Inference for Two-Stage Extremum Estimators*

Aristide Houndetoungan†[a] and Abdoul Haki Maoude[b]

[a]*Cy Cergy Paris University and Thema*
[b]*Concordia University, Department of Statistics*

February 6, 2024

Abstract

We present a simulation-based approach to approximate the asymptotic variance and asymptotic distribution function of two-stage estimators. We focus on extremum estimators in the second stage and consider a large class of estimators in the first stage. This class includes extremum estimators, high-dimensional estimators, and other types of estimators (e.g., Bayesian estimators). We accommodate scenarios where the asymptotic distributions of both the first- and second-stage estimators are non-normal. We also allow for the second-stage estimator to exhibit a significant bias due to the first-stage sampling error. We introduce a debiased plug-in estimator and establish its limiting distribution. Our method is readily implementable with complex models. Unlike resampling methods, we eliminate the need for multiple computations of the plug-in estimator. Monte Carlo simulations confirm the effectiveness of our approach in finite samples. We present an empirical application with peer effects on adolescent fast-food consumption habits, where we employ the proposed method to address the issue of biased instrumental variable estimates resulting from the presence of many weak instruments.

Keywords: Hypothesis Testing, Two-stage Estimators, Semiparametric and Nonparametric Methods, Resampling Methods, High-Dimensional Asymptotics

JEL Classification: C12, C13, C14, C15, C55.

*For comments and suggestions, we are grateful to Arnaud Dufays, Ulrich Hounyo, Mathieu Marcoux, Antoine Djogbenou, Désiré Kédagni, Pamela Giustinelli, Florian Pelgrin, and Davide Giraudo. This research uses data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a program that is directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by Grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other US federal agencies and foundations. Special acknowledgment is given to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain Add Health data files is available on the Add Health website (www.cpc.unc.edu/addhealth). No direct support was received from Grant P01-HD31921 for this research. Replication codes for the results from this research are available at https://github.com/ahoundetoungan/InferenceTSE.

†Corresponding author - Email addresses: aristide.houndetoungan@cyu.fr (A. Houndetoungan), abdoulhaki.maoude@concordia.ca (A. H. Maoude)

1 Introduction

Two-stage (or multiple-stage) estimation is a widely used technique for tackling issues such as endogeneity, selection, non-identification, missing values, and high-dimensional data (e.g., Hirano et al., 2003; Jofre-Bonet and Pesendorfer, 2003; Newey and Powell, 2003; Ackerberg et al., 2012; Freyberger and Larsen, 2022; Chernozhukov et al., 2022; Houndetoungan, 2022; Ichimura and Newey, 2022; Boucher and Houndetoungan, 2023). This approach consists of estimating a function (or parameter) in the first stage, followed by "plugging" this estimator into a model, and using a conventional method to estimate other parameters in the final stage. The estimator of the final stage is called a *two-stage* or *plug-in* estimator. Due to the sampling error from the initial stage, asymptotic normality is not always guaranteed in the final stage (e.g., Newey, 1984; Johansen, 1991). Moreover, the plug-in estimator may exhibit a significant first-order bias when the initial-stage sampling error is substantial (Chernozhukov et al., 2017). It can also be challenging to compute the asymptotic variance at the final stage when taking this sampling error into account (Ackerberg et al., 2012).

In this paper, we introduce a novel simulation-based approach for estimating the asymptotic variance and cumulative distribution function (CDF) of plug-in estimators. We focus on cases where the second-stage estimator is an extremum estimator, which encompasses M-estimators, generalized method of moment (GMM) estimators, and minimum distance (MD) estimators. We also consider a large class of estimators in the first stage, including extremum estimators, high-dimensional estimators, and other types of estimators (e.g., Bayesian estimators). Our main assumption is that the *conditional distribution* of the plug-in estimator, *given any realization of the first-stage estimator*, is asymptotically normal (see Fligner and Hettmansperger, 1979; Rubshtein, 1996, for examples of conditional asymptotic normality). This assumption is weak because, by treating the first-stage estimator as a predetermined sequence, the plug-in estimator can be viewed as a single-step estimator, which is generally asymptotically normally distributed (see Newey and McFadden, 1994). Leveraging the asymptotic distribution of the first-stage estimator and the normality at the second stage conditional on the first stage, we simulate the *unconditional* asymptotic CDF of the second-stage estimator.

Our method is versatile and is applicable to many frameworks. First, unlike classical inference methods (e.g., Newey, 1984; Murphy and Topel, 2002), we do not require the first-stage estimator to be root-n consistent, where n is the sample size. The first-stage estimator may converge slowly as is the case in nonparametric modeling or where fewer observations are used in the first stage.

Moreover, our method accommodates situations in which the asymptotic distribution of the first-stage estimator is not normal, leading to a non-normal unconditional asymptotic distribution in the second stage. For instance, in some complex models, Bayesian approaches are used for the inference in the first stage (see Breza et al., 2020; Lubold et al., 2023). However, Zellner and Rossi (1984)

demonstrate that a Bayesian estimator may not be normally distributed. Non-normal asymptotic distributions can also occur with many estimators for time-series models. One example is vector autoregressive models that are estimated using a maximum likelihood method (see Johansen, 1991).

Furthermore, our method can be used even when the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ does not have a zero mean, where $\boldsymbol{\theta}_0$ is the parameter of interest and $\hat{\boldsymbol{\theta}}_n$ is the plug-in estimator. This problem occurs when the sampling error from the first stage induces a significant bias in the second stage (see Belloni et al., 2014b, 2017; Cattaneo et al., 2019). Using the estimate of the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, we show that it is possible to reduce the bias of the plug-in estimator. We then introduce a *debiased plug-in estimator* and demonstrate that its limiting estimator has a zero mean bias.

We illustrate the effectiveness of the debiased approach by studying peer effects on adolescent fast-food consumption habits. We employ an instrumental variable (IV) approach with *many weak instruments* (see Belloni et al., 2017; Mikusheva and Sun, 2022). The large number of weak instruments leads to biased estimates and we reduce the bias using our method.

From a computational standpoint, our approach is suitable for complex models. Our debiased estimator is easily implementable, given that it does not require computing new statistics other than those that are used to estimate the limiting distribution. Unlike resampling methods, we eliminate the need for multiple computations of the plug-in estimator. In practice, our approach requires practitioners to possess a valid estimator for the asymptotic distribution of the first-stage estimator. Since this estimator is typically obtained through a single-step approach, its distribution can be derived using a frequentist method (Amemiya, 1985) or a Bayesian approach such as the Gibbs sampler and Metropolis-Hastings algorithm (Casella and George, 1992; Chib and Greenberg, 1995).

We present a simulation study employing various models where classical asymptotic inference methods *cannot be* applied, including IV models with many instruments (Cattaneo et al., 2019), a latent variable model that is estimated in two steps where the number of observations in the first stage grows slowly with respect to n, and a Copula-GARCH model where the number of returns increases with the sample size (Gonçalves et al., 2023). Our method demonstrates strong performance in finite samples. Even without bias correction for the plug-in estimator, we can construct reliable confidence intervals (CIs) for $\boldsymbol{\theta}_0$. The CIs are not centered on the estimates due to the finite sample bias. We further show that the debiased estimation approach effectively addresses this issue.

Our method also performs as well as the classical asymptotic inference approach when the latter is applicable, for instance, when the first- and second-stage estimators are finite-dimensional extremum estimators that are root-n consistent. This feature is interesting since necessary conditions for the validity of the classical asymptotic inference method may not be easily testable in certain contexts. Our approach can help to prevent potential biases of which practitioners may not be aware.

Related Literature

This paper contributes to the extensive and growing literature on sequential estimators, which addresses issues of asymptotic inference, biased two-stage estimators, and resampling methods.

Inference for Two-Stage Estimators. Most inference methods for two-stage estimators impose regularity conditions to obtain a plug-in estimator that is asymptotically normally distributed. Examples include situations where both the first- and second-stage estimators are finite-dimensional extremum estimators that converge at the same rate (Newey, 1984; Hotz and Miller, 1993), cases where the first-stage estimator is $\sqrt{n}$-consistent and asymptotically normally distributed (Murphy and Topel, 2002), and scenarios where the second-stage estimator is asymptotically invariant to infinitesimal variations in the first-stage estimator (Andrews, 1994; Belloni et al., 2014a; Chernozhukov et al., 2015; Houndetoungan and Kouame, 2023).

We contribute to this literature by proposing a new method that does not require the asymptotic distribution of the first- and second-stage estimators to be normal. Moreover, we do not impose a specific class of estimators in the first stage, making our approach more general. Even though the plug-in estimator is asymptotically normally distributed, accurately computing its variance while considering the sampling error from the first stage can be intricate. Ackerberg et al. (2012) propose a numerical method to approximate this variance when the first-stage estimator is nonparametric. Our simulation method can also be used to compute the asymptotic variance for a broad range of models.

High-Dimensional Modeling and Debiasing Approaches. Our framework is also linked to the literature on two-stage high-dimensional modeling (e.g., Belloni et al., 2014a, 2017; Farrell, 2015; Chernozhukov et al., 2015, 2018; Mikusheva and Sun, 2022). In this context, the number of covariates in the first stage can be of order $\sqrt{n}$. Although the plug-in estimator can still be consistent, the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ may not have a zero mean. The literature proposes several approaches to tackle this issue (e.g., Chernozhukov et al., 2017, 2018; Fernández-Val and Weidner, 2018; Cattaneo et al., 2019).

We contribute to this literature in that we impose no restrictions on the expectation of the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Our approach is applicable when this asymptotic distribution is not centered at zero. We show that this flexibility enables bias reduction for the plug-in estimator. We introduce a debiased plug-in estimator and establish its asymptotic distribution.

Importantly, while many bias reduction methods are designed for specific models, our approach can be applied to a broad class of models. We demonstrate its efficacy in diverse settings; for instance, it yields strong performance with IV models where the number of instruments grows at the same rate as $\sqrt{n}$. Moreover, it excels with latent variable models that are estimated in two steps, where the number of observations in the first stage grows slowly with respect to n. Finally, it also performs well with Copula-GARCH models in which the number of returns is increasing in n. However, like

any method that relies upon asymptotics and contrary to resampling methods, the precision of our method can be mitigated when the sample size is too low.

Resampling Methods. Resampling methods such as bootstrap and jackknife can also be used for the inference of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ (e.g, Efron, 1982; Davidson and MacKinnon, 1999; Andrews, 2002; Cattaneo et al., 2019). Unlike classical inference methods relying on asymptotic normality, some resampling techniques directly approximate the CDF of two-stage estimators, making them applicable even when the asymptotic distribution is not normal. Yet, these approaches require multiple computations of the first- and second-stage estimators, which can result in slow or even infeasible computations. We contribute to this literature by proposing a flexible inference method that requires a single computation of both the first- and second-stage estimators. Our approach can then be applied to complex models when resampling methods can be time-consuming or infeasible.

It is worth noting that the literature also proposes fast bootstrap approaches that are designed to circumvent the problem of multiple optimizations in either one or both stages (e.g., see Andrews, 2002; Hong and Scaillet, 2006; Kline and Santos, 2012; Armstrong et al., 2014; Gonçalves et al., 2023). In addition, Honoré and Hu (2017) make the bootstrap process faster by relying on the estimation of one-dimensional parameters, irrespective of the dimension of $\hat{\boldsymbol{\theta}}_n$. Nevertheless, these recent methods do not handle the issue of finite sample bias that can occur in the second stage. Despite the computational advantage of our approach, we allow for the limiting distribution of the plug-in estimator not to be centered at zero. This flexibility enables us to reduce the bias of the plug-in estimator.

Application. To demonstrate the effectiveness of our method, we revisit the empirical analysis conducted by Fortin and Yazbeck (2015) on peer effects (influence of friends) in adolescent fast-food consumption habits. There is a large and growing literature on peer effects in economics (see De Paula, 2017; Bramoullé et al., 2020, for a review). The instrumental variable (IV) approach is widely adopted in the literature for estimating linear-in-means peer effects models. This approach is popular and easy to implement because the instruments are directly generated from the model using *friends of friends* (Bramoullé et al., 2009). Yet, these instruments may be weak in some cases, leading to estimates biased toward an ordinary least squares (OLS) estimate (Andrews et al., 2019).

We address this issue by using both close- and long-distance friends to construct the instruments, thereby expanding our instrumental variable pool (250 excluded instruments for a sample size of $n = 2,736$). This pool includes *many potentially weak instruments*. We correct the resulting finite sample bias of the IV estimator and provide valid inference using the approach that is proposed in this paper. Our findings indicate that a one-point increase in the average friend's fast-food consumption frequency leads to a 0.23 increase in one's fast-food consumption frequency. This result suggests that a policy focusing on key players in the network can be efficient in combating obesity resulting from fast-food consumption in schools (Ballester et al., 2006; Zenou, 2016; Lee et al., 2021).

Plan of the Paper

The remainder of the paper is organized as follows. In Section 2, we present our framework. Section 3 provides an overview of our approach using a leading example. In Section 4, we present our main results. Section 5 provides a simulation study to assess the finite sample performance of our approach. In Section 6, we present an empirical analysis with peer effects. Section 7 concludes the paper.

Notation

The symbols $\mathbb{E}$ and $\mathbb{V}$ denote expectation and variance, respectively. $\|.\|$ is the ℓ_2-norm. ∂_x is the derivative with respect to some x. If $\boldsymbol{a} = (a_1, \ldots, a_d)'$, $\boldsymbol{b} = (b_1, \ldots, b_d)' \in \mathbb{R}^d$, $\boldsymbol{a} \preceq \boldsymbol{b}$ is equivalent to $a_k \leqslant b_k$ for any integer $k \in [1, k]$. $\boldsymbol{I}_d$ is the d-dimensional identity matrix. plim is the limit in probability as the sample size n grows to infinity. $\operatorname{plim}_\kappa$ is the limit in probability when some κ grows to infinity (n set fixed). $\operatorname{plim}_{\kappa,n}$ for the limit in probability when κ and n grow to infinity. We use the symbol $\lim$ for the classical limit. For a positive definite matrix $\mathbf{M}$, we use $\mathbf{M}^{1/2}$ to denote its Cholesky decomposition and $\mathbf{M}^{-1/2}$ to denote the Cholesky decomposition of its inverse.

2 The Class of Conditional Extremum Estimators

This section introduces the class of plug-in estimators that are studied in this paper. For expositional ease, we consider the case of two-stage estimators. However, our findings can be generalized to multiple-stage estimators, given that what we refer to as the first-stage estimator may encompass many single-step estimators. Moreover, we expose our argument assuming an M-estimator in the second stage. We can extend this to any extremum estimator since inference methods for extremum estimators are similar, regardless of whether it is an M-estimator, GMM estimator, or MD estimator (see Amemiya, 1985). Due to this similarity, we interchangeably use the terms M-estimator and extremum estimator, hoping that this will not confuse the reader.

In the second stage, we assume that the practitioner maximizes an objective function given by

$$Q_n(\boldsymbol{\theta}, \mathbf{y}_n, \mathbf{X}_n, \hat{\mathbf{B}}_n) = \frac{1}{n} \sum_{i=1}^n q(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_{n,i}), \tag{1}$$

where $\mathbf{y}_n = (y_1, \ldots, y_n)'$, $\mathbf{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, $\hat{\mathbf{B}}_n = (\hat{\boldsymbol{\beta}}_{n,1}, \ldots, \hat{\boldsymbol{\beta}}_{n,n})'$, and q is a known function. In Equation (1), y_i and $\boldsymbol{x}_i$ are observed variables for the i-th unit in the sample (e.g., y_i is a dependent variable and $\boldsymbol{x}_i$ are explanatory variables). $\hat{\boldsymbol{\beta}}_{n,1}, \ldots, \hat{\boldsymbol{\beta}}_{n,n}$ are estimators from some first-stage regression (e.g., prediction of some variable in a preliminary regression). The estimator $\hat{\boldsymbol{\beta}}_{n,i}$ may be a scalar or finite-dimensional vector. The subscript i may also refer to time in time-series models.

Let $\hat{\boldsymbol{\theta}}_n$ be the estimator that maximizes the objective function (1). $\hat{\boldsymbol{\theta}}_n$ is called *plug-in* (or *two-stage*) estimator. We also refer to it as a *conditional M-estimator* (or *conditional extremum estimator*) because, given $\hat{\mathbf{B}}_n$ obtained in the first stage, $\hat{\boldsymbol{\theta}}_n$ is simply an extremum estimator. We will refer to $\hat{\mathbf{B}}_n$ as the

first-stage estimator. We do not require $\hat{\boldsymbol{\beta}}_{n,i}$ to originate from an extremum estimation, or to have a particular asymptotic distribution (like the normal distribution). However, we assume that $\hat{\boldsymbol{\beta}}_{n,i}$ uniformly converges in probability to some $\boldsymbol{\beta}_{0,i}$, the true value of the parameter that it is designed to estimate (see Assumption 2.1). We also denote by $\boldsymbol{\theta}_0$ the true value of the parameter $\boldsymbol{\theta}$ (i.e., the value taken by $\boldsymbol{\theta}$ in the data-generating process).

Special cases within our framework arise when $\boldsymbol{\beta}_{0,i} = f(\boldsymbol{z}_i, \boldsymbol{\gamma}_0)$ for some function f, where $\boldsymbol{z}_i$ is a control variable that may overlap components of $\boldsymbol{x}_i$, and $\boldsymbol{\gamma}_0$ is a parameter. In this case, we have $\hat{\boldsymbol{\beta}}_{n,i} = f(\boldsymbol{z}_i, \hat{\boldsymbol{\gamma}}_n)$, where $\hat{\boldsymbol{\gamma}}_n$ is an estimator of $\boldsymbol{\gamma}_0$. An example of this situation is the instrumental variable (IV) approach with $\boldsymbol{z}_i$ being the instrument and $\hat{\boldsymbol{\beta}}_{n,i}$ is the predicted value of the endogenous variable to be plugged into the second stage (Cattaneo et al., 2019). The function f may not depend on i; that is, $\boldsymbol{\beta}_{0,i} = \boldsymbol{\beta}_0$ for any i, where $\boldsymbol{\beta}_0$ is a finite-dimensional vector to be estimated in the first stage (Murphy and Topel, 2002). In the case of semiparametric or nonparametric specification, we can have $\boldsymbol{\beta}_{0,i} = f_n(\boldsymbol{z}_i, \boldsymbol{\gamma}_{0,n})$, where the specification of the function f_n and the dimension of the parameter $\boldsymbol{\gamma}_{0,n}$ depends on the sample size n. Examples of estimators in this situation are power series, splines, and Fourier series approximations (see Belloni et al., 2015).

Let $\mathcal{Y} \subseteq \mathbb{R}^{K_y}$, $\mathcal{X} \subseteq \mathbb{R}^{K_x}$, and $\mathcal{B} \subseteq \mathbb{R}^{K_\beta}$ be the supports of y, $\boldsymbol{x}$, and $\boldsymbol{\beta}_{0,i}$, respectively, where K_y, K_x, and K_β are the corresponding dimensions. Let also $\boldsymbol{\Theta} \subset \mathbb{R}^{K_\theta}$ be the space of $\boldsymbol{\theta}_0$, where K_θ is the dimension of $\boldsymbol{\theta}_0$. We introduce the following assumptions.

Assumption 2.1 (First-Stage). *$\hat{\boldsymbol{\beta}}_{n,i} - \boldsymbol{\beta}_{0,i}$ converges in probability to zero, uniformly in i, in the sense that:* $\max_i \|\hat{\boldsymbol{\beta}}_{n,i} - \boldsymbol{\beta}_{0,i}\| = o_p(1)$.

Assumption 2.2 (Regularity Conditions)**.**

(i) For all $\boldsymbol{\theta}$, $q(\boldsymbol{\theta},\ y,\ \boldsymbol{x},\ \boldsymbol{b})$ is a measurable function of $(y,\ \boldsymbol{x}',\ \boldsymbol{b}')'$ in the space $\mathcal{Y} \times \mathcal{X} \times \mathcal{B}$.

