effect was found to be root-n consistent and asymptotically normal. Simulation studies have demonstrated that the LIVE-based method performs better than TSLS with a post-LASSO estimator. Empirical studies have shown that the LIVE method can be applied to deal with the dummy endogenous variables problem with many potential instruments, without knowing which ones are significant. We also developed a DS-LIVE for cases when there are many control variables. The proposed methods can be extended to nonparametric models in future studies.

ACKNOWLEDGEMENTS

We thank the editorial team for the insightful comments that helped us to greatly improve the earlier version of this article. We are grateful to Christian Hansen, Whitney Newey, and the seminar and conference participants from Australia National University, Monash University, Peking University, AMES 2019, 11th ICSA International Conference, and others, for their helpful comments, and Jing Zhou (RUC) for sharing the education data. Zhong's work is supported by National Natural Science Foundation of China (11922117 and 11671334) and Fujian Provincial Natural Science Fund for Distinguish Young Scholars (2019J06004). Fan's research, in part, was supported by the National Natural Science Foundation of China Grants 71671149, 71801183 and 72033008 (Key Project), the Science Foundation of the Ministry of Education of China (18YJC790073) and the start-up fund from The Chinese University of Hong Kong.

REFERENCES

Amemiya, T. (1974). The non-linear two-stage least squares estimator. *Journal of Econometrics*, 2, 105–110.

Angrist, J. D. & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90, 431–442.

Baade, R. & Matheson, V. (2016). Going for the gold: The economics of the Olympics. *Journal of Economic Perspectives*, 30, 201–218.

Bai, J. & Ng, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26, 1577–1606.

Bazzi, S. & Clemens, M. (2013). Instruments: Avoiding common pitfalls in identifying the causes of economic growth. *American Economic Journal: Macroeconomics*, 5, 152–186.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80, 2369–2429.

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

Bruckner, M. & Pappa, E. (2015). News shocks in the data: Olympic games and their macroeconomic effects. *Journal of Money, Credit and Banking*, 47, 1339–1367.

Cai, Z., Das, M., Xiong, H., & Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, 133, 207–241.

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

Caner, M. & Fan, Q. (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive LASSO. *Journal of Econometrics*, 187, 256–274.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170, 383–398.

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., & Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educational Research Review*, 14, 33–46.

Coates, D. (2007). Stadiums and arenas: Economic development or economic redistribution? *Contemporary Economic Policy*, 25, 565–577.

Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, 124, 335–361.

De la Peña, V. H., Lai, T., & Shao, Q. (2009). Self-normalized processes: Limit theory and statistical applications. *Probability and its Applications*, Springer-Verlag, Berlin.

Dohl, G. & Lochner, L. (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *American Economic Review*, 102, 1927–1956.

Donald, S. G. & Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69, 1161–1191.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and it oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. & Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics*, 42, 872–917.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32, 928–961.

Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38, 3567–3604.

Fan, Q. & Zhong, W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *Journal of Business and Economic Statistics*, 36, 388–399.

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

Gautier, E. & Tsybakov, A. B. (2018). High-dimensional instrumental variables regression and confidence sets, https://arxiv.org/abs/1812.11330.

Hansen, C. & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182, 290–308.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46, 931–959.

Huang, J., Horowitz, J., & Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38, 2282–2313.

Jing, B., Shao, Q., & Wang, Q. (2003). Self-normalized Cramer-type large deviations for independent random variables. *Annals of Probability*, 31, 2167–2215.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111, 132–144.

Kloek, T. & Mennes, L. B. M. (1960). Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica*, 28, 45–61.

Lin, W., Feng, R., & Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110, 270–288.

Mai, Q. & Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100, 229–234.

Newey, W. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58, 809–837.

Owen, J. G. (2005). Estimating the cost and benefit of hosting Olympic games. *The Industrial Geographer*, 1, 1–18.

Rose, A. K. & Spiegel, M. M. (2011). The Olympic effect. *Economic Journal*, 121, 652–677.

Tibshirani, R. (1996). Regression shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H., Li, R., & Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553–568.

Windmeijer, F., Farbmacher, H., Davies, N., & Smith, G. D. (2019). On the use of the LASSO for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114, 1339–1350.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge.

Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics*, 182, 226–234.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

Zhang, C. & Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*, 36, 1567–1594.

Zhong, W., Gao, Y., Zhou, W., & Fan, Q. (2021). Endogenous treatment effect estimation using high-dimensional instruments and double selection. *Statistics and Probability Letters*, 169, 108967.

Zimbalist, A. (2015). *Circus Maximus: The Economic Gamble Behind Hosting the Olympics and the World Cup*. Brookings Institution Press, Washington, DC.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

APPENDIX: PROOFS

We first need to clarify that the probability density function $f_n(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n)$ in (4), denoted for notational simplicity as $f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)$ when the true underlying parameter $\boldsymbol{\gamma}_n^*$ is considered, satisfies the following regularity conditions (A1)–(C1):

(A1) The instrumental variables $\mathbf{z}_i$, $i = 1, \dots, n$ are independent and identically distributed with the probability density function $f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)$, which has a common support, and the usual identification condition holds. Furthermore, the first and second derivatives of the likelihood function satisfy the following equations:

$$\mathrm{E}_{\gamma_n^*}\left\{\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nj}}\right\} = 0, \quad j = 0, 1, \dots, p,$$

and

$$\mathrm{E}_{\gamma_n^*}\left\{\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nj}}\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nk}}\right\} = -\mathrm{E}_{\gamma_n^*}\left\{\frac{\partial^2 \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nj}\partial \gamma_{nk}}\right\}$$

for $j, k = 0, 1, \dots, p$.

(B1) The first $q_n \times q_n$ submatix $\mathbf{I}_n^{(1)}(\boldsymbol{\gamma}_n^*)$ of the Fisher information matrix is

$$\mathbf{I}_n^{(1)}(\boldsymbol{\gamma}_n^*) = \mathrm{E}_{\gamma_n^*}\left[\left(\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \boldsymbol{\gamma}_n}\right)\left(\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \boldsymbol{\gamma}_n}\right)'\right].$$

Then there exist $C_1, C_2, C_3, C_4 > 0$ such that

$$0 < C_1 < \lambda_{\min}\left(\mathbf{I}_n^{(1)}(\boldsymbol{\gamma}_n^*)\right) \le \lambda_{\max}\left(\mathbf{I}_n^{(1)}(\boldsymbol{\gamma}_n^*)\right) < C_2 < \infty.$$

There also exist $C_3, C_4 > 0$, and $j, k = 0, 1, \dots, p$ such that

$$\mathrm{E}_{\gamma_n^*}\left\{\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nj}}\frac{\partial \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nk}}\right\}^2 < C_3 < \infty$$

and

$$\mathrm{E}_{\gamma_n^*}\left\{\frac{\partial^2 \log f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n^*)}{\partial \gamma_{nj}\partial \gamma_{nk}}\right\}^2 < C_4 < \infty,$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of the given matrix A, respectively.

(C1) There is an open subset $\omega_n \in \Omega_n \in \mathrm{R}^p$ which contains the true parameter value of $\boldsymbol{\gamma}_n^*$, such that for almost all $\mathbf{z}_i$ the density function admits all third derivatives $\partial^3 f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n)/\partial \gamma_{nj}\partial \gamma_{nk}\partial \gamma_{nl}$ for all $\boldsymbol{\gamma}_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that

$$\left|\frac{\partial^3 f_n(\mathbf{z}_i, \boldsymbol{\gamma}_n)}{\partial \gamma_{nj}\partial \gamma_{nk}\partial \gamma_{nl}}\right| \le M_{njkl}(\mathbf{z}_i)$$

for all $\boldsymbol{\gamma}_n \in \omega_n$ and $\mathrm{E}_{\boldsymbol{\gamma}_n}(M_{njkl}(\mathbf{z}_i))^2 < C_5 < \infty$ for all p, n and j, k, l and $k \ge 1$.

These regularity conditions (A1)–(C1) are the conditions imposed in Fan & Peng (2004). In particular, conditions (B1) and (C1) impose the second and fourth moments of the likelihood function. The information matrix of the likelihood function is assumed to be positive definite, and its eigenvalues are uniformly bounded. Considering that the density function $f_n(D_i, \mathbf{z}_i; \boldsymbol{\gamma}_n) = \frac{\exp\left(D_i \boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i\right)}{1+\exp\left(1+e^{\boldsymbol{\gamma}_n' \tilde{\mathbf{z}}_i}\right)}$ given in (4) satisfies the regularity conditions (A1)–(C1) automatically as condition (C) holds.