(ii) For all $(y,\ \boldsymbol{x}',\ \boldsymbol{b}')' \in \mathcal{Y} \times \mathcal{X} \times \mathcal{B}$, $q(\boldsymbol{\theta},\ y,\ \boldsymbol{x},\ \boldsymbol{b})$ is twice continuously differentiable in $\boldsymbol{\theta}$ in the space $\boldsymbol{\Theta}$.

Assumption 2.1 is a common requirement when dealing with a first-stage estimator that may be infinite-dimensional (see Chen et al., 2003; Ichimura and Lee, 2010). The condition will hold in many applications. For the case where $\boldsymbol{\beta}_{0,i} = f(\boldsymbol{z}_i, \boldsymbol{\gamma}_0)$, Assumption 2.1 requires $\hat{\boldsymbol{\gamma}}_n$ to be a consistent estimator and $f(\boldsymbol{z}_i, \boldsymbol{\gamma})$ to be continuously differentiable in $\boldsymbol{\gamma}$, with bounded derivative uniformly in i.[1] For nonparametric sieve estimators, flexible regularity conditions can be imposed to obtain Assumption 2.1 (see Belloni et al., 2015). Certain of these conditions are discussed by Cattaneo et al. (2019) in their online appendix. Assumption 2.1 also holds in the case where $\boldsymbol{\beta}_{0,i}$ represents fixed effects from some first-stage modeling (e.g., Dzemski, 2019; Yan et al., 2019).

Assumption 2.2 sets regularity conditions on the objective function's behavior. These conditions

[1] The result follows from the mean value theorem: $\hat{\boldsymbol{\beta}}_{n,i} - \boldsymbol{\beta}_{0,i} = \partial_{\boldsymbol{\gamma}'} f(\boldsymbol{z}_i, \boldsymbol{\gamma}^+)(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)$, for some $\boldsymbol{\gamma}^+$ that lies between $\hat{\boldsymbol{\gamma}}_n$ and $\boldsymbol{\gamma}_0$.

are generally imposed for classical M-estimators and do not involve the first-stage estimator (see Amemiya, 1985). Our approach requires $\hat{\boldsymbol{\theta}}_n$ to be a consistent estimator. We acknowledge this as a high-level assumption.

Assumption 2.3 (Consistency). *$\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$.*

Assumption 2.3 is generally verified even when the first-stage estimator is asymptotically infinite dimensional (e.g., see Chen et al., 2003; Cattaneo et al., 2019). The proof of this consistency is context-dependent, and the required conditions may vary. In Online Appendix (OA) S.1.1, we present primitive conditions for Assumption 2.3. We adapt Theorem 4.1.1 of Amemiya (1985) to our framework by accommodating a wide range of first-stage estimators.

3 Overview of our Approach

Before delving into the theory behind our approach and presenting formal results, this section provides an overview using an illustrative example with a latent variable model.

Example 3.1 (Latent variable model). *We consider the following model:*

$$y_i = \theta_0 \beta_{0,i} + \varepsilon_i, \quad \beta_{0,i} = f(\boldsymbol{z}_i, \boldsymbol{\gamma}_0) = \boldsymbol{z}_i' \boldsymbol{\gamma}_0, \quad d_i = \mathbb{1}\{\beta_{0,i} > v_i\}, \quad v_i \sim \textit{Uniform}(0,\ 1),$$

where $\boldsymbol{\gamma}_0$ is an unknown parameter, θ_0 is the parameter of interest, $\beta_{0,i}$ is an unobserved probability, and ε_i's are independent and identically distributed (i.i.d) random errors with mean zero and variance $\sigma^2_{0,\varepsilon}$. Assume that we observe an i.i.d. sample of $(y_i,\ d_i,\ \boldsymbol{z}_i')'$, for $i = 1,\ \ldots,\ n$. We can use a two-stage approach to estimate θ_0. In the first stage, we estimate $\boldsymbol{\gamma}_0$ by regressing d_i on $\boldsymbol{z}_i$. Let $\hat{\boldsymbol{\gamma}}_n$ be the ordinary least squares (OLS) estimator of $\boldsymbol{\gamma}_0$. In the second stage, we estimate θ_0 using the regression of y_i on $\hat{\beta}_{n,i} = \boldsymbol{z}_i' \hat{\boldsymbol{\gamma}}_n$.

The objective function to be maximized in the second stage is $Q_n(\theta,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n) = -\frac{1}{n}\sum_{i=1}^n (y_i - \theta \hat{\beta}_{n,i})^2$, where $\hat{\mathbf{B}}_n = (\hat{\beta}_{n,1},\ \ldots,\ \hat{\beta}_{n,n})'$. The first-order condition of this maximization is $\frac{2}{n}\sum_{i=1}^n (y_i - \hat{\theta}_n \hat{\beta}_{n,i})\hat{\beta}_{n,i} = 0$, where $\hat{\theta}_n$ is the estimator of θ_0. By the mean value theorem, this condition solves to $\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{A}_n^{-1} \dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)$, where

$$\hat{A}_n = \tfrac{2}{n}\textstyle\sum_{i=1}^n \hat{\beta}^2_{n,i} \quad \textit{and} \quad \dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n) = \tfrac{2}{\sqrt{n}}\textstyle\sum_{i=1}^n (y_i - \theta_0 \hat{\beta}_{n,i})\hat{\beta}_{n,i}.$$

We will refer to $\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ as the influence function (IF). Assume that $A_0 = \operatorname{plim} \hat{A}_n$ exists and that the IF has an asymptotic variance denoted by Σ_0. Consequently, the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is given by $A_0^{-1}\Sigma_0 A_0^{-1}$. A consistent estimator of A_0 is $\hat{A}_n$. However, in the general case, it is not always straightforward to construct a consistent estimator for Σ_0 because one needs to account for the sampling error of $\hat{\boldsymbol{\gamma}}_n$.

For the sake of simplicity, we treat $\boldsymbol{z}_i$ as a nonstochastic variable. We define the conditional expectation and conditional variable of the IF, given $\hat{\mathbf{B}}_n$, as follows:

$$\begin{aligned} \mathcal{E}_n &:= \mathbb{E}(\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n) = \tfrac{2}{\sqrt{n}}\theta_0 \textstyle\sum_{i=1}^n (\beta_{0,i} - \hat{\beta}_{n,i})\hat{\beta}_{n,i}, \\ V_n &:= \mathbb{V}(\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n) = \tfrac{4}{n}\sigma^2_{0,\varepsilon} \textstyle\sum_{i=1}^n \hat{\beta}^2_{n,i}. \end{aligned} \tag{2}$$

Using the law of iterated variances, we have $\mathbb{V}(\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)) = \mathbb{E}(V_n) + \mathbb{V}(\mathcal{E}_n)$. *As* V_n *converges in probability to some nonstochastic quantity* $V_0 = 4\sigma^2_{0,\varepsilon}\boldsymbol{\gamma}_0' \lim(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i'/n)\boldsymbol{\gamma}_0$, *we show that*

$$\Sigma_0 = \lim \mathbb{V}(\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n)) = V_0 + \lim \mathbb{V}(\mathcal{E}_n). \tag{3}$$

Equation (3), disentangles the sampling errors from both stages. V_0 *accounts for the sampling error due to the error term* ε_i *in the second stage, whereas* $\lim \mathbb{V}(\mathcal{E}_n)$ *captures the variability that originates from* $\hat{\boldsymbol{\gamma}}_n$.

For some large integer κ, *imagine we can generate the variables* $\mathcal{E}_{n,1}, \ldots, \mathcal{E}_{n,\kappa}$, *that are i.i.d. as* $\mathcal{E}_n$. *By the Law of Large Numbers (LLN), we can estimate* $\mathbb{V}(\mathcal{E}_n)$ *using the empirical variance of* $\mathcal{E}_{n,1}, \ldots, \mathcal{E}_{n,\kappa}$. *Unfortunately, we cannot obtain such variables since the finite sample distribution of* $\mathcal{E}_n$ *is unknown. However, the good news is that we can estimate the asymptotic distribution of* $\mathcal{E}_n$. *Consequently, we can construct empirical analogs for* $\mathcal{E}_{n,1}, \ldots, \mathcal{E}_{n,\kappa}$ *using simulations from an estimator of the asymptotic distribution of* $\mathcal{E}_n$.

Given that the first stage is an OLS regression, the estimator of the asymptotic distribution of $\hat{\boldsymbol{\gamma}}_n$ *is a normal distribution with mean* $\hat{\boldsymbol{\gamma}}_n$ *and variance* $\hat{\mathbb{V}}(\hat{\boldsymbol{\gamma}}_n) = (\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i')^{-1}(\sum_{i=1}^n \hat{\nu}_i^2\boldsymbol{z}_i\boldsymbol{z}_i')(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i')^{-1}$, *where* $\hat{\nu}_i = d_i - \boldsymbol{z}_i'\hat{\boldsymbol{\gamma}}_n$. *For* $s = 1,\ \ldots,\ \kappa$, *let* $\bar{\beta}_{n,i}^{(s)} = \boldsymbol{z}_i'\bar{\boldsymbol{\gamma}}_n^{(s)}$, *where* $\bar{\boldsymbol{\gamma}}_n^{(s)} \sim N(\hat{\boldsymbol{\gamma}}_n,\ \hat{\mathbb{V}}(\hat{\boldsymbol{\gamma}}_n))$. *We define* $\hat{\mathcal{E}}_{n,s} = \frac{2\hat{\theta}_n}{\sqrt{n}}\sum_{i=1}^n(\hat{\beta}_{n,i} - \bar{\beta}_{n,i}^{(s)})\bar{\beta}_{n,i}^{(s)}$ *and* $\hat{V}_n = \frac{4}{n}\hat{\sigma}^2_{n,\varepsilon}\sum_{i=1}^n \hat{\beta}^2_{n,i}$, *where* $\hat{\sigma}^2_{n,\varepsilon}$ *is the estimator of* $\sigma^2_{0,\varepsilon}$. *We show that a consistent estimator of the asymptotic variance of* $\hat{\theta}_n$ *is:*

$$\frac{\hat{V}_n + \hat{\mathbb{V}}(\mathcal{E}_n)}{n\hat{A}_n^2}, \quad \textit{where} \quad \hat{\mathbb{V}}(\mathcal{E}_n) = \frac{1}{\kappa - 1}\sum_{s=1}^{\kappa}(\hat{\mathcal{E}}_{n,s} - \hat{\mathbb{E}}(\mathcal{E}_n))^2 \quad \textit{and} \quad \hat{\mathbb{E}}(\mathcal{E}_n) = \frac{1}{\kappa}\sum_{s=1}^{\kappa}\hat{\mathcal{E}}_{n,s}. \tag{4}$$

We further use the idea of separating the sampling errors between the first and second stages to approximate the asymptotic CDF of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. *We define the standardized IF, conditional on* $\hat{\mathbf{B}}_n$, *as follows:*

$$u_n := V_n^{-\frac{1}{2}}(\dot{q}_n(\theta_0,\ \mathbf{y}_n,\ \hat{\mathbf{B}}_n) - \mathcal{E}_n) = \sum_{i=1}^n \hat{a}_{n,i}(y_i - \theta_0\beta_{0,i}), \quad \textit{where} \quad \hat{a}_{n,i} = \sigma_{0,\varepsilon}^{-1}\hat{\beta}_{n,i}\big(\textstyle\sum_{i=1}^n \hat{\beta}^2_{n,i}\big)^{-\frac{1}{2}}.$$

The expectation of u_n *is zero and the variance is one. For* $\hat{\mathbf{B}}_n$ *set fixed as a predetermined sequence in* n, *the variables* $a_{n,i}$*'s are nonstochastic and* $\sum_{i=1}^n \hat{a}_{n,i}(y_i - \theta_0\beta_{0,i})$ *is a sum of independent variables. Consequently, by a conditional central limit theorem (CLT), the conditional distribution of* u_n, *given* $\hat{\mathbf{B}}_n$, *converges to* $N(0,\ 1)$, *for almost all* $\hat{\mathbf{B}}_n$.[2] *Since* $\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{A}_n^{-1}(V_n^{1/2}u_n - \mathcal{E}_n)$, *we show that the unconditional asymptotic distribution of* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *can be approximated by the empirical CDF of the sample:*

$$\hat{\psi}_{n,s} = \hat{A}_n^{-1}(\hat{V}_n^{1/2}\zeta_s + \hat{\mathcal{E}}_{n,s}), \quad s = 1,\ \ldots,\ \kappa,$$

where $\zeta_1,\ \ldots,\ \zeta_\kappa$ *are* κ *independent draws from* $N(0,\ 1)$. *The* 2.5% *and* 97.5% *empirical quantiles of the sample* $\{\hat{\theta}_n - \hat{\psi}_{n,s}/\sqrt{n},\ s = 1,\ \ldots,\ \kappa\}$ *are the bounds of the* 95% *confidence interval (CI) of* θ_0.

Figure 1 depicts average estimates of the asymptotic CDF of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *for 10,000 replications, where* $\theta_0 = 1$, $\varepsilon_i \sim$ *Uniform*$[-1,\ 1]$, $\boldsymbol{z}_i = (1, z_{i,2})'$, $z_{i,2} \sim$ *Uniform*$[0,\ 1]$, *and* $\boldsymbol{\gamma}_0 = (0.1,\ 0.8)'$. *The sample size takes the values* $1,000$ *and* $2,000$. *The black line (*F_0*) represents the actual CDF of the sample comprising the 10,000 replications of* $\sqrt{n}(\hat{\theta}_n - \theta_0)$. *We use this CDF as the benchmark against which we compare our estimates.*

[2] See an example of conditional CLT in Rubshtein (1996). Indeed, Lyapunov's condition is verified if $\sum_{i=1}^n \hat{a}_{n,i}^{2+\nu}\mathbb{E}(|\varepsilon_i|^{2+\nu}) = o_p(1)$, for some $\nu > 0$. A similar condition is also required in the case where $\beta_{0,i}$ is known and θ_0 is estimated using a single-step approach.

The dotted blue line ($\hat{H}_1$) is the average Gaussian estimate of the CDF (assuming asymptotic normality in the second stage) when we disregard the first-stage sampling error. The dashed blue line ($\hat{H}_2$) represents the average Gaussian estimate when the asymptotic variance is estimated using our simulation approach; i.e., asymptotic variance is estimated by $\hat{A}_n^{-1}(\hat{V}_n + \hat{\mathbb{V}}(\mathcal{E}_n))\hat{A}_n^{-1}$. The dashed red line ($\hat{F}_n$) corresponds to the average estimate using our simulation approach; i.e., the average empirical CDF of the sample $\{\hat{\psi}_{n,s},\ s = 1, \ldots, \kappa\}$. We also enclose in parentheses the L_1-Wasserstein distance between each CDF estimate and the actual CDF F_0.[3]

By comparing the distances between each estimate and the true CDF, $\hat{H}_1$ seems biased given that it overlooks the first-stage sampling error. The Gaussian approximation, accounting for the first-stage sampling error, and our simulation approach yield strong performance. The Gaussian approximation slightly fits F_0 better. This is not surprising because in this simple case of OLS estimations at both steps, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is normally distributed asymptotically with a zero mean (e.g., see Murphy and Topel, 2002).

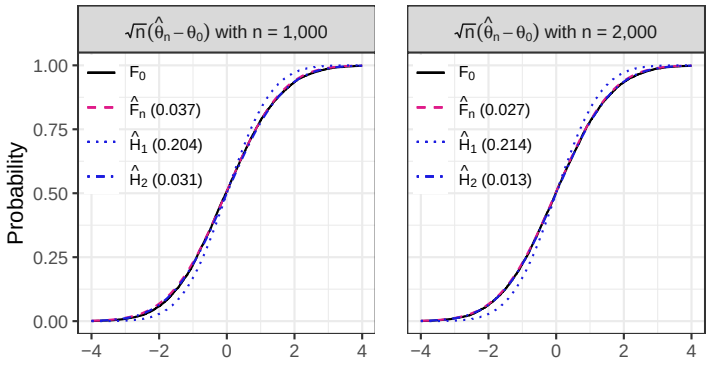

Figure 1: Illustration of the asymptotic CDF estimates

This figure displays the average estimates of the CDF of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. The L_1-Wasserstein distance between each estimated CDF and the true sampling CDF, F_0, is enclosed in parentheses.

A key ingredient of our approach lies in computing the conditional variance of the IF. One important simplification in the above example is that $\boldsymbol{y}_n$ is independent of the first-stage estimator. This is employed when computing $\mathcal{E}_n$ and V_n in (2). In a more general context, computing the conditional moments of the IF may be challenging. We will later discuss this situation in Section 4.1.2. We argue that one can disregard the dependence between the first-stage estimator and $\boldsymbol{y}_n$ because the first-stage estimator converges in probability to a constant.

Moreover, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normally distributed with a zero mean because $\mathcal{E}_n$ is also asymptotically normally distributed with a zero mean. Yet, our approach does not require this restriction. The limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is centered at $A_0^{-1} \lim \mathbb{E}(\mathcal{E}_n)$. If $\lim \mathbb{E}(\mathcal{E}_n)$ is not zero, the plug-in estimator can exhibit significant bias in finite samples. We address this issue by

[3]The L_1-Wasserstein distance between a CDF $\hat{R}_n$ and F_0 is $\|\hat{R}_n - F_0\|_w = \int_{\mathbb{R}} |\hat{R}_n(t) - F_0(t)| dt$.

proposing the debias estimator $\theta^*_{n,\kappa} = \hat{\theta}_n - (\sqrt{n}\hat{A}_n)^{-1}\hat{\mathbb{E}}(\mathcal{E}_n)$. We demonstrate the limiting distribution of $\sqrt{n}(\theta^*_{n,\kappa} - \theta_0)$ has a zero mean.

4 Inference for Conditional Extremum Estimators

We present our main results in this section. Technical details of proofs can be found in Appendix A. The first-order condition of the maximization of (1) is $\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}} q(\hat{\boldsymbol{\theta}}_n,\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i}) = 0$. By applying the mean value theorem to $\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}} q(\boldsymbol{\theta},\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i})$, we obtain:

$$\Delta_n := \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_n^{-1}(1/\sqrt{n})\sum_{i=1}^n \dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i}), \tag{5}$$

where $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i}) = \partial_{\boldsymbol{\theta}} q(\boldsymbol{\theta}_0,\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i})$ and $\mathbf{A}_n = -\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}^+_n,\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i})$, for some $\boldsymbol{\theta}^+_n$ that lies between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. Given that $\operatorname{plim}\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0$, we also have $\operatorname{plim}\boldsymbol{\theta}^+_n = \boldsymbol{\theta}_0$. In large samples, $\mathbf{A}_n$ is assumed to be nonsingular (see Assumption 4.2). Let $\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) = (1/\sqrt{n})\sum_{i=1}^n \dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$. We will refer to $\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ as the *influence function* (IF).

In the case of a single-step estimator, the central limit theorem (CLT) implies (under regularity conditions) that the IF is asymptotically normally distributed with zero mean (see Amemiya, 1985, Theore 4.1.3). A crucial condition that is required by the CLT is that the dependence among the variables $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$'s is "weak". Roughly speaking, if we define a certain order between the subscripts i's (e.g., if i is time), the correlation between $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$ and $\dot{\boldsymbol{q}}_{n,j}(y_j,\ \hat{\boldsymbol{\beta}}_{n,j})$ must vanish at a certain rate as $|i-j|$ grows to infinity (see Withers, 1981; Romano and Wolf, 2000; Ekström, 2014). For two-stage estimators, the variables $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$'s are dependent on each other because they all depend on the same first-stage estimator. Consequently, the weak dependence condition does not hold in general, even though $\hat{\boldsymbol{\beta}}_{n,i}$ uniformly converges in probability to $\boldsymbol{\beta}_{0,i}$. Without imposing additional conditions, there is no general CLT that guarantees asymptotic normality in this case.

Our approach does not require asymptotic normality for the IF. Instead, we will impose that the conditional distribution of the IF, given $\hat{\mathbf{B}}_n$, is asymptotically normal. We will later argue why such a condition is weak and would hold in many contexts (see Section 4.2).

We introduce the following regularity assumptions.