Proof of Lemma 1. According to (3), it follows from the Taylor expansion that

$$
\begin{aligned}
&\widehat{p}(\mathbf{z}_i, \widehat{\boldsymbol{\gamma}}_n) - p(\mathbf{z}_i, \boldsymbol{\gamma}_n^*) \\
&\quad = \frac{\exp(\widehat{\boldsymbol{\gamma}}_n' \tilde{\mathbf{z}}_i)}{1 + \exp(\widehat{\boldsymbol{\gamma}}_n' \tilde{\mathbf{z}}_i)} - \frac{\exp(\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i)}{1 + \exp(\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i)} \\
&\quad = \frac{\exp(\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i)}{(1 + \exp(\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i))^2} (\widehat{\boldsymbol{\gamma}}_n' \tilde{\mathbf{z}}_i - \boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i) + \frac{e^{\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i}(1 - e^{2\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i})}{2(1 + \exp(\boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i))^4} (\widehat{\boldsymbol{\gamma}}_n' \tilde{\mathbf{z}}_i - \boldsymbol{\gamma}_n^* \tilde{\mathbf{z}}_i)^2 (1 + o(1)) \\
&\quad \leq (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)' \tilde{\mathbf{z}}_i + ((\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)' \tilde{\mathbf{z}}_i)^2 (1 + o(1)) = O_p(\sqrt{p_n^2/n})
\end{aligned}
$$

as $(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)' \tilde{\mathbf{z}}_i \leq \|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*\| \cdot \|\tilde{\mathbf{z}}_i\| = O_p(\sqrt{p_n/n}) O_p(\sqrt{p_n}) = O_p(\sqrt{p_n^2/n})$ holds by applying the Cauchy–Schwarz inequality and condition (C). The proof of Lemma 1 is completed. ■

Lemma A1. *Let* $X_1, \dots, X_n$ *be the triangular array of i.i.d. zero-mean random variable. Suppose that* $\mathrm{M}_n = (\mathrm{E}\, X_1^2)^{1/2} / (\mathrm{E}\, |X_1|^3)^{1/3} > 0$ *and that for some* $b_n \to \infty$ *slowly,* $n^{1/6}\, \mathrm{M}_n / b_n \geq 1$. *Then, uniformly on* $0 \leq x \leq n^{1/6}\, \mathrm{M}_n / b_n - 1$, *we have*

$$
\left| \frac{\mathrm{P}(|S_n / V_n \geq x| \geq x)}{2(1 - \Phi(x))} - 1 \right| \leq \frac{A}{b_n^3} \to 0,
$$

where $S_n = \sum_{i=1}^n X_i$, $V_n = \sum_{i=1}^n X_i^2$, $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution, and A is a positive constant.*

The moderate deviation inequality for self-normalized sums was originally from Jing, Shao & Wang (2003), and used later in De la Pẽna, Lai & Shao (2009), Belloni et al. (2012), Belloni, Chernozhukov & Hansen (2014) and Fan & Zhong (2018).

Proof of Theorem 1. We present the matrix form of (8)

$$
\mathbf{Y} = \mathbf{D}^* \beta_0 + \mathbf{X} \boldsymbol{\theta}_0 + \boldsymbol{\nu}. \tag{A1}
$$

Substituting (A1) into (9), we have

$$
\begin{aligned}
\widehat{\beta} &= \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right)^{-1} \widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \left(\mathbf{D}^* \beta_0 + \mathbf{X} \boldsymbol{\theta}_0 + \boldsymbol{\nu} \right) \\
&= \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right)^{-1} \widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \left((\widehat{\mathbf{D}}^* + \mathbf{D}^* - \widehat{\mathbf{D}}^*) \beta_0 + \mathbf{X} \boldsymbol{\theta}_0 + \boldsymbol{\nu} \right) \\
&= \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right)^{-1} \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right) \beta_0 + \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right)^{-1} \widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \left[(\mathbf{D}^* - \widehat{\mathbf{D}}^*) \beta_0 + \boldsymbol{\nu} \right] \\
&= \beta_0 + \left(\widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \widehat{\mathbf{D}}^* \right)^{-1} \widehat{\mathbf{D}}^{*\prime} \mathbf{Q} \left[(\mathbf{D}^* - \widehat{\mathbf{D}}^*) \beta_0 + \boldsymbol{\nu} \right],
\end{aligned} \tag{A2}
$$

where the third equality follows that $\mathbf{QX} = (I_n - \mathcal{P}_{\mathbf{X}})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = 0$, which yields

$$\sqrt{n}(\hat{\beta} - \beta_0) = \left(\hat{\mathbf{D}}^{*\prime}\mathbf{Q}\hat{\mathbf{D}}^{*}/n\right)^{-1}\hat{\mathbf{D}}^{*\prime}\mathbf{Q}\left[(\mathbf{D}^{*} - \hat{\mathbf{D}}^{*})\beta_0 + \mathbf{v}\right]/\sqrt{n}$$
$$=: T_2^{-1} \cdot T_1. \tag{A3}$$

For notational simplicity, denote $\zeta = \hat{\mathbf{D}}^{*} - \mathbf{D}^{*}$, and for a vector $\mathbf{W} \in \mathbb{R}^n$, let $\alpha_{\mathbf{W}}(\mathbf{X}) := \arg\min_{\mathbf{b}\in\mathbb{R}^m} ||\mathbf{W} - \mathbf{X}\mathbf{b}||^2$. First we consider the T_2 term in (A3),

$$\begin{aligned} T_2 &= \hat{\mathbf{D}}^{*\prime}\mathbf{Q}\hat{\mathbf{D}}^{*}/n \\ &= \left(\hat{\mathbf{D}}^{*} - \mathbf{D}^{*} + \mathbf{D}^{*}\right)'\mathbf{Q}\left(\hat{\mathbf{D}}^{*} - \mathbf{D}^{*} + \mathbf{D}^{*}\right)/n \\ &= \frac{\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^{*}}{n} + \frac{\zeta'\mathbf{Q}\zeta}{n} + \frac{2\zeta'\mathbf{Q}\mathbf{D}^{*}}{n} \\ &=: \frac{\mathbf{D}^{*\prime}\mathbf{Q}\mathbf{D}^{*}}{n} + T_{21} + T_{22}. \end{aligned} \tag{A4}$$

Note that we decompose T_{21} into two parts as follows:

$$T_{21} = \frac{\zeta'\zeta}{n} - \frac{\zeta'\mathcal{P}_{\mathbf{X}}\zeta}{n} = T_{21,a} - T_{21,b}. \tag{A5}$$

Combining with Lemma 1, we have

$$|T_{21,a}| = \left|\frac{\zeta'\zeta}{n}\right| = \frac{\|\hat{\mathbf{D}}^{*} - \mathbf{D}^{*}\|_2^2}{n} = O_p\left(\frac{p_n^2}{n^2}\right). \tag{A6}$$

Similar to the treatment in Belloni, Chernozhukov & Hansen (2014), we deal with $T_{21,b}$:

$$\begin{aligned} |T_{21,b}| &= \left|\frac{\zeta'\mathcal{P}_{\mathbf{X}}\zeta}{n}\right| = \left|\frac{\zeta'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\zeta}{n}\right| = \left|\frac{\alpha_\zeta'(\mathbf{X})\mathbf{X}'\zeta}{n}\right| \\ &\le \left\|\alpha_\zeta'(\mathbf{X})\right\|_1 \left\|\frac{\mathbf{X}'\zeta}{n}\right\|_\infty, \end{aligned} \tag{A7}$$

where it follows from condition (D) that

$$\begin{aligned} \left\|\alpha_\zeta'(\mathbf{X})\right\|_1 &\le \sqrt{m}\left\|\alpha_\zeta'(\mathbf{X})\right\| = \sqrt{m}\left\|\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\zeta}{n}\right\| \\ &\le \sqrt{m}c_3^{-1}\left\|\frac{\mathbf{X}'\zeta}{n}\right\|_\infty. \end{aligned} \tag{A8}$$

To derive the bound of $\|\mathbf{X}'\zeta\|_\infty$, we apply Lemma A1 on the tail bound for self-normalized deviations:

$$\mathrm{P}\left(\max_{1\le j\le m}\left|\frac{\sum_{i=1}^n x_{ij}\zeta_i}{\sqrt{\sum_{i=1}^n x_{ij}^2\zeta_i^2}}\right| > \sqrt{\log n}\right)$$

$$\begin{aligned}
&\leq \max_{1\leq j\leq m} \mathrm{P}\left(\left|\frac{\sum_{i=1}^{n} x_{ij}\zeta_i}{\sqrt{\sum_{i=1}^{n} x_{ij}^2\zeta_i^2}}\right| > \sqrt{\log n}\right)\\
&\leq 2\,(1-\Phi(\log n))\,(1+o(1)) = \frac{2\exp\left(-\frac{\log n}{2}\right)}{\sqrt{2\pi}\log n}(1+o(1))\\
&= \frac{1}{\sqrt{n}\log n}(1+o(1)) \to 0, \qquad \text{(A9)}
\end{aligned}$$

as $\mathrm{P}(U > u) \leq \exp(-u^2/2)/(u\sqrt{2\pi})$ holds for a standard normal random variable U, which entails that

$$\max_{1\leq j\leq m}\left|\frac{\sum_{i=1}^{n} x_{ij}\zeta_i}{\sqrt{\sum_{i=1}^{n} x_{ij}^2\zeta_i^2}}\right| = O_p(\sqrt{\log n}). \qquad \text{(A10)}$$