Assumption 4.1 (Influence Function)**.**
(i) $\mu_\nu(\hat{\mathbf{B}}_n) = \mathbb{E}(\|\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\|^\nu|\hat{\mathbf{B}}_n)$ *and* $\mathbb{E}(\mu_\nu(\hat{\mathbf{B}}_n))$ *exist for some* $\nu > 2$*, with* $\mathbb{E}(\mu_\nu(\hat{\mathbf{B}}_n))$ *bounded.*
(ii) $\mathbf{V}_n := \mathbb{V}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n)$ *converges in probability to some nonstochastic quantity* $\mathbf{V}_0$ *and* $\mathcal{E}_n := \mathbb{E}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n)$ *converges in distribution to some random variable* $\mathcal{E}_0$*.*

Assumption 4.2 (Hessian Matrix)**.** *For any estimator* $\boldsymbol{\theta}^+_n$ *such that* $\operatorname{plim}\boldsymbol{\theta}^+_n = \boldsymbol{\theta}_0$*, the Hessian of the objective function at* $\boldsymbol{\theta}^+_n$*, given by* $\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}^+_n,\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i})$*, converges in probability to a finite nonsingular matrix* $\mathbf{A}_0 = \lim \mathbb{E}\big(\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} q(\boldsymbol{\theta}_0,\ y_i,\ \boldsymbol{x}_i,\ \boldsymbol{\beta}_{0,i})\big)$*.*

Assumptions 4.1 and 4.2 introduce weak regularity requirements. Condition (i) of Assumption 4.1 implies that the second conditional and unconditional moments of the IF exist and are bounded. We impose this condition for the asymptotic variance of the IF to be finite. Condition (ii) and some other regularity requirements that we will later introduce allow for the IF function to have an asymptotic distribution. This condition will hold in general because $\mathbf{V}_n$ can be expressed as a sampling mean, whereas $\mathcal{E}_n$ is a sum of n random variables divided by $\sqrt{n}$.

By the Law of Large Numbers (LLN), if $\mathbf{V}_n$ is smooth in $\hat{\mathbf{B}}_n$, then it will converge in probability to a constant because the first-stage estimator is consistent (see Example 3.1). The existence of a limiting distribution for $\mathcal{E}_n$ prevents $\mathcal{E}_n$ from asymptotically exploding. In many cases, it would be possible to write $\mathcal{E}_n$ as a function of $\mathbf{C}_n(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{b}_n)$, for some sequences $\mathbf{C}_n$ and $\boldsymbol{b}_n$ that converge in probability to nonstochastic quantities and some estimator $\hat{\boldsymbol{\gamma}}_n$ such that $\mathbf{C}_n(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{b}_n)$ has a limiting distribution. In Example 3.1, $\mathcal{E}_n$ can be approximated using a first-order Taylor expansion around $\boldsymbol{\gamma}_0$ as $\mathcal{E}_n \approx -(\frac{2\theta_0}{n}\sum_{i=1}^n \boldsymbol{z}_i'\boldsymbol{\gamma}_0\boldsymbol{z}_i')\sqrt{n}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)$. Consequently, $\mathcal{E}_0$ is normally distributed with a zero mean. Note that Condition (ii) is weaker than the assumption that is generally imposed in the literature. Specifically, we do not require $\mathcal{E}_0$ to be normally distributed, a situation that can occur when the first-stage estimator is not normally distributed.

Assumption 4.2 ensures the consistency of the Hessian matrix. Under some weak regularity conditions, for instance, $\partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'}q(\boldsymbol{\theta}_0,\ y_i,\ \boldsymbol{x}_i,\ \boldsymbol{\beta}_{0,i})$ is ergodic stationary across i with a finite variance, the LLN implies that $\operatorname{plim}\frac{1}{n}\sum_{i=1}^n \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'}q(\boldsymbol{\theta}_0,\ y_i,\ \boldsymbol{x}_i,\ \boldsymbol{\beta}_{0,i}) = \mathbf{A}_0$. Assumption 4.2 imposes this convergence when $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_{0,i}$ are replaced with consistent estimators. It extends Conditions (B) in Theorem 4.1.3 of Amemiya (1985) to two-stage estimation approaches. We discuss primitive conditions for Assumption 4.2 in OA S.1.2. These conditions require the Hessian at $\boldsymbol{\theta}_n^+$ to be smooth in $\boldsymbol{\theta}_n^+$ and $\hat{\boldsymbol{\beta}}_{n,i}$.

4.1 Asymptotic Variance of Plug-in Estimators

4.1.1 Expression of the Asymptotic Variance

In this section, we present an expression of the asymptotic variance that can be easily approximated. Let $\boldsymbol{\Sigma}_n = \mathbb{V}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n))$, $\boldsymbol{\Sigma}_0 = \lim \boldsymbol{\Sigma}_n$, and $\mathbf{A}_0 = \operatorname{plim} \mathbf{A}_n$. Assumptions 4.1 and 4.2 do not ensure that $\Delta_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ has a limiting distribution. For now, we assume that such a limiting distribution exists. The sufficient condition for this assumption is imposed later in Assumption 4.4. The asymptotic variance of Δ_n is given by $\mathbb{V}(\Delta_0) := \mathbf{A}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{A}_0^{-1}$. This expression is similar to the asymptotic variance formula for single-step M-estimators. Yet, a notable difference here is that the sampling error from the first-stage estimator is incorporated into $\boldsymbol{\Sigma}_0$. The sources of variability of $\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ are the data $\{\mathbf{y}_n,\ \mathbf{X}_n\}$ and the first-stage estimator $\hat{\mathbf{B}}_n$. The model error in the second stage is captured by $\{\mathbf{y}_n,\ \mathbf{X}_n\}$, whereas $\hat{\mathbf{B}}_n$ accounts for the sampling error of the first-stage estimator.

Estimating the asymptotic variance of Δ_n requires consistent estimators of $\mathbf{A}_0$ and $\mathbf{\Sigma}_0$. By Assumption 4.2, since $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$, a consistent estimator for $\mathbf{A}_0$ can simply be $\hat{\mathbf{A}}_n := -\frac{1}{n}\sum_{i=1}^{n} \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} q(\hat{\boldsymbol{\theta}}_n,\ y_i,\ \boldsymbol{x}_i,\ \hat{\boldsymbol{\beta}}_{n,i})$. To construct a consistent estimator for $\mathbf{\Sigma}_0$, we rely on the law of iterated variances to disentangle the sampling error of the first stage from the model error in the second stage. We demonstrate that this approach leads to an expression that makes it easier to consistently estimate $\mathbf{\Sigma}_0$.

By the law of iterated variances, we have $\mathbf{\Sigma}_n = \mathbb{E}(\mathbf{V}_n) + \mathbb{V}(\mathcal{E}_n)$. As $\mathbf{V}_n$ is a semi-positive definite matrix with bounded expectation (Assumptions 4.1), it is also uniformly integrable. Therefore, Assumption 4.1 implies that $\mathbb{E}(\mathbf{V}_n)$ converges to $\mathbf{V}_0$ (see Lebesgue–Vitali theorem in Bogachev and Ruas, 2007, Theorem 4.5.4). Moreover, since $\mathcal{E}_n$ converges in distribution to $\mathcal{E}_0$ and $\mathbb{E}(\|\mathcal{E}_n\|^{\nu}) < \infty$ for some $\nu > 2$, if follows that $\mathbb{V}(\mathcal{E}_n)$ converges to $\mathbb{V}(\mathcal{E}_0)$ (see Chung, 2001, Theorem 4.5.2). As a result,

$$\mathbf{\Sigma}_0 = \mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0). \tag{6}$$

The first term in the right-hand side (RHS) of Equation (6) is similar to the variance of the IF in the case of standard M-estimators. This variance is only due to the second stage, as it is the limit of the conditional variance, given $\hat{\mathbf{B}}_n$. The second term is the variance that is due to the first-stage estimation because the only source of uncertainty in $\mathcal{E}_n$ is $\hat{\mathbf{B}}_n$. Ignoring this second term results in a downward biased estimate of $\mathbf{\Sigma}_0$.

For some integer $\kappa \geqslant 1$, let $\mathbf{B}_{n,1},\ \ldots,\ \mathbf{B}_{n,\kappa}$ be κ variables that are independent and identically distributed (i.i.d) as $\hat{\mathbf{B}}_n$. For any integer $s \in [1,\ \kappa]$, let $\mathcal{E}_{n,s} = \mathbb{E}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \mathbf{B}_{n,s})|\mathbf{B}_{n,s})$, i.e., we define $\mathcal{E}_{n,s}$ by replacing $\hat{\mathbf{B}}_n$ in $\mathcal{E}_n$ with $\mathbf{B}_{n,s}$. Consequently, $\mathcal{E}_n$ and $\mathcal{E}_{n,s}$ are i.i.d as well. The following theorem provides the exact expression of the asymptotic variance of Δ_n.

Theorem 4.1 (Asymptotic Variance)**.** *Let* $\mathbf{\Omega}_n^{\kappa} = \frac{1}{\kappa}\sum_{s=1}^{\kappa} \mathcal{E}_{n,s}$ *and* $\mathbf{\Sigma}_n^{\kappa} = \mathbf{V}_n + \frac{1}{\kappa-1}\sum_{s=1}^{\kappa}(\mathcal{E}_{n,s} - \mathbf{\Omega}_n^{\kappa})(\mathcal{E}_{n,s} - \mathbf{\Omega}_n^{\kappa})'$*. Under Assumptions 2.1–4.2, we have* $\operatorname{plim}_{n,\kappa} \mathbf{\Sigma}_n^{\kappa} = \mathbf{\Sigma}_0$ *and* $\mathbb{V}(\Delta_0) = \mathbf{A}_0^{-1}\left(\operatorname{plim}_{n,\kappa} \mathbf{\Sigma}_n^{\kappa}\right)\mathbf{A}_0^{-1}$.

Proof. The formal proof is presented in Appendix A.1. As $\mathcal{E}_{n,s}$ is i.i.d across s, the LLN implies that $\frac{1}{\kappa-1}\sum_{s=1}^{\kappa}(\mathcal{E}_{n,s} - \mathbf{\Omega}_n^{\kappa})(\mathcal{E}_{n,s} - \mathbf{\Omega}_n^{\kappa})'$ converges in probability to $\mathbb{V}(\mathcal{E}_n)$, as κ grows to infinity. □

The expression of the asymptotic variance given by Theorem 4.1 is the exact asymptotic variance. In the next section, we discuss how to obtain a finite sample approximation.

4.1.2 Finite Sample Approximation of the Asymptotic Variance

Theorem 4.1 suggests that a consistent estimator of the asymptotic variance $\mathbb{V}(\Delta_0) = \mathbf{A}_0^{-1}\mathbf{\Sigma}_0\mathbf{A}_0^{-1}$ can be obtained by replacing $\mathbf{A}_0$ with $\hat{\mathbf{A}}_n$ and $\mathbf{\Sigma}_0$ with a proxy of $\mathbf{\Sigma}_n^{\kappa}$. The expression of $\mathbf{\Sigma}_n^{\kappa}$ is a function of $\boldsymbol{\theta}_0$, $\mathbf{B}_0$, $\hat{\mathbf{B}}_n$, and $\mathbf{B}_{n,s}$. This is because $\mathcal{E}_n$ and $\mathbf{V}_n$ depend on $\boldsymbol{\theta}_0$, $\mathbf{B}_0$, and $\hat{\mathbf{B}}_n$, and $\mathcal{E}_{n,s}$ depends $\boldsymbol{\theta}_0$, $\mathbf{B}_0$, and $\mathbf{B}_{n,s}$ (see Example 3.1). To construct a proxy for $\mathbf{\Sigma}_n^{\kappa}$, we can replace $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ with their

respective estimators $\hat{\boldsymbol{\theta}}_n$ and $\hat{\mathbf{B}}_n$. However, this is not sufficient given that $\boldsymbol{\Sigma}_n^{\kappa}$ also depends on $\mathbf{B}_{n,s}$. The variable $\mathbf{B}_{n,s}$ follows the *true* finite sample distribution of $\hat{\mathbf{B}}_n$ which is unknown.

Before dealing with this issue, we first discuss how $\mathcal{E}_n$ and $\mathbf{V}_n$ can be computed as functions of $\boldsymbol{\theta}_0$, $\mathbf{B}_0$, and $\hat{\mathbf{B}}_n$. We can encounter diverse situations in practice. The simplest one is when $\hat{\mathbf{B}}_n$ is independent of $\mathbf{y}_n$ (see Example 3.1). In such cases, it is possible to compute $\mathbf{V}_n$ and $\mathcal{E}_n$ analytically by generally substituting $\mathbf{y}_n$ with its specification. Frameworks that fall within this consideration include instances where the first-stage model is disconnected from the second-stage model (e.g., Breza et al., 2020; Lubold et al., 2023; Boucher and Houndetoungan, 2023). Conditioning on $\hat{\mathbf{B}}_n$ in $\mathbb{V}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n)$ and $\mathbb{E}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)|\hat{\mathbf{B}}_n)$ simply results in treating $\hat{\mathbf{B}}_n$ as a constant. Importantly, this conditioning does not alter the distribution of $\mathbf{y}_n$.

The second more challenging situation is when $\hat{\mathbf{B}}_n$ and $\mathbf{y}_n$ are not independent. This situation arises in IV approaches and when $\mathbf{y}_n$ is used in the first stage (e.g., see Dufays et al., 2022). For large samples, we can obtain approximations of the conditional moments of the IF by assuming that $\hat{\mathbf{B}}_n$ and $\mathbf{y}_n$ are independent. We can impose this assumption because $\hat{\boldsymbol{\beta}}_{n,i}$ converges in probability to a nonstochastic quantity. The convergence makes $\hat{\boldsymbol{\beta}}_{n,i}$ asymptotically independent of unobserved factors in the first stage, such as error terms, which could be correlated with $\mathbf{y}_n$. Assuming that $\hat{\mathbf{B}}_n$ and $\mathbf{y}_n$ are independent makes it possible to compute $\mathbf{V}_n$ and $\mathcal{E}_n$ (or at least obtain a numerical approximation). See examples in our simulation study in Section 5.

When $\mathbf{V}_n$ and $\mathcal{E}_n$ do not have a closed-form expression, particularly for certain nonlinear models, we can find large-sample approximations by using empirical definitions. For instance, $\mathcal{E}_n$ can be approximated by $(1/\sqrt{n})\sum_{i=1}^{n}\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$ evaluated at $\hat{\boldsymbol{\theta}}_n$. Furthermore, the empirical variance can be used for $\mathbf{V}_n$. If $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$'s are not independent across i, we can estimate $\mathbf{V}_n$ using a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix (Andrews, 1991), as is the case in a single-step M-estimation. See data-generating process (DGP) D in our simulation study.

The analytical expression (or numerical approximation) of $\mathcal{E}_{n,s}$ is the same as that of $\mathcal{E}_n$ by replacing $\hat{\mathbf{B}}_n$ with $\mathbf{B}_{n,s}$. Since we generally have an asymptotic characterization of the distribution of the first-stage estimator, but not the true distribution, we impose the following assumption.

Assumption 4.3 (Asymptotic Distribution of the First-Stage Estimator)**.** *The practitioner can simulate realizations from a consistent estimator of the asymptotic distribution of* $\hat{\mathbf{B}}_n$.

Assumption (4.3) requires the practitioner to possess a consistent estimator of the joint asymptotic distribution of $\hat{\boldsymbol{\beta}}_{n,1},\ \ldots,\ \hat{\boldsymbol{\beta}}_{n,n}$. This distribution can be obtained for a large class of models. For first-stage estimators of type $\hat{\boldsymbol{\beta}}_{n,i} = f(\boldsymbol{z}_i, \hat{\boldsymbol{\gamma}}_n)$, which encompass nonparametric methods, a simulation from an estimator of the asymptotic distribution of $\hat{\mathbf{B}}_n$ is $\bar{\mathbf{B}}_n = (\bar{\boldsymbol{\beta}}_{n,1},\ \ldots,\ \bar{\boldsymbol{\beta}}_{n,n})'$, where $\bar{\boldsymbol{\beta}}_{n,i} = f(\boldsymbol{z}_i, \bar{\boldsymbol{\gamma}}_n)$ and $\bar{\boldsymbol{\gamma}}_n$ is simulated from an estimator of the asymptotic distribution of $\hat{\boldsymbol{\gamma}}_n$. From a frequentist perspec-

tive, the estimator of the asymptotic distribution of $\hat{\gamma}_n$ is typically derived through an asymptotic analysis (e.g., a normal distribution centered at $\hat{\gamma}_n$ with some covariance matrix). In the Bayesian paradigm, we can assume that the posterior distribution that is obtained from a Gibbs sampler or Metropolis-Hastings is a valid estimator (see Casella and George, 1992; Chib and Greenberg, 1995).

Let $\bar{\mathbf{B}}_{n,1}, \ldots, \bar{\mathbf{B}}_{n,\kappa}$ be independent variables that are simulated from the estimator of the distribution of $\hat{\mathbf{B}}_n$. To construct a proxy for $\mathbf{\Sigma}_n^\kappa$, we replace $\mathbf{B}_{n,s}$ with the simulations $\bar{\mathbf{B}}_{n,s}$. Let $\hat{\mathbf{V}}_n$ be the variable that is obtained by replacing $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ in $\mathbf{V}_n$ with $\hat{\boldsymbol{\theta}}_n$, $\hat{\mathbf{B}}_n$, respectively. Let also $\hat{\mathcal{E}}_{n,s}$ be the variables resulting from replacing $\boldsymbol{\theta}_0$, $\mathbf{B}_0$, and $\mathbf{B}_{n,s}$ in $\mathcal{E}_{n,s}$ with $\hat{\boldsymbol{\theta}}_n$, $\hat{\mathbf{B}}_n$, and $\bar{\mathbf{B}}_{n,s}$, respectively. An estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$ is given by:

$$\hat{\mathbb{V}}_n(\hat{\boldsymbol{\theta}}_n) = \frac{\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{\Sigma}}_n^\kappa\hat{\mathbf{A}}_n^{-1}}{n}, \tag{7}$$

where $\hat{\mathbf{\Sigma}}_n^\kappa = \hat{\mathbf{V}}_n + \frac{1}{\kappa-1}\sum_{s=1}^{\kappa}(\hat{\mathcal{E}}_{n,s} - \hat{\mathbf{\Omega}}_n^\kappa)(\hat{\mathcal{E}}_{n,s} - \hat{\mathbf{\Omega}}_n^\kappa)'$ and $\hat{\mathbf{\Omega}}_n^\kappa = \frac{1}{\kappa}\sum_{s=1}^{\kappa}\hat{\mathcal{E}}_{n,s}$.

We must discuss some necessary conditions for $n\hat{\mathbb{V}}_n(\hat{\boldsymbol{\theta}}_n)$ to be a consistent estimator of $\mathbb{V}(\Delta_0)$. First, replacing $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ in $\mathbf{V}_n$ and $\mathcal{E}_{n,s}$ with consistent estimators is a common practice for approximating quantities depending on unknown parameters. This requires both $\mathbf{V}_n$ and $\mathcal{E}_{n,s}$ to be smooth in $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$. For example, we also replace $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ with their estimators in Assumption 4.2 and discuss lower-level conditions for consistency in AO S.1.2.

Nevertheless, a more important concern in Equation (7) is that we simulate from an estimator of the asymptotic distribution of $\hat{\mathbf{B}}_n$ instead of the actual distribution, as required by Theorem 4.1. For the sake of simplicity, let us assume that $\mathcal{E}_{n,s}$ is smooth in $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$, so that we can overlook the problem raised by replacing $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ with their estimators. Now, let us focus on the implications of the use of the estimated distribution of $\hat{\mathbf{B}}_n$. Let $\tilde{\mathcal{E}}_{n,s}$ be the variable that is obtained by replacing $\mathbf{B}_{n,s}$ in $\mathcal{E}_{n,s}$ with $\bar{\mathbf{B}}_{n,s}$. The difference between this new variable and the former $\hat{\mathcal{E}}_{n,s}$ is that the new variable depends on $\boldsymbol{\theta}_0$ and $\mathbf{B}_0$, rather than on their empirical counterparts.

If we compute $\hat{\mathbf{\Sigma}}_n^\kappa$ using $\tilde{\mathcal{E}}_{n,s}$ instead of $\hat{\mathcal{E}}_{n,s}$, the resulting statistic can be a consistent estimator of $\mathbf{\Sigma}_0$ if $\hat{\mathcal{E}}_{n,s}$ and $\mathcal{E}_n$ are asymptotically identically distributed (a.i.d), and $\mathbb{E}(\|\tilde{\mathcal{E}}_{n,s}\|^\nu) = O(1)$ for some $\nu > 2$. The first condition is trivial since $\bar{\mathbf{B}}_{n,s}$ and $\mathbf{B}_{n,s}$ are a.i.d. The second condition is required to extend the convergence in distribution to the convergence in quadratic mean (see Chung, 2001, Theorem 4.5.2). We also impose the same condition for $\mathcal{E}_n$ in Assumption 4.1, Condition (i). This condition is likely to be satisfied in many contexts, including cases where the estimator of the asymptotic distribution is normal or a mixture of normals.

In most models, policymakers are interested in the estimates of marginal effects (MEs) or response functions and not directly in $\hat{\boldsymbol{\theta}}_n$. Our method can be combined with the Delta approach to compute standard errors of (nonlinear) smooth functions in $\hat{\boldsymbol{\theta}}_n$. We discuss this point in the following remark.

Remark 4.1 (Delta Method)**.** *MEs (or response functions) are often defined as $f(\mathbf{X}_n, \boldsymbol{\theta}_0)$, for some function*

f that is differentiable in $\boldsymbol{\theta}_0$. An estimator is then given by $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)$. The function f may be nonlinear in $\boldsymbol{\theta}_0$, especially for nonlinear models. The Delta method is generally used to estimate the asymptotic variance of $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)$. A first-order Taylor approximation of $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}_0$ implies that $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n) - f(\mathbf{X}_n,\ \boldsymbol{\theta}_0) \approx n\dot{f}(\mathbf{X}_n,\ \boldsymbol{\theta}_0)\Delta_n$, where $\dot{f}$ is the derivative of $f(\mathbf{X}_n,\ \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The asymptotic variance of $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)$ can be estimated by $\dot{f}(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)\hat{\mathbb{V}}_n(\hat{\boldsymbol{\theta}}_n)(\dot{f}(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n))'$, where $\hat{\mathbb{V}}_n(\hat{\boldsymbol{\theta}}_n)$ is the asymptotic variance of $\hat{\boldsymbol{\theta}}_n$ given by Equation (7).

Our approach is computationally attractive. In the first stage, we only need a single estimate of $\hat{\mathbf{B}}_n$ and an estimator of its distribution. In the second stage, we also need a single estimate of $\hat{\boldsymbol{\theta}}_n$. The integer κ must be set as large as possible for the estimator $\hat{\boldsymbol{\Sigma}}_n^{\kappa}$ to be efficient. In general, this will not raise a computational issue. Compared to the inference for single-step M-estimators, the additional computational routine required in our method consists only of approximating the variance of $\mathcal{E}_n$ using simulations from the distribution that was obtained at the first stage.

4.2 Asymptotic Distribution of Plug-in Estimators

We extend the simulation approach that was introduced in the preceding section to estimate the entire asymptotic distribution function of Δ_n. Recall that the reason why we cannot directly apply a CLT to the IF is that each variable $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$ depends on the first-stage estimator, thereby introducing dependencies among them. To address this problem, we first analyze the conditional distribution of the IF, given $\hat{\mathbf{B}}_n$. In fact, given $\hat{\mathbf{B}}_n$ set fixed as a predetermined sequence in n, we no longer encounter the issue that all the $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$'s are mutually dependent. This allows us to treat the problem as in the case of classical single-step M-estimators, where a CLT can be used under regularity conditions. Following that, we use a similar approach as for the case of the asymptotic variance to incorporate the sampling error of $\hat{\mathbf{B}}_n$ into the limiting conditional distribution of the IF.

Because of the first-stage estimator, the IF may not have a zero mean, even asymptotically, thereby leading to a limiting distribution of Δ_n that is not centered at zero (e.g., see Chernozhukov et al., 2018). Therefore, we define the *standardized* IF as $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) := \mathbf{V}_n^{-1/2}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) - \mathcal{E}_n)$, which has a zero mean and variance $\boldsymbol{I}_{K_\theta}$. We introduce the following assumptions.

Assumption 4.4 (Conditional Asymptotic Normality)**.** *The conditional distribution of the standardized influence function $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$, given $\hat{\mathbf{B}}_n$, converges in distribution to $N(0,\ \boldsymbol{I}_{K_\theta})$ almost surely; in the sense that for all $\boldsymbol{t} \in \mathbb{R}^{K_\theta}$, we have $\operatorname{plim} \mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n) = \Phi(\boldsymbol{t})$, where Φ is the CDF of $N(0,\ \boldsymbol{I}_{K_\theta})$.*

Assumption 4.4 requires the conditional distribution of the standardized IF, given $\hat{\mathbf{B}}_n$, to be asymptotically normal, for almost all $\hat{\mathbf{B}}_n$. This assumption is weak because the conditions for asymptotic normality generally hold when treating $\hat{\mathbf{B}}_n$ as nonstochastic, i.e., a predetermined sequence in n.[4] In

[4] A similar interpretation of Assumption 4.4 by Kato (2011) is that $\sup_{g \in LB} |\mathbb{E}[g(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n))|\hat{\mathbf{B}}_n] - \int_{\mathbb{R}} g(t)d\Phi(\boldsymbol{t})| = o_p(1)$,

this scenario, $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ can be viewed as the standardized IF of a single-step estimator and a conditional CLT can be applied (see Rubshtein, 1996, Theorem 1). For example, if y_i is independent across i, conditional on $\hat{\mathbf{B}}_n$, the variables $\dot{\boldsymbol{q}}_{n,i}(y_i,\ \hat{\boldsymbol{\beta}}_{n,i})$'s would also be independent across i. Consequently, the Lyapunov CLT or Lindeberg CLT can imply Assumption 4.4 under regularity conditions. When dealing with time series data where y_i's are dependent, we may use a more general CLT for dependent processes if the dependence among y_i and y_j disappear as $|i-j|$ increases. The same condition is also required for a single-step estimator.

Assumption 4.4 enables us to separate the sampling error in the first stage from the model error in the second stage. Importantly, since the first two moments of $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ do not depend on $\hat{\mathbf{B}}_n$, then the limiting conditional distribution also does not depend on $\hat{\mathbf{B}}_n$. Consequently, even *unconditionally*, $\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n)$ follows a standard normal distribution asymptotically (see Lemma A.1 in Appendix A.2).[5] The following theorem establishes the asymptotic distribution of Δ_n.

Theorem 4.2 (Aymptotic Distribution)**.** *Let $\boldsymbol{\psi}_n = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}\mathcal{E}_n$, where $\boldsymbol{\zeta} \sim N(0,\ \boldsymbol{I}_{K_\theta})$. Let F be the limiting distribution function of $\boldsymbol{\psi}_n$; that is, $F(\boldsymbol{t}) = \lim \mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^{K_\theta}$. Under Assumptions 2.1–4.2, and 4.4, we have $\lim \mathbb{P}(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \preceq \boldsymbol{t}) = F(\boldsymbol{t})$.*

Proof. A formal proof is presented in Appendix A.2. We have $\Delta_n = \mathbf{A}_n^{-1}\mathbf{V}_n^{1/2}\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) + \mathbf{A}_n^{-1}\mathcal{E}_n$. By assumption 4.4, $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ is asymptotically normally distributed. We show that $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ can thus be substituted with $\boldsymbol{\zeta}$. By the Slutsky theorem, we also substitute $\mathbf{A}_n$ with its limit. □

Theorem 4.2 states that Δ_n and $\boldsymbol{\psi}_n$ are asymptotically identically distributed (a.i.d).[6] In the definition of $\boldsymbol{\psi}_n$, the sampling error from the first stage is captured by the term $\mathbf{A}_0^{-1}\mathcal{E}_n$. Because $\mathcal{E}_n$ may not be a centered variable, the limiting distribution of Δ_n may not have a zero mean. In addition, the asymptotic distribution function F may not be normal. This depends on the distribution of $\mathbf{A}_0^{-1}\mathcal{E}_0$.

We can approximate F using the empirical CDF of a sample of many independent variables that are a.i.d as $\boldsymbol{\psi}_n$. Specifically, let $\boldsymbol{\psi}_{n,s} = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta}_s + \mathbf{A}_0^{-1}\mathcal{E}_{n,s}$, for $s = 1,\ \ldots,\ \kappa$, where $\boldsymbol{\zeta}_1,\ \ldots,\ \boldsymbol{\zeta}_\kappa$ are independent variables from $N(0,\ \boldsymbol{I}_{K_\theta})$. By the LLN, the empirical distribution function that is defined by $F_n^\kappa(\boldsymbol{t}) = \frac{1}{\kappa}\sum_{s=1}^{\kappa} \mathbb{1}\{\boldsymbol{\psi}_{n,s} \preceq \boldsymbol{t}\}$, converges in probability to F, as n and κ grow to infinity.

A direct implication of Theorem 4.2 is that the sample of $\boldsymbol{\psi}_{n,s}$ can be used to construct confidence intervals (CIs). For any vector $\boldsymbol{b}$, we denote by $\boldsymbol{b}(\iota)$ the ι-th component of $\boldsymbol{b}$.

Corollary 4.1 (Confidence Intervals)**.** *Assume the conditions of Theorem 4.2 hold. Let T_α be the α empirical quantile of the sample $\{\hat{\boldsymbol{\theta}}_n(\iota) - \boldsymbol{\psi}_{n,s}(\iota)/\sqrt{n},\ s = 1,\ \ldots,\ \kappa\}$. Then, $[T_{\frac{\alpha}{2}},\ T_{1-\frac{\alpha}{2}}]$ is a consistent estimator of*

where LB is the set of all functions on $\mathbb{R}$ with Lipschitz norm bounded by one. See also Fligner and Hettmansperger (1979); Van der Vaart (2000) for examples and discussion on conditional asymptotic normality.

[5] This does not extend to the non-standardized IF because its conditional moments depend on $\hat{\mathbf{B}}_n$.

[6] In Appendix A.3, we extend Theorem 4.2 to the uniform convergence. We show that $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}} |\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_0\Delta_n \preceq \boldsymbol{t}) - G(\boldsymbol{t})| = o_p(1)$, where G is the limiting distribution function of $\boldsymbol{\zeta} + \mathbf{V}_n^{-1/2}\mathcal{E}_n$.

the $(1-\alpha)$ *CI of* $\boldsymbol{\theta}_0(\iota)$*, in the sense that* $\lim_{n,\kappa} \mathbb{P}(\boldsymbol{\theta}_0(\iota) \in [T_{\frac{\alpha}{2}},\ T_{1-\frac{\alpha}{2}}]) = 1-\alpha$.

Proof. See Appendix A.4. □

In practice, we can construct the variables $\boldsymbol{\psi}_{n,s}$ by replacing $\mathbf{A}_0$, $\mathcal{E}_{n,s}$ and $\mathbf{V}_n$ with their empirical counterparts. Let $\hat{\boldsymbol{\psi}}_{n,s} = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{V}}_n^{1/2}\boldsymbol{\zeta}_s + \hat{\mathbf{A}}_n^{-1}\hat{\mathcal{E}}_{n,s}$. We can estimate F by the empirical distribution function $\hat{F}_n^{\kappa}$ defined as

$$\hat{F}_n^{\kappa}(\boldsymbol{t}) = \tfrac{1}{\kappa}\textstyle\sum_{s=1}^{\kappa} \mathbb{1}\{\hat{\boldsymbol{\psi}}_{n,s} \preceq \boldsymbol{t}\}.$$

Unlike the case of the asymptotic variance, the fact that we do not use simulation from the true distribution $\hat{\mathbf{B}}_n$, but rather from an estimator of the asymptotic distribution, raises no issue here. To see why, let $\tilde{\boldsymbol{\psi}}_{n,s} = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}\tilde{\mathcal{E}}_n$, where $\tilde{\mathcal{E}}_{n,s}$ is obtained by replacing $\mathbf{B}_{n,s}$ in $\mathcal{E}_{n,s}$ with $\bar{\mathbf{B}}_{n,s}$ ($\boldsymbol{\theta}_0$ and $\mathbf{B}_0$ are not replaced with their empirical counterparts). Also, let $\tilde{F}_n^{\kappa}(\boldsymbol{t}) = \frac{1}{\kappa}\sum_{s=1}^{\kappa} \mathbb{1}\{\tilde{\boldsymbol{\psi}}_{n,s} \preceq \boldsymbol{t}\}$. As the ν-th moment of $\mathbb{1}\{\tilde{\boldsymbol{\psi}}_{n,s} \preceq \boldsymbol{t}\}$, for some $\nu > 1$, is necessarily finite, sufficient condition for $\tilde{F}_n^{\kappa}(\boldsymbol{t})$ to converge in probability to the theoretical CDF F (as κ and n grow to infinity) is that $\mathcal{E}_{n,s}$ and $\tilde{\mathcal{E}}_{n,s}$ have the same asymptotic distribution. This holds true because $\mathbf{B}_{n,s}$ and $\bar{\mathbf{B}}_{n,s}$ are a.i.d.

From Theorem 4.2, we also provide sufficient conditions to obtain asymptotic normality at the second stage. As highlighted earlier, this depends on the limiting distribution of $\mathbf{A}_0^{-1}\mathcal{E}_n$.

Corollary 4.2 (Asymptotic Normality)**.** *Under Assumptions 2.1–4.2, and 4.4, if* $\mathcal{E}_0 \sim N\big(\mathbb{E}(\mathcal{E}_0),\ \mathbb{V}(\mathcal{E}_0)\big)$*, then* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *converges in distribution to* $N\big(\mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0),\ \mathbf{A}_0^{-1}(\mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0))\mathbf{A}_0^{-1}\big)$.

Proof. See Appendix A.5. □

The expectation of the limiting distribution of Δ_n is given by $\mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)$ and may not be zero. For example, this can happen when the first stage involves estimating a high-dimensional parameter or when the first stage estimator converges slowly. Corollary 4.2 shares similarities with Theorem 1 of Cattaneo et al. (2019). Under regularity conditions, they show that $\mathbf{V}_n^{-1/2}\mathbf{A}_0(\Delta_n - \mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_n))$ is asymptotically normally distributed. The same result also follows from Corollary 4.2. Indeed, we have $\mathbf{V}_n^{-1/2}\mathbf{A}_n(\boldsymbol{\psi}_n - \mathbf{A}_n^{-1}\mathbb{E}(\mathcal{E}_n)) = \boldsymbol{\zeta} + \mathbf{V}_n^{-1/2}(\mathcal{E}_n - \mathbb{E}(\mathcal{E}_n))$, which is asymptotically normally distributed as the sum of two independent variables that are asymptotically normally distributed.

The following remark extends our discussion in Remark 4.1 on how to use the Delta approach to compute standard errors for estimates of marginal effects (MEs) or response functions. We now discuss the estimator of the asymptotic distribution.

Remark 4.2 (Inference for Functions of Plug-in Estimators)**.** *Krinsky and Robb (1990) propose a simulation approach for inferring* $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n)$ *using the asymptotic distribution of* Δ_n*. In contrast to the Delta method, their approach avoids the Taylor approximation. They consider a sample of estimated MEs given by* $f(\mathbf{X}_n,\ \hat{\boldsymbol{\theta}}_n^{(s)})$, $s = 1,\ \ldots,\ \kappa$*, where* $\hat{\boldsymbol{\theta}}_n^{(s)}$ *is a simulation from the estimator of the asymptotic distribution of*

$\hat{\boldsymbol{\theta}}_n$. *In their case, this estimator is generally a normal distribution centered at* $\hat{\boldsymbol{\theta}}_n$. *They demonstrate that the* $\frac{\alpha}{2}$ *and* $(1-\frac{\alpha}{2})$ *quantiles of this sample is the* $(1-\alpha)$ *CI of the ME. A similar approach can be used in our framework. A simulation from the estimator of the asymptotic distribution of* $\hat{\boldsymbol{\theta}}_n$ *is* $\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\psi}}_{n,s}/\sqrt{n}$, *where* $\hat{\boldsymbol{\psi}}_{n,s} = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{V}}_s^{1/2}\boldsymbol{\zeta}_s + \hat{\mathbf{A}}_n^{-1}\hat{\mathcal{E}}_{n,s}$. *Let* $\hat{\tau}_s = f(\mathbf{X}_n, \hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\psi}}_{n,s}/\sqrt{n})$. *The* $\frac{\alpha}{2}$ *and* $(1-\frac{\alpha}{2})$ *quantiles of the sample* $\{\hat{\tau}_s,\ s = 1,\ \ldots,\ \kappa\}$ *are the bounds of the* $(1-\alpha)$ *CI of the ME.*

4.3 Biased Plug-in Estimators

Plug-in estimators may exhibit significant bias due to a large imprecision of the first-stage estimator. Examples of situations where this issue can occur are when many covariates are involved in the first-stage estimation or when the number of observations in the first stage grows slowly with respect to n. While $\hat{\boldsymbol{\theta}}_n$ can still be consistent in such situations, the limiting distribution of Δ_n may not have a zero mean (Belloni et al., 2014b, 2017; Cattaneo et al., 2019). In this section, we discuss how our method can be used to handle such a situation.

One interesting feature of our approach is that it does not require the limiting distribution of Δ_n to have a zero mean. The asymptotic mean of Δ_n is $\mathbb{E}(\Delta_0) = \mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)$ and Condition (ii) of Assumption 4.1 only imposes that $\mathcal{E}_n$ has a limiting distribution, but $\mathbb{E}(\mathcal{E}_0)$ may not be zero. Having $\mathbb{E}(\mathcal{E}_0) \neq 0$ suggests a plug-in estimator with an important finite sample bias. Nevertheless, even in this condition, our approach can lead to reliable CIs for $\boldsymbol{\theta}_0$, although the CIs will not be centered at $\hat{\boldsymbol{\theta}}_n$ due to the finite sample bias. For instance, the CI of the ι-th component of $\boldsymbol{\theta}_0$ will be centered at the ι-th component of $\hat{\boldsymbol{\theta}}_n + \mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)/\sqrt{n}$, where $\mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)/\sqrt{n}$ approximates the finite sample bias of $\hat{\boldsymbol{\theta}}_n$. We illustrate this feature through Monte Carlo simulations with an IV model, where the number of instruments is of order $\sqrt{n}$. See DGP D in Section 5.

The good news is that we can estimate $\mathbf{A}_0$ and $\mathbb{E}(\mathcal{E}_0)$ and, therefore, the bias $\mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)/\sqrt{n}$. Consequently, we can recenter the limiting distribution of Δ_n to obtain a distribution with a zero mean. Our inference method thus offers a way to reduce the finite sample bias $\hat{\boldsymbol{\theta}}_n$. Using an estimator of the finite sample bias, we propose a debiased estimator and establish its asymptotic distribution. Specifically, we consider the estimator

$$\boldsymbol{\theta}^*_{n,\kappa} = \hat{\boldsymbol{\theta}}_n - \hat{\mathbf{A}}_n^{-1}\hat{\boldsymbol{\Omega}}_n^{\kappa}/\sqrt{n}, \tag{8}$$

where $\hat{\boldsymbol{\Omega}}_n^{\kappa} = \frac{1}{\kappa}\sum_{s=1}^{\kappa}\hat{\mathcal{E}}_{n,s}$ is an estimator of $\mathbb{E}(\mathcal{E}_0)$ as defined in Equation (7). The following theorem establishes the consistency of $\boldsymbol{\theta}^*_{n,\kappa}$ and its limiting distribution.

Theorem 4.3 (Debiased Estimator). *Assume that Assumptions 2.1–4.4 hold. Assume also that* $\frac{1}{\kappa}\sum_{s=1}^{\kappa}\hat{\mathcal{E}}_{n,s}$ *converges in probability to* $\mathbb{E}(\mathcal{E}_0)$ *as* n *and* κ *grow to infinity.*

(i) $\boldsymbol{\theta}^*_{n,\kappa}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\theta}_0$.