Note that by condition (C), there exists a constant c_4 such that

$$\max_{1\leq j\leq m}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\zeta_i^2} \leq c_4\|\widehat{\mathbf{D}}^* - \mathbf{D}^*\|/\sqrt{n} = O_p\left(\frac{p_n}{n}\right). \qquad \text{(A11)}$$

By combining (A9), (A10) and (A11), we have

$$\left\|\frac{\mathbf{X}'\zeta}{\sqrt{n}}\right\|_\infty = O_p(\sqrt{\log n}), \qquad \text{(A12)}$$

which is plugged into (A8) to give

$$\left\|\alpha_\zeta(\mathbf{X})\right\|_1 = O_p\left(\sqrt{\frac{\log n}{n}}\right). \qquad \text{(A13)}$$

It follows from (A7), (A8), (A12) and (A13) that

$$|T_{21,b}| \leq \sqrt{m}c_3^{-1}\left(\left\|\frac{\mathbf{X}'\zeta}{n}\right\|_\infty\right)^2 = O_p\left(\frac{\log n}{n}\right), \qquad \text{(A14)}$$

which combines with (A6) to derive that

$$|T_{21}| = O_p\left(\frac{p_n^2}{n^2}\right) + O_p\left(\frac{\log n}{n}\right). \qquad \text{(A15)}$$

Next, we rewrite the term T_{22} in (A4) as

$$T_{22} = \frac{2\zeta'\mathbf{D}^*}{n} - \frac{2\zeta'\mathcal{P}_{\mathbf{X}}\mathbf{D}^*}{n} =: T_{22,a} - T_{22,b}. \qquad \text{(A16)}$$

The first term in (A16) follows from Lemma 1:

$$|T_{22,a}| = \left|\frac{2\sum_{i=1}^{n}(\widehat{\mathbf{D}}_i^* - \mathbf{D}_i^*)\mathbf{D}_i^*}{n}\right| \leq \left|\frac{2\sum_{i=1}^{n}(\widehat{\mathbf{D}}_i^* - \mathbf{D}_i^*)}{n}\right| = O_p\left(\sqrt{\frac{p_n^2}{n}}\right). \tag{A17}$$

Combining with (A13), we have

$$\begin{aligned}|T_{22,b}| &= \left|\frac{\zeta'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^*}{n}\right| \leq \left\|\alpha_\zeta(\mathbf{X})\right\|_1 \left\|\frac{\mathbf{X}'\mathbf{D}^*}{n}\right\|_\infty \\ &= O_p\left(\sqrt{\frac{\log n}{n}}\right)\max_{1\leq j\leq m}\left|\frac{\sum_{i=1}^{n}\mathbf{X}_{ij}\mathbf{D}_i^*}{n}\right| \\ &= O_p\left(\sqrt{\frac{\log n}{n}}\right)\max_{1\leq j\leq m}\left|\frac{\sum_{i=1}^{n}\mathbf{X}_{ij}}{n}\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right).\end{aligned} \tag{A18}$$

It follows from (A16), (A17) and (A18) that

$$|T_{22}| = O_p\left(\sqrt{\frac{p_n^2}{n}}\right) + O_p\left(\sqrt{\frac{\log n}{n}}\right). \tag{A19}$$

Combining with (A4), (A15) and (A19), we derive

$$T_2 = \frac{\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*}{n} + o_p(1). \tag{A20}$$