(ii) Let $\boldsymbol{\psi}^*_n = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}(\mathcal{E}_n - \mathbb{E}(\mathcal{E}_n))$ *and let* F^* *be the limiting distribution function of* $\boldsymbol{\psi}^*_n$; *that is,*

$F^*(\boldsymbol{t}) = \lim \mathbb{P}(\boldsymbol{\psi}_n^* \preceq \boldsymbol{t})$ *for all* $\boldsymbol{t} \in \mathbb{R}^{K_\theta}$*, then* $\lim_{n,\kappa} \mathbb{P}(\sqrt{n}(\boldsymbol{\theta}_{n,\kappa}^* - \boldsymbol{\theta}_0) \preceq \boldsymbol{t}) = F^*(\boldsymbol{t})$.

(iii) The limiting distribution of $\sqrt{n}(\boldsymbol{\theta}_{n,\kappa}^* - \boldsymbol{\theta}_0)$ *has a zero mean and a variance given by* $\mathbf{A}_0^{-1}(\mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0))\mathbf{A}_0^{-1}$.

Proof. See Appendix A.6. □

Assuming that $\frac{1}{\kappa}\sum_{s=1}^{\kappa} \hat{\mathcal{E}}_{n,s}$ converges in probability to $\mathbb{E}(\mathcal{E}_0)$ as n and κ grow to infinity is a weak condition. We have previously discussed this point when estimating $\boldsymbol{\Sigma}_0$ by $\boldsymbol{\Sigma}_n^{\kappa}$ in Equation (7).

As was the case of the standard plug-in estimator, we can construct an empirical sample for $\boldsymbol{\psi}_n^*$ to approximate the CDF F^*. Let $\hat{\mathbf{A}}_n^*$, $\hat{\mathcal{E}}_{n,s}^*$, and $\hat{\mathbf{V}}_n^*$, which are defined as $\hat{\mathbf{A}}_n$, $\hat{\mathcal{E}}_{n,s}$, and $\hat{\mathbf{V}}_{n,}$, respectively, with the difference that they are computing using $\boldsymbol{\theta}_{n,\kappa}^*$ and not $\hat{\boldsymbol{\theta}}_n$. Let also

$$\hat{\boldsymbol{\psi}}_{n,s}^* = (\hat{\mathbf{A}}_n^*)^{-1}(\hat{\mathbf{V}}_n^*)^{1/2}\boldsymbol{\zeta}_s + (\hat{\mathbf{A}}_n^*)^{-1}(\hat{\mathcal{E}}_{n,s}^* - \hat{\boldsymbol{\Omega}}_n^{*\kappa}), \quad \text{where} \quad \hat{\boldsymbol{\Omega}}_n^{*\kappa} = \tfrac{1}{\kappa}\textstyle\sum_{s=1}^{\kappa} \hat{\mathcal{E}}_{n,s}^*.$$

We can estimate $F^*(\boldsymbol{t})$ by $\hat{F}_n^{*\kappa}(\boldsymbol{t}) = \frac{1}{\kappa}\sum_{s=1}^{\kappa} \mathbb{1}\{\hat{\boldsymbol{\psi}}_{n,s}^* \preceq \boldsymbol{t}\}$. We can also estimate the variance of the asymptotic distribution by $(\hat{\mathbf{A}}_n^*)^{-1}\{\hat{\mathbf{V}}_n^* + \frac{1}{\kappa-1}\sum_{s=1}^{\kappa}(\hat{\mathcal{E}}_{n,s}^* - \hat{\boldsymbol{\Omega}}_n^{*\kappa})(\hat{\mathcal{E}}_{n,s}^* - \hat{\boldsymbol{\Omega}}_n^{*\kappa})'\}(\hat{\mathbf{A}}_n^*)^{-1}$.

The debiased estimator $\boldsymbol{\theta}_{n,\kappa}^*$ is more general and would closely resemble the classical estimator if the latter does not suffer from finite sample bias. This feature is interesting since necessary conditions for the validity of the classical asymptotic inference method may not be easily testable in certain contexts. Our approach can help to prevent potential biases of which practitioners may not be aware. A practical way of knowing whether the debiased estimator should be preferred is to verify if the CIs from Corollary 4.1 are centered at $\hat{\boldsymbol{\theta}}_n$. If this is not the case, the debiased estimator can be employed to alleviate the bias in the classical estimator.

One issue regarding the debiased estimator is that the estimate of the finite sample bias of $\hat{\boldsymbol{\theta}}_n$ may be biased if the estimator of the first-stage distribution is biased. This situation can arise in small samples when the first-stage estimate is bounded in some interval with a large variance. In such cases, draws from the first-stage distribution may be equal to the bounds, leading to a biased estimate for $\mathbb{E}(\mathcal{E}_0)$. We illustrate this issue in our simulation study with a Copula-GARCH model. To mitigate the bias of the estimate of $\mathbb{E}(\mathcal{E}_0)$, we replace the sampling mean $\hat{\boldsymbol{\Omega}}_n^{\kappa}$ in Equation (8) with the sampling median of $\{\hat{\mathcal{E}}_{n,s},\ s = 1, \ldots, \kappa\}$. The median correction yields better performance and avoids the problem of outliers in $\hat{\mathcal{E}}_{n,s}$. Importantly, Theorem 4.3 still holds with the median correction if the distribution of $\mathcal{E}_0$ is symmetric (given that the mean and the median of $\mathcal{E}_0$ would be equal).

Furthermore, our approach requires the practitioner to possess a reliable estimate for the asymptotic distribution of $\hat{\mathbf{B}}_n$. As the first stage would generally be a single-stage estimation, a consistent estimator of the asymptotic distribution can be achieved. Even though many covariates are included in the first stage, existing literature suggests various methods for valid inference (see Zhang and Cheng, 2017; Chernozhukov et al., 2015; Belloni et al., 2016, and references therein).

Finally, our approach can also be used with other bias reduction methods that fall within the class of conditional extremum estimators that are considered in this paper. An example is the double de-

biased technique that is commonly used in machine learning (Chernozhukov et al., 2017, 2018). This method involves splitting the first-stage sample into multiple samples and constructing an estimator on each sample. The second stage employs a classical extremum estimation by using the estimators that are obtained in the first stage. While this debiased technique requires asymptotic normality in the second stage, it can be combined with our approach to be used in more general cases where the asymptotic distribution of the first-stage estimator is not normal.

5 Simulation Study

In this section, we conduct a simulation study using various data-generating processes (DGPs) to assess the finite sample performance of the proposed CDF estimator and the debiased estimator.

5.1 Data-Generating Processes

We study four DGP, denoted as DGP A–D. The sample size n takes values in $\{250,\ 500,\ 1000,\ 2000\}$. DGP A is a treatment effect model with endogeneity. The model is defined as follows:

$$y_i = \theta_0 d_i + \varepsilon_i, \quad d_i = \mathbb{1}\{z_i > 0.5(\varepsilon_i + 1.2)\}, \quad z_i \sim \text{Uniform}[0,\ 1], \quad \varepsilon_i \sim \text{Uniform}[-1, 1],$$

where d_i is a treatment status indicator ($d_i = 1$ if i is treated), z_i is an instrument for the treatment, and $\theta_0 = 1$. The treatment is endogenous given that it is correlated with the error term ε_i. The practitioner observes an i.i.d sample of $(y_i,\ d_i,\ z_i)$. We estimate θ_0 by the IV method. In the first stage, we predict $\mathbb{E}(y_i|z_i)$ using an OLS regression of y_i on z_i. We also predict $\mathbb{E}(d_i|z_i)$ using an OLS regression of d_i on z_i. The vector combining both OLS estimators constitutes the first-stage estimator. In the second stage, we regress the prediction of $\mathbb{E}(y_i|z_i)$ on the prediction of $\mathbb{E}(d_i|z_i)$.[7] Importantly, because the first stage estimator combines two estimators, we must simulate from the estimator of the joint asymptotic distribution of both estimators and not from each marginal distribution (see details in OA S.2).

DGP B is similar to DGP A with the difference that many regressors are involved in the first stage. Assume that the practitioner has access to $k_n = O(\sqrt{n})$ instruments, $z_{1,i},\ \ldots,\ z_{k_n,i}$. We maintain the specification of DGP A, but we change the treatment status for DGP B as follows:

$$d_i = \mathbb{1}\{0.2 + z_{1,i} + z_{2,i} + z_{3,i} + z_{4,i} > 0.5(\varepsilon_i + 1.2)\}, \quad z_{1,i},\ \ldots,\ z_{k_n,i} \overset{i.i.d}{\sim} \text{Uniform}[0,\ 0.2].$$

Only four instruments, $z_{1,i},\ \ldots,\ z_{4,i}$, are relevant for d_i. The others are superfluous variables that are independent of d_i. Yet, we include the k_n instruments in the first-stage regressions. We consider the cases $k_n = \lfloor 2\sqrt{n} \rceil$ and $k_n = \lfloor 4\sqrt{n} \rceil$, where $\lfloor . \rceil$ is the rounding to the nearest integer. DGP B falls within the frameworks of Cattaneo et al. (2019) and Mikusheva and Sun (2022).

[7] We could also regress y_i on the prediction of $\mathbb{E}(d_i|z_i)$. However, applying our method is easier when both y_i and d_i are projected in the space of $(1,\ z_i)'$ as we avoid the need for handling the correlation between $\mathbf{y}_n$ and $\hat{\mathbf{B}}_n$. See discussion on this point in Section 4.1.2.

DGP C is a Poisson model with a latent covariate that is defined as:

$$y_i \sim \text{Poisson}(\exp(\theta_{0,1} + \theta_{0,2} p_i)), \quad p_i = \sin^2(\pi z_i), \quad z_i \sim \text{Uniform}[0,\ 10], \quad d_i \sim \text{Bernoulli}(p_i),$$

where p_i is an unobserved probability and $\boldsymbol{\theta}_0 = (\theta_{0,1},\ \theta_{0,2})' = (-0.8,\ 2)'$. Consider an i.i.d. sample comprising data points $(y_i,\ z_i,\ d_i)$. The practitioner observes the pairs $(y_i,\ z_i)$ for all i but only observes d_i for a representative subsample of size $n^* = \lfloor n^{\alpha_n} \rfloor$. The parameter α_n takes values in $\{1,\ 0.985,\ 0.945,\ 0.91\}$, in the same order as n, i.e., $\alpha_n = 1$ if $n = 250$, $\alpha_n = 0.985$ if $n = 500$, and so forth. As p_i is not observed, $\boldsymbol{\theta}_0$ can be estimated in two stages. We assume that the practitioner only knows that p_i is a function of z_i but they do not know the exact specification. In the first stage, we estimate $p_i = \mathbb{E}(d_i|z_i)$ using a nonparametric regression of d_i on z_i in the subsample of size n^* where d_i is observed. We rely on a piecewise cubic spline approximation (see Hastie, 2017). The regression results can be used to compute $\hat{p}_i$, the estimator of p_i, for all i in the full sample, as we observe z_i for all i. The second stage is a standard Poisson regression after replacing p_i with its estimator.

DGP D is a multivariate time-series model similar to the model used in the simulation study by Gonçalves et al. (2023). We consider k_n returns $y_{1,i},\ \ldots,\ y_{k_n,i}$, where i is time and $k_n \geqslant 2$. Each $y_{p,i}$, for $p = 2,\ \ldots,\ k_n$, follows an AR(1)-GARCH(1, 1) defined as:

$$y_{p,i} = \phi_{p,0} + \phi_{p,1} y_{p,i-1} + \sigma_{p,i}\varepsilon_{p,i}, \quad \sigma^2_{p,i} = \beta_{p,0} + \beta_{p,1}\sigma^2_{p,i-1}\varepsilon^2_{p,i-1} + \beta_{p,2}\sigma^2_{p,i-1},$$

where $\phi_{p,0} = 0$, $\phi_{p,i-1} = 0.4$, $\beta_{p,0} = 0.05$, $\beta_{p,1} = 0.05$, $\beta_{p,2} = 0.9$, and $\varepsilon_{p,i}$ follows a standardized Student distribution of degree-of-freedom $\nu_p = 6$. The number of returns, k_n, increases with the sample size and takes values in $\{2,\ 3,\ 5,\ 8\}$, in the same order as n, i.e., $k_n = 2$ if $n = 250$, $k_n = 3$ if $n = 500$, and so forth. We account for the correlation between the returns using the Clayton copula (see Nelsen, 2006). The joint density function of $y_i = (y_{1,i},\ \ldots,\ y_{p,i})'$ conditional on $\mathcal{F}^{i-1}$ (information set at $i-1$) is given by $c_i(G_{1,i}(\boldsymbol{\beta}_{0,1}),\ \ldots,\ G_{k_n,i}(\boldsymbol{\beta}_{0,k_n}),\ \theta_0)$, where $\boldsymbol{\beta}_{0,p} = (\phi_{p,0},\ \phi_{p,1},\ \beta_{p,0},\ \beta_{p,1},\ \beta_{p,2},\ \nu_p)'$, $G_{p,i}(\boldsymbol{\beta}_{0,p})$ is the CDF of $y_{p,i}$ conditional on $\mathcal{F}^{i-1}$, and c_i is the PDF of k_n-dimensional Clayton copula of parameter $\theta_0 = 4$. The practitioner observes the sample $y_1,\ \ldots,\ y_n$. We rely on a multiple-stage estimation strategy to estimate $\log(\theta_0)$.[8] In the first k_n stages, we separately estimate each $\boldsymbol{\beta}_{0,p}$ by applying an AR(1)-GARCH(1, 1) model to the sample $y_{p,1},\ \ldots, y_{p,n}$. In the last stage, we estimate θ_0 by maximum likelihood (ML) after replacing $\boldsymbol{\beta}_{0,p}$ in the density function of y_i with its estimator.[9]

5.2 Simulation Results

We perform 10,000 simulations and set κ to 1,000.[10] We begin with the estimates of the CDF. We use the true sample distribution of Δ_n as the benchmark against which we compare our estimates.

[8] We consider $\log(\theta_0)$ for DGP C because the Clayton copula parameter must remain strictly positive.

[9] As for DGP A, We must consider the joint asymptotic distribution of the k_n first-stage estimators in the asymptotic distribution at the final stage, rather than the marginal distributions at each stage. An estimator of the joint asymptotic distribution of the GARCH estimators can be readily constructed using the score functions (see OA S.2).

[10] Replication codes can be found at https://github.com/ahoundetoungan/InferenceTSE.

Figures 2 and 3 display the average estimates of the CDF of Δ_n. The curve F_0 is the actual CDF of Δ_n, whereas the curve $\hat{F}_n$ represents the average estimate by simulation method. We also estimate the asymptotic CDF using Gaussian approximations. We consider the case where the sampling error from the first-stage estimation is disregarded (curve $\hat{H}_1$), and the case where we account for this sampling error using the simulation method that is proposed in Section 4.1 (curve $\hat{H}_2$). The L_1-Wasserstein distance between each estimated CDF and F_0 is in parentheses.[11]

For DGP A, both the $\hat{H}_2$ and $\hat{F}_n$ approximations yield strong performance, with the $\hat{H}_2$ providing a better fit of the actual CDF according to the Wasserstein distance. The result is not surprising since the first- and second-stage estimators are of type M, resulting in a normal asymptotic distribution (see Murphy and Topel, 2002). However, even when the asymptotic normality is verified, accurately computing the variance of a plug-in estimator can be intricate. The proposed method in Section 4.1, which is used for $\hat{H}_2$, excels in approximating this variance. In contrast, $\hat{H}_1$ approximation falls short as it overlooks the sampling error from the first stage.

Including many superfluous variables in the first stage of an IV approach can lead to biased estimates (see Cattaneo et al., 2019). We can observe this result with DGP B given that the true distribution of Δ_n is not centered at zero. Yet, our inference method captures this bias, whereas classical inference methods fail. The bias of our estimated CDF is larger when $k_n = \lfloor 4\sqrt{n} \rfloor$, but vanishes as n grows. Even though $\hat{H}_2$ accounts for the first-stage sampling error, it does not control for the bias at the second stage. In terms of Wasserstein distance to F_0, both $\hat{H}_1$ and $\hat{H}_2$ look similar. This result suggests that accounting for the first-stage sampling error can be as important as addressing the finite sample bias of the plug-in estimator.

In the case of DGP C, the size of the first-stage sample does not grow at the same rate as n. The first-stage estimator is $\sqrt{n^*}$-consistent, whereas the second-stage estimator is $\sqrt{n}$-consistent. This makes most classical inference approaches inapplicable. Even when both the first- and second-stage estimators are of type M, asymptotic normality may not be guaranteed in the second stage and the plug-in estimator can be biased. Our simulation approach performs well for both $\sqrt{n}(\hat{\theta}_{n,1} - \theta_{0,1})$ and $\sqrt{n}(\hat{\theta}_{n,2} - \theta_{0,2})$, where $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ represent the respective estimators for $\theta_{0,1}$ and $\theta_{0,2}$. Because of the low convergence rate in the first stage, the CDFs of $\sqrt{n}(\hat{\theta}_{n,1} - \theta_{0,1})$ and $\sqrt{n}(\hat{\theta}_{n,2} - \theta_{0,2})$ are not centered at zero and our approach captures this feature. Conversely, the normal approximations perform poorly. Once again, Wasserstein distances to F_0 reveal that accounting for the first-stage sampling error seems to be as important as addressing the finite sample bias in the second stage.

A notable distinction between DGP D and the other models is that $y_{p,i}$ is used in the p-th stage. This makes the estimator of $\boldsymbol{\beta}_{0,p}$ dependent on $y_{p,i}$. To compute the expectation and variance of

[11] The L_1-Wasserstein distance between a CDF $\hat{R}_n$ and F_0 is $\|\hat{R}_n - F_0\|_w = \int_{\mathbb{R}} |\hat{R}_n(t) - F_0(t)| dt$.

$\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ conditional on $\hat{\mathbf{B}}_n$, we assume that $\mathbf{y}_n$ and $\hat{\mathbf{B}}_n$ are independent. This condition, although not verified here, is innocuous because the first-stage estimators converge in probability to nonstochastic quantities. Simulation results show that our approach performs well but is somewhat less accurate in small samples. This discrepancy arises because inference for GARCH models requires a large number of observations.[12] Moreover, as the number of returns increases with n, $\sqrt{n}(\log(\hat{\theta}_n) - \log(\theta_0))$ exhibits bias. Our approach captures this bias, whereas the normal approximations perform poorly.

We now turn to the finite sample performance of the debiased estimator. Table 1 provides a summary of the estimates. In DGP A, where the classical plug-in estimator does not exhibit finite sample bias, the debiased estimator closely aligns with the classical estimator. This result confirms the generality of the debiased approach, as discussed in Section 4.3. The small discrepancy between the debiased estimator and the plug-in one is due to the finite number of simulations from the first-stage distribution. Increasing κ can reduce this gap.

For DGP B, our debiased estimator significantly reduces the bias of the classical estimator. For example, with $2\sqrt{n}$ instrumental variables involved in the first stage, the bias of the classical plug-in estimator is substantial at -0.101 for $n = 250$. In contrast, the debiased estimator reduces this bias to -0.017. Not surprisingly, the bias reduction performs less well with $4\sqrt{n}$ instrumental variables. The bias is -0.183 for the classical estimator in the smallest samples and is three times lower (-0.067) for the debiased estimator. Even in the largest sample, the classical estimator's bias remains higher at -0.074, while the debiased estimator exhibits reduced bias at -0.011.

The results are similar for DGP C. In the smallest sample, the biases of the estimates for $\theta_{0,1}$ and $\theta_{0,2}$ are 0.114 and -0.184, respectively. The debiased estimator mitigates these biases to 0.009 and -0.017, respectively. Given that the first-stage sample increases slowly, the classical estimates for $\theta_{0,1}$ and $\theta_{0,2}$ still display biases of 0.033 and -0.053 respectively for $n = 2,000$. In contrast, these biases are negligible for the debiased estimator.

The bias correction performs well for DGP D when $n = 2,000$. The classical estimator exhibits a bias of -0.086, while the debiased estimator's bias is ten times lower in absolute value (0.007). However, the correction is less effective in small samples. For $n = 250$, the debiased estimator exhibits a larger bias than the classical estimator. This result aligns with the discussion in Section 4.3 that the estimate of the finite sample bias may be substantially biased when the first-stage estimates are constrained with a large variance. The standard deviation of the debiased estimator is 15.520, which is 40 times higher than that of the classical estimator. This suggests large variances in the first stage, which can tighten the constraints imposed on the simulations from the first-stage asymptotic distributions

[12] Additional results, omitted from the paper, emphasize that the estimate of the asymptotic distribution in the first stage is inaccurate when $n = 250$. This inaccuracy yields a biased estimate of the asymptotic CDF in the second stage

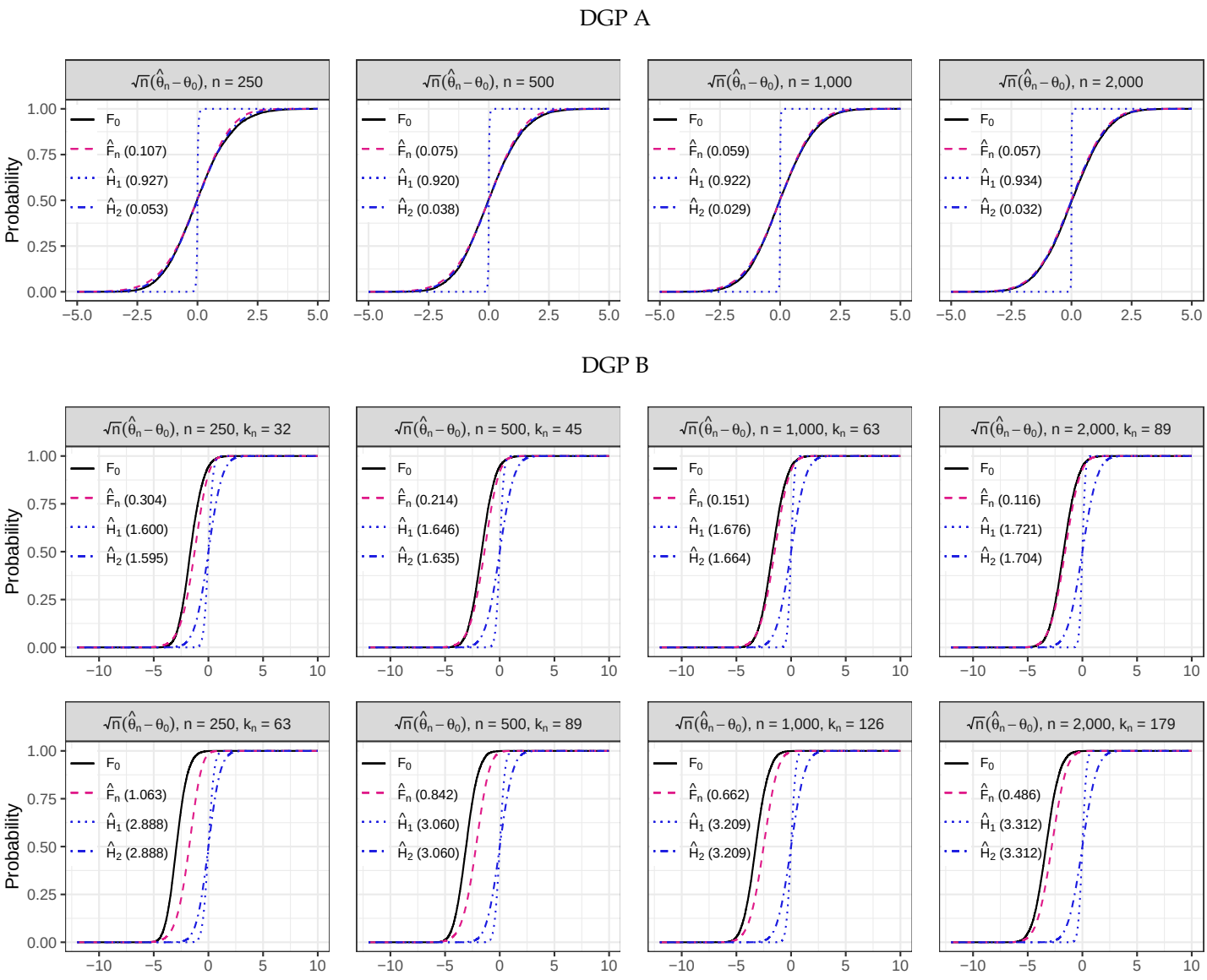

Figure 2: Monte Carlo Simulations: Estimates of asymptotic CDFs (DGPs A and B)

This figure illustrates the average estimates of the CDF of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ for DGPs A and B. F_0 represents the true sampling CDF. $\hat{H}_1$ denotes the CDF estimate when the first-stage sampling error is disregarded. $\hat{H}_2$ represents the estimate based on asymptotic normality, with the asymptotic variance estimated using our method. $\hat{F}_n$ corresponds to the CDF estimate obtained through our simulation approach. The L_1-Wasserstein distance between each estimated CDF and F_0 is enclosed in parentheses.

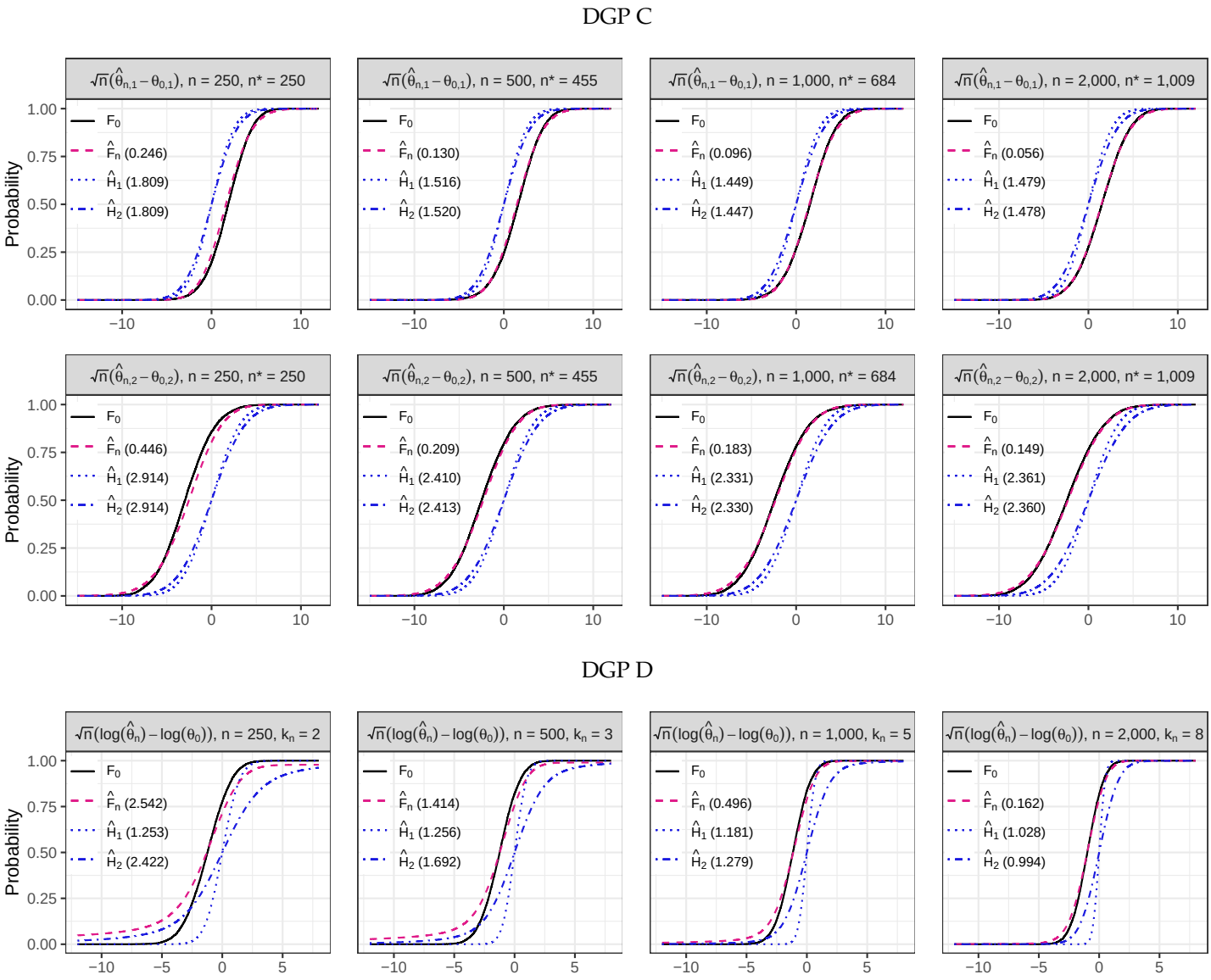

Figure 3: Monte Carlo Simulations: Estimates of asymptotic CDFs (DGPs C and D)

This figure illustrates the average estimates of the CDF of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ for DGPs C and D. F_0 represents the true sampling CDF. $\hat{H}_1$ denotes the CDF estimate when the first-stage sampling error is disregarded. $\hat{H}_2$ represents the estimate based on asymptotic normality, with the asymptotic variance estimated using our method. $\hat{F}_n$ corresponds to the CDF estimate obtained through our simulation approach. The L_1-Wasserstein distance between each estimated CDF and F_0 is enclosed in parentheses.

(constraints on GARCH model parameters include $\phi_{0,1}^2 < 1$ and $0 < \beta_{p,1} + \beta_{p,3} < 1$).

To address this issue, we compute the debiased estimator with the median correction (see Table 1). The result is much better than that of the mean correction. For $n = 250$, the debiased estimator now exhibits a bias of 0.072, that is, 11 times lower than the bias of the debiased estimator with the mean correction and about 4 times less important than the bias of the classical estimator. The standard deviation also decreases from 15.520 to 0.646.

In OA S.2.2, we present estimates of the asymptotic CDF of $\Delta_{n,\kappa}^* := \sqrt{n}(\boldsymbol{\theta}_{n,\kappa}^* - \theta_0)$. Unlike the case of the classical plug-in estimator, the true sampling CDFs are asymptotically centered at zero.

Table 1: Parameter Estimates Summary

	Classical plug-in estimator					Debiased plug-in estimator				
	Mean	Bias	Sd	RMSE	MAE	Mean	Bias	Sd	RMSE	MAE
DGP A: $\theta_0 = 1$										
N = 250	1.002	0.002	0.076	0.076	0.060	1.007	0.007	0.077	0.077	0.061
n = 500	1.001	0.001	0.053	0.053	0.042	1.003	0.003	0.053	0.053	0.042
n = 1000	1.001	0.001	0.037	0.037	0.030	1.002	0.002	0.037	0.037	0.030
n = 2000	1.001	0.001	0.026	0.026	0.021	1.001	0.001	0.026	0.026	0.021
DGP B: $k_n = \lfloor 2\sqrt{n} \rfloor$, $\theta_0 = 1$										
N = 250	0.899	−0.101	0.061	0.118	0.104	0.983	−0.017	0.076	0.077	0.062
n = 500	0.927	−0.073	0.045	0.086	0.076	0.992	−0.008	0.053	0.054	0.043
n = 1000	0.947	−0.053	0.033	0.062	0.054	0.996	−0.004	0.037	0.037	0.030
n = 2000	0.962	−0.038	0.024	0.045	0.039	0.998	−0.002	0.026	0.026	0.021
DGP B: $k_n = \lfloor 4\sqrt{n} \rfloor$, $\theta_0 = 1$										
N = 250	0.817	−0.183	0.053	0.190	0.183	0.933	−0.067	0.072	0.098	0.083
n = 500	0.863	−0.137	0.041	0.143	0.137	0.962	−0.038	0.052	0.064	0.053
n = 1000	0.899	−0.101	0.031	0.106	0.101	0.979	−0.021	0.037	0.043	0.035
n = 2000	0.926	−0.074	0.022	0.077	0.074	0.989	−0.011	0.026	0.028	0.023
DGP C: $\theta_{0,1} = -0.8$										
N = 250	−0.686	0.114	0.137	0.179	0.147	−0.791	0.009	0.158	0.158	0.126
n = 500	−0.732	0.068	0.103	0.124	0.101	−0.800	0.000	0.141	0.141	0.091
n = 1000	−0.754	0.046	0.076	0.089	0.072	−0.801	−0.001	0.082	0.082	0.065
n = 2000	−0.767	0.033	0.056	0.065	0.053	−0.800	0.000	0.059	0.059	0.047
DGP C: $\theta_{0,2} = 2$										
N = 250	1.816	−0.184	0.174	0.253	0.212	1.983	−0.017	0.210	0.211	0.168
n = 500	1.892	−0.108	0.131	0.170	0.140	1.999	−0.001	0.188	0.188	0.120
n = 1000	1.926	−0.074	0.098	0.122	0.100	2.001	0.001	0.108	0.108	0.085
n = 2000	1.947	−0.053	0.073	0.090	0.073	2.000	0.000	0.079	0.079	0.063
DGP D: $\theta_0 = 4$										
N = 250	3.720	−0.280	0.381	0.473	0.389	4.809	0.809	15.520	15.540	1.052
n = 500	3.793	−0.207	0.232	0.311	0.257	4.128	0.128	0.487	0.503	0.291
n = 1000	3.859	−0.141	0.147	0.204	0.169	4.034	0.034	0.216	0.218	0.147
n = 2000	3.914	−0.086	0.095	0.129	0.105	4.007	0.007	0.103	0.103	0.082
DGP D: $\theta_0 = 4$ (with a median correction for the debiased estimator)										
N = 250	3.720	−0.280	0.381	0.473	0.389	4.072	0.072	0.646	0.650	0.409
n = 500	3.793	−0.207	0.232	0.311	0.257	4.023	0.023	0.316	0.317	0.230
n = 1000	3.859	−0.141	0.147	0.204	0.169	4.003	0.003	0.170	0.170	0.133
n = 2000	3.914	−0.086	0.095	0.129	0.105	4.001	0.001	0.101	0.101	0.081

This table displays the summary of the parameter estimates, including their mean, bias, standard deviations (sd), root mean square error (RMSE), and mean absolute error (MAE).

6 Peer Effects in Adolescent Fast-Food Consumption Habits

In this section, we revisit the empirical analysis conducted by Fortin and Yazbeck (2015) on peer effects in adolescent fast-food consumption habits. Given the potential externalities that are associated with fast-food consumption and its link to overweight issues among adolescents, there may be justification for introducing a consumption tax on fast food. The optimal tax level hinges on the social multiplier of eating habits, emphasizing the need for an accurate measure of peer effects. Furthermore, the presence of peer effects implies that key players in the network can play a significant role in policies aimed at reducing the frequency of fast-food consumption in schools (Ballester et al., 2006). To address potential biases in estimation, we propose an IV approach using a large set of instruments, including *many weak instruments*. We reduce the bias of the estimate using our approach.

6.1 Estimation of Linear-in-Means Peer Effect Models

This section presents the model used in this application. We consider a set of R schools, where the number of students in the r-th school is denoted by n_r. Students within the same school interact. The network in the r-th school is represented by an adjacency matrix $\mathbf{G}_r = [g_{r,ij}]_{\substack{i=1,\,...,\,n_r \\ j=1,\,...,\,n_r}}$, where $g_{r,ij} = 1$ if student j is a friend of student i and $g_{r,ij} = 0$ otherwise. We restrict friendships to the same school; i.e., students from different schools cannot be friends. Moreover, self-friendships are not allowed, in the sense that $g_{r,ii} = 0$ for all r and i. We consider the following linear-in-means peer effect model:

$$y_{r,i} = \alpha_{0,r} + \theta_{0,1} \sum_{j=1}^{n_r} \frac{g_{r,ij}}{n_{r,i}} y_{r,j} + \boldsymbol{x}'_{r,i}\boldsymbol{\theta}_{0,2} + \sum_{j=1}^{n_r} \frac{g_{r,ij}}{n_{r,i}} \boldsymbol{x}'_{r,j}\boldsymbol{\theta}_{0,3} + \varepsilon_{r,i}, \tag{9}$$

where $y_{r,i}$ is the weekly fast-food consumption frequency of student i (reported frequency in days of fast-food restaurant visits in the past week), $\boldsymbol{x}_{r,i}$ is a vector of student i's observable characteristics, $\varepsilon_{r,i}$ is an error term assumed to be independent of $\boldsymbol{x}_{r,i}$ and $\mathbf{G}_r$, and $n_{r,i} = \sum_{j=1}^{n_r} g_{r,ij}$ is the number of friends of student i. The parameter $\theta_{0,1}$ captures peer effects, which measure the influence of an increase in the average friend's fast-food consumption frequency on one's fast-food consumption frequency.[13] The parameter $\boldsymbol{\theta}_{0,2}$ reflects the effect of student's characteristics, whereas $\boldsymbol{\theta}_{0,3}$ captures contextual effects, i.e., the influence of the average observable characteristics among friends. The parameter $\alpha_{0,r}$ accounts for unobserved effects of school characteristics, such as school location, regional taxes, and pricing policies.

Given that the number of schools (here $R = 16$) is relatively small compared to the sample size ($n = 2,535$), we do not face the incidental parameter issue by including school dummy variables for the fixed effects α_r.[14] It can be shown that the average friend's fast-food consumption frequency,

[13] The uniqueness of equilibrium in this model requires $|\theta_{0,1}| < 1$. That is, students do not increase their consumption frequency greater than the increase in their average friends' consumption frequency (see Bramoullé et al., 2009).

[14] It is possible to eliminate α_r by taking Equation (9) in difference with the average student at the school level. We do not use this approach because we observe a small number of schools.

which is measured by $\bar{y}_{r,i} = \frac{\sum_{j=1}^{n_r} g_{r,ij} y_{r,j}}{n_{r,i}}$, at the RHS of Equation (9) is correlated with the error term $\varepsilon_{r,i}$. Consequently, the classical OLS estimates of the parameters in (9) are likely to be inconsistent.

Fortunately, one does not need to seek instruments for $\bar{y}_{r,i}$ elsewhere; they can be generated from the model. For the sake of clarity, we rewrite Equation (9) in a matrix form for school r. Let $\mathbf{y}_r = (y_{r,1},\ \ldots,\ y_{r,n_s})'$, $\mathbf{X}_r = (\boldsymbol{x}_{r,1},\ \ldots,\ \boldsymbol{x}_{r,n_r})'$, and $\boldsymbol{\varepsilon}_r = (\varepsilon_{r,1},\ \ldots,\ \varepsilon_{r,n_s})'$.[15] Let also the row-normalized adjacency matrix $\tilde{\mathbf{G}}_r = [\tilde{g}_{r,ij}]_{\substack{i=1,\ \ldots,\ n_r \\ j=1,\ \ldots,\ n_r}}$, where $\tilde{g}_{r,ij} = 1/\sum_{j=1}^{n_r} g_{r,ij}$ if j is an i's friend and $\tilde{g}_{r,ij} = 0$ otherwise. The linear-in-means peer effect model at the school level is:

$$\mathbf{y}_r = \alpha_{0,r}\mathbf{1}_{n_r} + \theta_{0,1}\tilde{\mathbf{G}}_r\mathbf{y}_r + \mathbf{X}_r\boldsymbol{\theta}_{0,2} + \tilde{\mathbf{G}}_r\mathbf{X}_r\boldsymbol{\theta}_{0,3} + \boldsymbol{\varepsilon}_r, \tag{10}$$

where $\mathbf{1}_{n_r}$ is an n_r-dimensional vector of ones. By premultiplying the terms of Equation (10) by $\tilde{\mathbf{G}}_r$, we can observe that $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ is correlated with the endogenous variable $\tilde{\mathbf{G}}_r\mathbf{y}_r$ if $\boldsymbol{\theta}_{0,3} \neq 0$. Since $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ is not an explanatory variable in Equation (10), it can thus be served as an excluded instrument (see Kelejian and Prucha, 1998; Bramoullé et al., 2009). This instrument is interpreted as the average friends of their average friends of the characteristics $\boldsymbol{x}_{r,i}$.