Next, we consider the T_1 term, which can be rewritten as follows:

$$\begin{aligned}T_1 &= \widehat{\mathbf{D}}^{*'}\mathbf{Q}\left[(\mathbf{D}^* - \widehat{\mathbf{D}}^*)\beta_0 + \boldsymbol{\nu}\right]/\sqrt{n} \\ &= \frac{\widehat{\mathbf{D}}^{*'}\mathbf{Q}(\mathbf{D}^* - \widehat{\mathbf{D}}^*)}{\sqrt{n}}\beta_0 + \frac{\widehat{\mathbf{D}}^{*'}\mathbf{Q}\boldsymbol{\nu}}{\sqrt{n}} \\ &= \frac{\mathbf{D}^{*'}\mathbf{Q}\boldsymbol{\nu}}{\sqrt{n}} - \frac{\zeta'\mathbf{Q}\zeta}{\sqrt{n}}\beta_0 - \frac{\mathbf{D}^{*'}\mathbf{Q}\zeta}{\sqrt{n}}\beta_0 + \frac{\zeta'\mathbf{Q}\boldsymbol{\nu}}{\sqrt{n}} \\ &= \frac{\mathbf{D}^{*'}\mathbf{Q}\boldsymbol{\nu}}{\sqrt{n}} - T_{11} - T_{12} + T_{13}.\end{aligned} \tag{A21}$$

It follows from (A15) and (A19) that

$$\left|T_{11}\right| = o_p(1) \quad \text{and} \quad \left|T_{12}\right| = o_p(1). \tag{A22}$$

Note that

$$T_{13} = \frac{\zeta'\boldsymbol{\nu}}{\sqrt{n}} - \frac{\zeta'\mathcal{P}_{\mathbf{X}}\boldsymbol{\nu}}{\sqrt{n}} =: T_{13,a} - T_{13,b}. \tag{A23}$$

It follows from Lemma 1, condition (E) and the Cauchy−Schwarz inequality that

$$|T_{13,a}| = \left| \frac{\sum_{i=1}^{n} (\widehat{\mathbf{D}}_i^* - \mathbf{D}_i^*) \nu_i}{\sqrt{n}} \right| \leq \frac{\|\widehat{\mathbf{D}}^* - \mathbf{D}^*\| \|\boldsymbol{\nu}\|}{\sqrt{n}} = O_p\left(\sqrt{\frac{p_n^2}{n}} \right). \tag{A24}$$

Similar to (A7), it follows from (A13) that

$$\begin{aligned} T_{13,b} &= \left| \frac{\zeta' \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\nu}}{\sqrt{n}} \right| = \left| \frac{\alpha_\zeta'(\mathbf{X}) \mathbf{X}' \boldsymbol{\nu}}{\sqrt{n}} \right| \\ &\leq \|\alpha_\zeta(\mathbf{X})\|_1 \left\| \frac{\mathbf{X}' \boldsymbol{\nu}}{\sqrt{n}} \right\|_\infty \\ &= O_p\left(\sqrt{\frac{\log n}{n}} \right) O_p(\sqrt{\log n}) = O_p\left(\sqrt{\frac{\log n^2}{n}} \right), \end{aligned} \tag{A25}$$

where the bound for $\|\mathbf{X}'\boldsymbol{\nu}/\sqrt{n}\|_\infty$ is the same as (A12). Combining with (A23), (A24) and (A25), we derive

$$T_1 = \frac{\mathbf{D}^{*'} \mathbf{Q} \boldsymbol{\nu}}{\sqrt{n}} + o_p(1),$$

which yields by (A20)

$$\sqrt{n}(\widehat{\beta} - \beta) = \left(\frac{\mathbf{D}^{*'} \mathbf{Q} \mathbf{D}^*}{n} + o_p(1) \right)^{-1} \left(\frac{\mathbf{D}^{*'} \mathbf{Q} \boldsymbol{\nu}}{\sqrt{n}} + o_p(1) \right). \tag{A26}$$

Note that $\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*/n = \frac{1}{n}\sum_{i,j=1}^{n} D_i^* Q_{i,j} D_j^* = \mathrm{E}(\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*) + o_p(1)$ by the weak law of large numbers. As $D_i^* Q_{i,j} \nu_j$ are i.i.d. with mean zero and variance $\sigma_n^2 = (\mathrm{E}(\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*))^{-1}\,\mathrm{E}(\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*\nu_i^2)$ $(\mathrm{E}(\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*))^{-1}$. Applying the central limit theorem and Slutsky's theorem, we have $\sigma_n^{-1}\sqrt{n}(\widehat{\beta} - \beta) \to N(0, 1)$. If $\mathrm{Var}(\nu_i) = \sigma_\nu^2$ satisfies the homoscedastic condition, $\sigma_n^2 = (\mathrm{E}(\mathbf{D}^{*'}\mathbf{Q}\mathbf{D}^*))^{-1}\sigma_\nu^2$ holds.

The proof is completed. ■

Received 24 October 2020
Accepted 29 March 2021