However, the instrument $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ might suffer from weakness if $\boldsymbol{\theta}_{0,3} \approx 0$. To address this concern, we can show from Equation (10) that:

$$\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r) = \tilde{\mathbf{G}}_r(\mathbf{I}_{n_r} - \theta_{0,1}\tilde{\mathbf{G}}_r)^{-1}(\alpha_{0,r}\mathbf{1}_{n_r} + \mathbf{X}_r\boldsymbol{\theta}_{0,2} + \tilde{\mathbf{G}}_r\mathbf{X}_r\boldsymbol{\theta}_{0,3}). \tag{11}$$

Thus, it is possible to use $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$ as an instrument. This approach is optimal because $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$ fully captures the exogenous component of endogenous variable $\tilde{\mathbf{G}}_r\mathbf{y}_r$. In practice, employing this instrument entails a two-stage IV approach. A first IV method with $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ as an instrument is used to estimate $\boldsymbol{\theta}_0 = (\alpha_{0,r},\ \theta_{0,1},\ \boldsymbol{\theta}_{0,2}',\ \boldsymbol{\theta}_{0,3}')'$. This estimate can also be employed to approximate $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$ by replacing $\boldsymbol{\theta}_0$ in Equation (10) with its estimate. A second IV method is performed with the estimate of $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$ as an instrument to estimate $\boldsymbol{\theta}_0$.

While the optimal IV approach has been thoroughly considered in the literature, it may not entirely resolve the issue of weak instruments. Specifically, if the instrument $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ that is used in the first IV approach is weak, the estimation of $\boldsymbol{\theta}_0$ can be biased, leading to a biased estimate for $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$. To circumvent this problem, we propose a new approach that consists of expanding the set of instruments for $\tilde{\mathbf{G}}_r\mathbf{y}_r$. As $(\mathbf{I}_{n_r} - \theta_{0,1}\tilde{\mathbf{G}}_r)^{-1} = \sum_{p=0}^{\infty}\theta_{0,1}^p\tilde{\mathbf{G}}_r^p$, it can be shown from Equation (11) that:

$$\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r) = \alpha_{0,r}\tilde{\mathbf{G}}_r(\mathbf{I}_{n_r} - \theta_{0,1}\tilde{\mathbf{G}}_r)^{-1}\mathbf{1}_{n_r} + \tilde{\mathbf{G}}_r\mathbf{X}_r\boldsymbol{\theta}_{0,2} + \sum_{p=0}^{\infty}\theta_{0,1}^p\tilde{\mathbf{G}}_r^{2+p}\mathbf{X}_r(\theta_{0,1}\boldsymbol{\theta}_{0,2} + \boldsymbol{\theta}_{0,3}).$$

This suggests the use of $\tilde{\mathbf{G}}_r^{2+p}\mathbf{X}_r$, for $p = 0,\ 1,\ 2,\ \ldots, k_{\max}$ as instruments, where $k_{\max}$ can be as large as possible for the matrix of instruments to be full rank. This set of instruments can be interpreted as averages of $\boldsymbol{x}_{r,i}$ among close- and long-distance friends. We control for 25 students characteristics in $\boldsymbol{x}_{r,i}$ (see below) and set $k_{\max} = 9$. This leads to 250 excluded instruments for $\tilde{\mathbf{G}}_r\mathbf{y}_r$. Despite this

[15] The notations $\mathbf{y}_r$ and $\mathbf{X}_r$ are only used in this section and must not be confused with $\mathbf{y}_n$ and $\mathbf{X}_n$ used elsewhere.

large number of instruments, our inference method can be used to construct CIs for the parameters of the model (e.g., see DGP B in Section 5). Moreover, following Theorem 4.3, we correct for the finite sample bias of the resulting IV estimator.

6.2 Add Health Data

We use data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) survey. The purpose of this survey was to investigate how various social contexts (families, friends, peers, schools, neighborhoods, and communities) influence adolescents' health and risk behaviors. With such an objective, the survey provides nationally representative detailed information on adolescents in grades 7–12 from 144 schools during the 1994-95 school year in the United States (US). All students (around 90,000) were asked to answer a short questionnaire on demographics, family backgrounds, academic performance, and health-related behaviors, as well as friendship links (best friends within the same school, up to 5 females and up to 5 males). Subsequently, an in-home sample (core sample) of about 20,000 students was randomly drawn from each school. These students were asked to participate in a more extensive questionnaire where detailed questions were asked. This subsample was followed in-home in the subsequent waves of the survey. The fifth wave is the most recent one and was conducted in 2016-18.

We use the Wave II dataset, which encompasses most of the variables that are relevant to this study. This wave targets a subsample of 20,000 students tracked over time. However, only 16 schools, comprising approximately 3,000 students, were completely surveyed in Wave II. To avoid the issue of sampling networks, we focus on the sample that was derived from these 16 schools, where we can observe the entire list of nominated best friends, and all nominated friends are also surveyed. In addition to the Wave II dataset, we gather information on each student's race and their mother's background from Wave I.

Our dependent variable is the weekly fast-food consumption frequency, measured by the reported frequency (in days) of fast-food restaurant visits in the past week. The final sample consists of 2,735 students, with an equal distribution between boys and girls. We control for 25 observable characteristics in $\mathbf{X}_r$ such as students' gender, grade, race, weekly allowance, and parents' education and occupation. On average, students report consuming fast food 2.35 days per week. The average age of students at the time of Wave II data collection is 16.62 years. Additional details on the data summary can be found in Table S.1 in OA S.3.1.

6.3 Estimation and Inference

Figure 4 displays estimates of peer effects using our approach and alternative methods, including the OLS estimator, the classical IV (CIV) estimator, the optimal IV (OIV) estimator, our IV estimator

with many instruments (IV-MI), and the corresponding debiased IV estimator with many instruments (DIV-MI).[16] The OLS approach overlooks the endogeneity issue, whereas the CIV method uses $\tilde{\mathbf{G}}_r^2\mathbf{X}_r$ as an instrument. The OIV estimator employs the estimate of $\mathbb{E}(\tilde{\mathbf{G}}_r\mathbf{y}_r|\mathbf{X}_r, \tilde{\mathbf{G}}_r)$ as an instrument, replacing unknown parameters in Equation (11) with their CIV estimates.

The OLS estimate indicates that the peer effect parameter is significant. The estimate decreases from 0.192 to 0.150 when we control for school-fixed effects. In contrast, the CIV estimator has a large variance, indicating that the coefficient is not statistically significant. This imprecision is a consequence of the weakness of the instruments (Mikusheva and Sun, 2022), leading to a biased estimator toward the OLS one. For instance, although it is known that the model suffers from an endogeneity problem, the Hausman-Wu endogeneity test (not reported here) indicates that the OLS and CIV estimators are not significantly different.

As discussed earlier, this issue also invalidates the OIV approach since biased CIV estimates are used to estimate the optimal instrument. Notably, we observe that the 95% confidence interval of the OIV is even larger than that of CIV. While the estimator becomes more precise when we control for school-fixed effects, the results still indicate that peer effects are not significant.

These findings align with the results of Fortin and Yazbeck (2015). Their OIV estimate of the peer effect parameter is 0.110 with a standard error approximated at 0.395, indicating non-significance. Additionally, they implemented a quasi-maximum likelihood (QML) method, estimating peer effects at 0.129. However, the coefficient is significant only at the 10% level.

After expanding the pool of instruments, the IV estimator estimate with many instruments reveals significant peer effects. The estimate decreases from 0.276 to 0.208 after accounting for school-fixed effects. Moreover, the results highlight evidence of finite sample bias, as the confidence intervals are not centered on the estimates. This bias is a consequence of the numerous instrumental variables in the first stage. After correcting for this bias, the estimates slightly increase to 0.300 and 0.218, respectively. The increase is smaller when controlling for school-fixed effects.

In summary, our results highlight the presence of peer effects in adolescent fast-food consumption habits. These results have two important implications. First, key players in the network can play a crucial role as channels for influencing adolescent habits. The more key players are influenced, the greater the potential spread of this influence in the network (see Ballester et al., 2006; Zenou, 2016). Second, the social multiplier becomes crucial in determining the impact of a policy on adolescent fast-food consumption frequency. The social multiplier coefficient, given by $1/(1-\theta_{0,1})$, is estimated at 1.279 for the model with fixed effects, considering the finite sample bias. This implies that the effect of a tax increase on fast-food consumption frequency, in an environment where adolescents do not interact with each other, must be multiplied by 1.279 when they interact (see Agarwal et al., 2021).

[16] See full results, including coefficients of control variables in OA S.3.2.

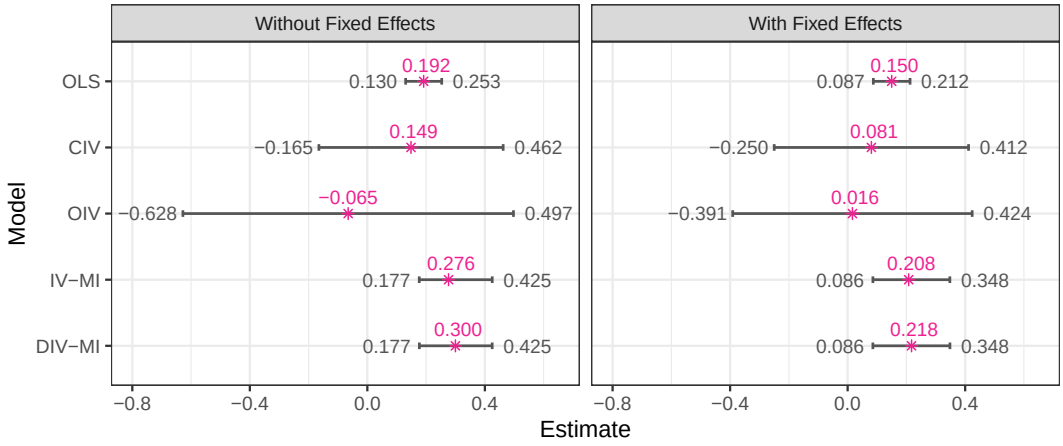

Figure 4: Peer Effect Estimates and Confidence Intervals

This figure presents peer effect estimates and confidence intervals. The asterisk symbols "$*$" denote the peer effect estimates, whereas the lines are the range of the 95% CIs of the peer effects.

7 Conclusion

This paper proposes a simulation-based approach for estimating the asymptotic variance and asymptotic CDF of two-stage estimators. We consider a large class of estimators in the first stage and an extremum estimator in the second stage. We show that researchers do not need restrictive conditions for the asymptotic distribution to be normally distributed with a zero mean.

A crucial issue regarding inference in two-stage estimation methods is that the asymptotic distribution of the plug-in estimator is influenced by the first-stage sampling error. We tackle this problem by disentangling the sampling error of the first stage from the model error in the second stage. Conditional on the first-stage estimator, the inference problem is similar to that of single-step extremum estimators, yielding asymptotic normality. We demonstrate that the practitioner can subsequently account for the sampling error from the first stage in the asymptotic distribution of the plug-in estimator using simulations from an estimator of the asymptotic distribution of the first stage.

We consider the possibility for the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ not to be normal, or centered at zero. We argue that this flexibility enables bias reduction for the plug-in estimator, particularly in cases where first-stage sampling error induces significant bias in the second stage. We introduce a debiased plug-in estimator and demonstrate that its limiting distribution has a zero mean. Finite sample performance confirms the finite sample performance of our debiased estimator.

We leverage the proposed approach to study peer effects on adolescents' fast-food consumption habits. We avoid the issue of weak instruments that can occur in this model by expanding the pool of instruments. This approach induces distortion in the estimates because the number of weak instru-

ments is large. Our method is designed to correct this bias and provide a valid inference.

This paper contributes to the emerging literature on inference methods when standard regularity conditions are violated. The proposed approach is straightforward to implement. We do not impose a particular class of estimator in the first stage. For instance, it may be a Bayesian estimator, which is not necessarily normally distributed. Our method is also suitable for complex models. Unlike resampling methods, we eliminate the need for multiple computations of both the first stage. We hope that the simplicity and flexibility of our approach will promote its use to infer multiple-stage estimators.

References

ACKERBERG, D., X. CHEN, AND J. HAHN (2012): "A practical asymptotic variance estimator for two-step semiparametric estimators," *Review of Economics and Statistics*, 94, 481–498.

AGARWAL, S., W. QIAN, AND X. ZOU (2021): "Thy neighbor's misfortune: Peer effect on consumption," *American Economic Journal: Economic Policy*, 13, 1–25.

AMEMIYA, T. (1985): *Advanced econometrics*, Harvard University Press.

ANDREWS, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica: Journal of the Econometric Society*, 59, 817–858.

——— (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, 62, 43–72.

——— (2002): "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

ANDREWS, I., J. H. STOCK, AND L. SUN (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.

ARMSTRONG, T. B., M. BERTANHA, AND H. HONG (2014): "A fast resample method for parametric and semiparametric models," *Journal of Econometrics*, 179, 128–133.

BALLESTER, C., A. CALVÓ-ARMENGOL, AND Y. ZENOU (2006): "Who's who in networks. Wanted: The key player," *Econometrica*, 74, 1403–1417.

BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND K. KATO (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

BELLONI, A., V. CHERNOZHUKOV, I. FERNANDEZ-VAL, AND C. HANSEN (2017): "Program evaluation and causal inference with high-dimensional data," *Econometrica*, 85, 233–298.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014a): "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28, 29–50.

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81, 608–650.

BELLONI, A., V. CHERNOZHUKOV, C. HANSEN, AND D. KOZBUR (2016): "Inference in high-dimensional panel models with an application to gun control," *Journal of Business & Economic Statistics*, 34, 590–605.

BOGACHEV, V. I. AND M. A. S. RUAS (2007): *Measure theory*, vol. 1, Springer.

BOUCHER, V. AND A. HOUNDETOUNGAN (2023): "Estimating peer effects using partial network data," *Unpublished manuscript*.

BRAMOULLÉ, Y., H. DJEBBARI, AND B. FORTIN (2009): "Identification of peer effects through social networks," *Journal of Econometrics*, 150, 41–55.

BRAMOULLÉ, Y., H. DJEBBARI, B. FORTIN, ET AL. (2020): "Peer Effects in Networks: A Survey," *Annual Review of Economics*, 12, 603–629.

BREZA, E., A. G. CHANDRASEKHAR, T. H. MCCORMICK, AND M. PAN (2020): "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, 110, 2454–84.

CASELLA, G. AND E. I. GEORGE (1992): "Explaining the Gibbs sampler," *The American Statistician*, 46, 167–174.

CATTANEO, M. D., M. JANSSON, AND X. MA (2019): "Two-step estimation and inference with possibly many included covariates," *The Review of Economic Studies*, 86, 1095–1122.

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, AND W. NEWEY (2017): "Double/debiased/Neyman machine learning of treatment effects," *American Economic Review*, 107, 261–265.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning," *The Econometrics Journal*, 21, 1–68.

CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2022): "Locally robust semiparametric estimation," *Econometrica*, 90, 1501–1535.

CHERNOZHUKOV, V., C. HANSEN, AND M. SPINDLER (2015): "Valid post-selection and post-regularization inference: An elementary, general approach," *Annu. Rev. Econ.*, 7, 649–688.

CHIB, S. AND E. GREENBERG (1995): "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, 49, 327–335.

CHUNG, K. L. (2001): *A Course in Probability Theory*, Academic Press, 3 ed.

DAVIDSON, R. AND J. G. MACKINNON (1999): "Bootstrap testing in nonlinear models," *International Economic Review*, 40, 487–508.

DE PAULA, A. (2017): "Econometrics of network models," in *Advances in economics and econometrics: Theory and applications, eleventh world congress*, Cambridge University Press, Cambridge, 268–323.

DUFAYS, A., E. A. HOUNDETOUNGAN, AND A. COËN (2022): "Selective Linear Segmentation for Detecting Relevant Parameter Changes," *Journal of Financial Econometrics*, 20, 762–805.

DZEMSKI, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

EFRON, B. (1982): *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics.

EKSTRÖM, M. (2014): "A general central limit theorem for strong mixing sequences," *Statistics & Probability Letters*, 94, 236–238.

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

FERNÁNDEZ-VAL, I. AND M. WEIDNER (2018): "Fixed effects estimation of large-T panel data models," *Annual Review of Economics*, 10, 109–138.

FLIGNER, M. A. AND T. P. HETTMANSPERGER (1979): "On the use of conditional asymptotic normality," *Journal of the Royal Statistical Society: Series B (Methodological)*, 41, 178–183.

FORTIN, B. AND M. YAZBECK (2015): "Peer effects, fast food consumption and adolescent weight gain," *Journal of health economics*, 42, 125–138.

FREYBERGER, J. AND B. J. LARSEN (2022): "Identification in ascending auctions, with an application to digital rights management," *Quantitative Economics*, 13, 505–543.

GONÇALVES, S., U. HOUNYO, A. J. PATTON, AND K. SHEPPARD (2023): "Bootstrapping two-stage quasi-maximum likelihood estimators of time series models," *Journal of Business & Economic Statistics*, 41, 683–694.

HASTIE, T. J. (2017): "Generalized additive models," in *Statistical models in S*, Routledge, 249–307.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

HONG, H. AND O. SCAILLET (2006): "A fast subsampling method for nonlinear dynamic models," *Journal of Econometrics*, 133, 557–578.

HONORÉ, B. E. AND L. HU (2017): "Poor (wo) man's bootstrap," *Econometrica*, 85, 1277–1301.

HOTZ, V. J. AND R. A. MILLER (1993): "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, 60, 497–529.

HOUNDETOUNGAN, E. A. (2022): "Count Data Models with Social Interactions under Rational Expectations," *Available at SSRN 3721250*.

HOUNDETOUNGAN, E. A. AND C. KOUAME (2023): "Identifying Peer Effects on Student Academic Effort," *Available at SSRN 4448048*.

ICHIMURA, H. AND S. LEE (2010): "Characterization of the asymptotic distribution of semiparametric M-estimators," *Journal of Econometrics*, 159, 252–266.

ICHIMURA, H. AND W. K. NEWEY (2022): "The influence function of semiparametric estimators," *Quantitative Economics*, 13, 29–61.

JIRAK, M. (2016): "Berry–Esseen theorems under weak dependence," *The Annals of Probability*, 44, 2024 – 2063.

JOFRE-BONET, M. AND M. PESENDORFER (2003): "Estimation of a dynamic auction game," *Econometrica*, 71, 1443–1489.

JOHANSEN, S. (1991): "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models," *Econometrica: Journal of the Econometric Society*, 59, 1551–1580.

KATO, K. (2011): "A note on moment convergence of bootstrap M-estimators," *Statistics & Decisions*, 28, 51–61.

KELEJIAN, H. H. AND I. R. PRUCHA (1998): "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances," *The Journal of Real Estate Finance and Economics*, 17, 99–121.

KLINE, P. AND A. SANTOS (2012): "A score based approach to wild bootstrap inference," *Journal of Econometric Methods*, 1, 23–41.

KRINSKY, I. AND A. L. ROBB (1990): "Approximating the Statistical Properties of Elasticities: A Correction," *Review of Economics and Statistics*, 72, 189–190.

LEE, L.-F., X. LIU, E. PATACCHINI, AND Y. ZENOU (2021): "Who is the key player? A network analysis of juvenile delinquency," *Journal of Business & Economic Statistics*, 39, 849–857.

LUBOLD, S., A. G. CHANDRASEKHAR, AND T. H. MCCORMICK (2023): "Identifying the latent space geometry of network models through analysis of curvature," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 240–292.

MIKUSHEVA, A. AND L. SUN (2022): "Inference with many weak instruments," *The Review of Economic Studies*, 89, 2663–2686.

MURPHY, K. M. AND R. H. TOPEL (2002): "Estimation and inference in two-step econometric models," *Journal of Business & Economic Statistics*, 20, 88–97.

NELSEN, R. B. (2006): *An Introduction to Copulas*, Springer.

NEWEY, W. K. (1984): "A method of moments interpretation of sequential estimators," *Economics Letters*, 14, 201–206.

NEWEY, W. K. AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

NEWEY, W. K. AND J. L. POWELL (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

RAIČ, M. (2019): "A multivariate Berry–Esseen theorem with explicit constants," *Bernoulli*, 25, 2824 – 2853.

ROMANO, J. P. AND M. WOLF (2000): "A more general central limit theorem for m-dependent random variables with unbounded m," *Statistics & probability letters*, 47, 115–124.

RUBSHTEIN, B.-Z. A. (1996): "A Central Limit Theorem for Conditional Distributions," in *Convergence in Ergodic Theory and Probability (Bergelson, March, Rosenblatt, eds.)*.

VAN DER VAART, A. W. (2000): *Asymptotic Statistics*, vol. 3, Cambridge University Press.

WITHERS, C. S. (1981): "Central limit theorems for dependent variables. I," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 509–534.

YAN, T., B. JIANG, S. E. FIENBERG, AND C. LENG (2019): "Statistical inference in a directed network model with covariates," *Journal of the American Statistical Association*, 114, 857–868.

ZELLNER, A. AND P. E. ROSSI (1984): "Bayesian analysis of dichotomous quantal response models," *Journal of Econometrics*, 25, 365–393.

ZENOU, Y. (2016): "Key players," *Oxford Handbook on the Economics of Networks*, 244–274.

ZHANG, X. AND G. CHENG (2017): "Simultaneous inference for high-dimensional linear models," *Journal of the American Statistical Association*, 112, 757–768.

A Appendix – Proofs

A.1 Proof of Theorem 4.1

Equation (5) is given by

$$\Delta_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \mathbf{A}_n^{-1}\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\,.$$

Let $\boldsymbol{\Sigma}_n = \mathbb{V}(\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n))$ and $\boldsymbol{\Sigma}_0 = \lim \boldsymbol{\Sigma}_n$. The existence of $\boldsymbol{\Sigma}_n$ and $\boldsymbol{\Sigma}_0$ is guaranteed by Assumption 4.1. As $\mathbb{E}(\|\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\|^\nu) < \infty$ for some $\nu > 2$, it follows that the limiting distribution of $\dot{\boldsymbol{q}}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ has variance $\boldsymbol{\Sigma}_0$ (see Chung, 2001, Theorem 4.5.2). Moreover, Assumption 4.2 implies that $\operatorname{plim} \mathbf{A}_n = \mathbf{A}_0$. Therefore, $\mathbb{V}(\Delta_0) = \mathbf{A}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{A}_0^{-1}$. We have

$$\boldsymbol{\Sigma}_n^\kappa = \mathbf{V}_n + \tfrac{1}{\kappa-1}\textstyle\sum_{s=1}^{\kappa}(\mathcal{E}_{n,s} - \boldsymbol{\Omega}_n^\kappa)(\mathcal{E}_{n,s} - \boldsymbol{\Omega}_n^\kappa)', \tag{12}$$

where $\mathbf{\Omega}_n^\kappa = \frac{1}{\kappa}\sum_{s=1}^{\kappa}\mathcal{E}_{n,s}$. The second term of the RHS of Equation (12) is the sampling variance of $\mathcal{E}_{n,s}$. The variables $\mathcal{E}_{n,1}, \ldots, \mathcal{E}_{n,\kappa}$ are independent and identically distributed as $\mathcal{E}_n$. By the LLN, $\frac{1}{\kappa-1}\sum_{s=1}^{\kappa}(\mathcal{E}_{n,s}-\mathbf{\Omega}_n^\kappa)(\mathcal{E}_{n,s}-\mathbf{\Omega}_n^\kappa)'$ converges in probability to $\mathbb{V}(\mathcal{E}_n)$ as κ grows to infinity; that is, $\operatorname{plim}_\kappa \mathbf{\Sigma}_n^\kappa = \mathbf{V}_n + \mathbb{V}(\mathcal{E}_n)$, where plim_k is the limit in probability as κ grows to infinity (n set fixed).

As $\mathbb{E}(\|\mathcal{E}_n\|^\nu) \leqslant \mathbb{E}(\|\dot{\boldsymbol{q}}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n)\|^\nu) < \infty$ for some $\nu > 2$, we have $\lim \mathbb{V}(\mathcal{E}_n) = \mathbb{V}(\mathcal{E}_0)$. Additionally, $\operatorname{plim} \mathbf{V}_n = \mathbf{V}_0$ by Condition (ii) of Assumption 4.1. As a result, $\operatorname{plim}_{n,\kappa} \mathbf{\Sigma}_n^\kappa = \mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0) = \mathbf{\Sigma}_0$ and $\mathbb{V}(\Delta_0) = \mathbf{A}_0^{-1}(\operatorname{plim}_{n,\kappa} \mathbf{\Sigma}_n^\kappa)\mathbf{A}_0^{-1}$.

A.2 Proof of Theorem 4.2

Let $F_{u,n}(\mathbf{t}) = \mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^{K_\theta}$. We first state and show the following lemma.

Lemma A.1 (Unconditional Asymptotic Normality)**.** *Under Assumption 4.4, the unconditional distribution of $\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n)$ is asymptotically normal, in the sense that $\lim F_{u,n}(\mathbf{t}) = \Phi(\boldsymbol{t})$ for each $\boldsymbol{t}$.*

Proof. We have $F_{u,n}(\mathbf{t}) = \mathbb{E}[\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n)]$, where the expectation is taken with respect to $\hat{\mathbf{B}}_n$. Thus, $\lim F_{u,n}(\mathbf{t}) = \lim \mathbb{E}\{\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n)\}$. Given that $|\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n)| \leqslant 1$, we can interchange the expectation and the limit in the previous equation (see Lebesgue–Vitali theorem in Bogachev and Ruas, 2007, Theorem 4.5.4). It follows that $\lim F_{u,n}(\mathbf{t}) = \mathbb{E}\{\lim \mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n)\}$. By Assumption 4.4, $\lim \mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\hat{\mathbf{B}}_n) = \Phi(\boldsymbol{t})$. Hence, $\lim F_{u,n}(\mathbf{t}) = \Phi(\boldsymbol{t})$. □

We now show Theorem 4.2. By substituting $\dot{\boldsymbol{q}}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) = \mathbf{V}_n^{1/2}\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) + \mathcal{E}_n$ in Equation (5) we obtain $\Delta_n = \mathbf{A}_n^{-1}\mathbf{V}_n^{1/2}\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n) + \mathbf{A}_n^{-1}\mathcal{E}_n$. By Lemma A.1, $\boldsymbol{u}_n(\mathbf{y}_n, \hat{\mathbf{B}}_n)$ converges in distribution to a standard normal distribution, independently of $\hat{\mathbf{B}}_n$. This convergence also holds independently of $\mathcal{E}_n$ since the latter is random only because of $\hat{\mathbf{B}}_n$. Recall also that $\operatorname{plim} \mathbf{A}_n = \mathbf{A}_0$ and $\operatorname{plim} \mathbf{V}_n = \mathbf{V}_0$, where $\mathbf{A}_0$ and $\mathbf{V}_0$ are nonstochastic. Therefore, by the Slutsky theorem, the conditional distribution of Δ_n, given $\mathcal{E}_n$, and the conditional distribution of $\boldsymbol{\psi}_n = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}\mathcal{E}_n$, given $\mathcal{E}_n$, have the same limit, where $\boldsymbol{\zeta} \sim N(0, \boldsymbol{I}_{K_\theta})$. Put differently, $\operatorname{plim} \mathbb{P}(\Delta_n \preceq \boldsymbol{t}|\mathcal{E}_n) = \operatorname{plim} \mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t}|\mathcal{E}_n)$. As this is true for almost all $\mathcal{E}_n$, it follows that $\mathbb{E}\big(\operatorname{plim} \mathbb{P}(\Delta_n \preceq \boldsymbol{t}|\mathcal{E}_n)\big) = \mathbb{E}\big(\operatorname{plim} \mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t}|\mathcal{E}_n)\big)$. As is the case in the proof of Lemma A.1, we can interchange the expectation and the limit because $|\mathbb{P}(\Delta_n \preceq \boldsymbol{t}|\mathcal{E}_n)| \leqslant 1$ and $|\mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t}|\mathcal{E}_n)| \leqslant 1$. Thus, $\lim \mathbb{E}\big(\mathbb{P}(\Delta_n \preceq \boldsymbol{t}|\mathcal{E}_n)\big) = \lim \mathbb{E}\big(\mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t}|\mathcal{E}_n)\big)$. As a result, $\lim \mathbb{P}(\Delta_n \preceq \boldsymbol{t}) = \lim \mathbb{P}(\boldsymbol{\psi}_n \preceq \boldsymbol{t})$. This completes the proof of Theorem 4.2.

A.3 Uniform Convergence of the Distribution of the Plug-in Estimator

We establish a more general result than Theorem 4.2 in terms of uniform convergence. To achieve this result, we consider a stronger version of Assumption 4.4 using uniform convergence.

Assumption A.1 (Conditional Asymptotic Normality). *The conditional distribution of the standardized influence function* $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$*, given* $\hat{\mathbf{B}}_n$*, uniformly converges in distribution to* $N(0,\ \boldsymbol{I}_{K_\theta})$*, almost surely; that is,* $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\hat{\mathbf{B}}_n)-\Phi(\boldsymbol{t})|=o_p(1)$.

Assumption A.1 can be implied by a strong version of the CLT to $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ conditional on $\hat{\mathbf{B}}_n$. This variant of the CLT, often referred to as the uniform CLT, is an improvement of the standard CLT, as introduced by Berry and Esseen (see Jirak, 2016; Raič, 2019). Given $\hat{\mathbf{B}}_n$ set fixed as a predetermined sequence, the uniform CLT can be applied as in the case of a single-step approach. As in Lemma A.1, Assumption A.1 also implies the uniform convergence of the unconditional distribution.

Lemma A.2 (Unconditional Asymptotic Normality). *Under Assumption A.1,* $\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)$ *uniformly converges to a standard normal distribution, in the sense that* $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t})-\Phi(\boldsymbol{t})|=o(1)$.

Proof. Let $m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n)=\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\hat{\mathbf{B}}_n)-\Phi(\boldsymbol{t})$. We have $\mathbb{E}(m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n))=\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t})-\Phi(\boldsymbol{t})$, where the expectation is taken with respect to $\hat{\mathbf{B}}_n$. As $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{E}(m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n))|\leqslant\mathbb{E}(\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n)|)$, it turns out that $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t})-\Phi(\boldsymbol{t})|\leqslant\mathbb{E}(\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n)|)$. Since $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n)|$ is bounded by one and is $o_p(1)$, then $\lim\mathbb{E}(\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|m_{\boldsymbol{t}}(\hat{\mathbf{B}}_n)|)=0$. This completes the proof. □

The following theorem establishes the uniform convergence of the distribution Δ_n.

Theorem A.1 (Aymptotic Distribution). *Assumptions 2.1–4.2, and A.1 hold. Let* $\boldsymbol{\chi}_n=\boldsymbol{\zeta}+\mathbf{V}_n^{-1/2}\mathcal{E}_n$*, where* $\boldsymbol{\zeta}\sim N(0,\ \boldsymbol{I}_{K_\theta})$*. Let* $G(\boldsymbol{t})=\lim\mathbb{P}(\boldsymbol{\chi}_n\preceq\boldsymbol{t})$ *for* $\boldsymbol{t}\in\mathbb{R}^{K_\theta}$ *be the limiting distribution function of* $\boldsymbol{\chi}_n$*. We have* $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}(\sqrt{n}\mathbf{V}_n^{-1/2}\mathbf{A}_0(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\preceq\boldsymbol{t})-G(\boldsymbol{t})|=o(1)$.

Proof. As $\{\boldsymbol{\chi}_n\preceq\boldsymbol{t}\}=\{\boldsymbol{\zeta}\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n\}$, we have $G(\boldsymbol{t})=\lim\mathbb{E}\big\{\mathbb{P}(\boldsymbol{\zeta}\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n|\mathcal{E}_n)\big\}$. Thus,

$$G(\boldsymbol{t})=\lim\mathbb{E}\{\Phi(\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n)\},\tag{13}$$

because $\operatorname{plim}\mathbf{V}_n$ is nonstochastic and $\boldsymbol{\zeta}\sim N(0,\ \boldsymbol{I}_{K_\theta})$. Moreover, as $\Delta_n=\mathbf{A}_n^{-1}\mathbf{V}_n^{1/2}\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)+\mathbf{A}_n^{-1}\mathcal{E}_n$, then $\{\mathbf{V}_n^{-1/2}\mathbf{A}_n\Delta_n\preceq\boldsymbol{t}\}=\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n\}$. This translates to $\lim\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_0\Delta_n\preceq\boldsymbol{t})=\lim\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n\}$ because $\operatorname{plim}\mathbf{A}_n=\mathbf{A}_0$ is nonstochastic. Since $\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n\}=\mathbb{E}\{\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n|\mathcal{E}_n)\}$, we thus have

$$\lim\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_n\Delta_n\preceq\boldsymbol{t})=\lim\mathbb{E}\big\{\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n|\mathcal{E}_n\}\big\}.\tag{14}$$

From (13) and (14), $\lim|\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_0\Delta_n\preceq\boldsymbol{t})-G(\boldsymbol{t})|=\lim|\mathbb{E}\big\{\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n|\mathcal{E}_n\}-\Phi(\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n)\big\}|$. In addition, For any $\boldsymbol{t}$ and $\mathcal{E}_n$, we have
$\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n|\mathcal{E}_n\}-\Phi(\boldsymbol{t}-\mathbf{V}_n^{-1/2}\mathcal{E}_n)\leqslant\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\mathcal{E}_n\}-\Phi(\boldsymbol{t})|$. Thus, $\lim|\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_0\Delta_n\preceq\boldsymbol{t})-G(\boldsymbol{t})|\leqslant\lim\mathbb{E}\big\{\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\mathcal{E}_n\}-\Phi(\boldsymbol{t})|\big\}$.

By Lemma A.2, $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\mathcal{E}_n\}-\Phi(\boldsymbol{t})|=o_p(1)$ because conditioning on $\mathcal{E}_n$ involves conditioning on $\hat{\mathbf{B}}_n$. As $\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}}|\mathbb{P}\{\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n)\preceq\boldsymbol{t}|\mathcal{E}_n\}-\Phi(\boldsymbol{t})|$ is bounded, this implies

that $\lim \mathbb{E}\{\sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}} |\mathbb{P}(\boldsymbol{u}_n(\mathbf{y}_n,\ \hat{\mathbf{B}}_n) \preceq \boldsymbol{t}|\mathcal{E}_n) - \Phi(\boldsymbol{t})|\} = 0$. As a result, $\lim \sup_{\boldsymbol{t}\in\mathbb{R}^{K_\theta}} |\mathbb{P}(\mathbf{V}_n^{-1/2}\mathbf{A}_0\Delta_n \preceq \boldsymbol{t}) - G(\boldsymbol{t})| = 0$. This completes the proof. □

A.4 Proof of Corollary 4.1

By the definition of Δ_n, we have $\boldsymbol{\theta}_0(\iota) = \hat{\boldsymbol{\theta}}_n(\iota) - \Delta_n/\sqrt{n}$. Thus , $\mathbb{P}(\boldsymbol{\theta}_0(\iota) \in [T_{\frac{\alpha}{2}},\ T_{1-\frac{\alpha}{2}}]) = \mathbb{P}(\Delta_n(\iota) \in [\sqrt{n}(\hat{\boldsymbol{\theta}}_n(\iota) - T_{1-\frac{\alpha}{2}}),\ \sqrt{n}(\hat{\boldsymbol{\theta}}_n(\iota) - T_{\frac{\alpha}{2}})])$, with $\sqrt{n}(\hat{\boldsymbol{\theta}}_n(\iota) - T_{1-\frac{\alpha}{2}})$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n(\iota) - T_{\frac{\alpha}{2}})$ being the empirical quantiles of the sample $\{\boldsymbol{\psi}_{n,s}(\iota),\ s = 1,\ \ldots,\ \kappa\}$ at $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$, respectively. Since $\Delta_n(\iota)$ and $\boldsymbol{\psi}_{n,s}(\iota)$ have the same limiting distribution when n and κ grow to infinity, these quantiles have the same limit as the theoretical quantiles of $\Delta_n(\iota)$ at $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$. This completes the proof.

A.5 Proof of Corollary 4.2

By Condition (ii) of Theorem 4.1, $\{\mathbf{V}_n,\ \mathcal{E}_n\}$ converges in distribution to $\{\mathbf{V}_0,\ \mathcal{E}_0\}$. As $\{\mathbf{V}_n,\ \mathcal{E}_n\}$ is independent of $\boldsymbol{\zeta}$, Theorem 4.2 implies that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to $\boldsymbol{\psi}_0 := \mathbf{A}_0^{-1}\mathbf{V}_0^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}\mathcal{E}_0$, which is the sum of two independent variables. As a result, $\boldsymbol{\psi}_0$ is normally distributed with mean $\mathbb{E}(\boldsymbol{\psi}_0) = \mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)$ and variance $\mathbb{V}(\boldsymbol{\psi}_0) = \mathbf{A}_0^{-1}(\mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0))\mathbf{A}_0^{-1}$.

A.6 Proof of Theorem 4.3

We have $\sqrt{n}(\boldsymbol{\theta}^*_{n,\kappa} - \boldsymbol{\theta}_0) = \Delta_n - \hat{\mathbf{A}}_n^{-1}\hat{\boldsymbol{\Omega}}_n^{\kappa}$. As $\Delta_n = O_p(1)$, $\operatorname{plim} \hat{\mathbf{A}}_n = \mathbf{A}_0$, and $\operatorname{plim}_{\kappa,n} \hat{\boldsymbol{\Omega}}_n^{\kappa} = \mathbb{E}(\mathcal{E}_0)$, it follows that $\sqrt{n}(\boldsymbol{\theta}^*_{n,\kappa} - \boldsymbol{\theta}_0) = O_p(1)$, that is, $\boldsymbol{\theta}^*_{n,\kappa}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$. This completes the proof of Statement (i).

Δ_n has the same asymptotic distribution as $\boldsymbol{\psi}_n = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}\mathcal{E}_n$ and $\hat{\mathbf{A}}_n^{-1}\hat{\boldsymbol{\Omega}}_n^{\kappa}$ converges in probability to the constant $\mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0)$, as κ and n grow to infinity. Thus, by the Slutsky theorem, $\sqrt{n}(\boldsymbol{\theta}^*_{n,\kappa} - \boldsymbol{\theta}_0) = \Delta_n - \hat{\mathbf{A}}_n^{-1}\hat{\boldsymbol{\Omega}}_n^{\kappa}$ has the same asymptotic distribution as $\boldsymbol{\psi}_n - \mathbf{A}_0^{-1}\mathbb{E}(\mathcal{E}_0) = \mathbf{A}_0^{-1}\mathbf{V}_n^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}(\mathcal{E}_n - \mathbb{E}(\mathcal{E}_0))$. As a result $\boldsymbol{\psi}^*_n$ and $\sqrt{n}(\boldsymbol{\theta}^*_{n,\kappa} - \boldsymbol{\theta}_0)$ have the same distribution, have κ and n grow to infinity. This completes the proof of Statement (ii).

Statement (iii) is a direct implication of Statement (ii). The variable $\boldsymbol{\psi}^*_n$ converges in distribution to $\boldsymbol{\psi}^*_0 := \mathbf{A}_0^{-1}\mathbf{V}_0^{1/2}\boldsymbol{\zeta} + \mathbf{A}_0^{-1}(\mathcal{E}_0 - \mathbb{E}(\mathcal{E}_0))$, with $\mathbb{E}(\boldsymbol{\psi}^*_0) = 0$ and $\mathbb{V}(\boldsymbol{\psi}^*_0) = \mathbf{A}_0^{-1}(\mathbf{V}_0 + \mathbb{V}(\mathcal{E}_0))\mathbf{A}_0^{-1}$. As $\boldsymbol{\psi}^*_n$ and $\sqrt{n}(\boldsymbol{\theta}^*_{n,\kappa} - \boldsymbol{\theta}_0)$ have the same distribution, the result follows.

B Online Appendix

Supplementary material related to this paper can be found online at https://ahoundetoungan.com/files/Papers/InferenceTSE_OA.pdf